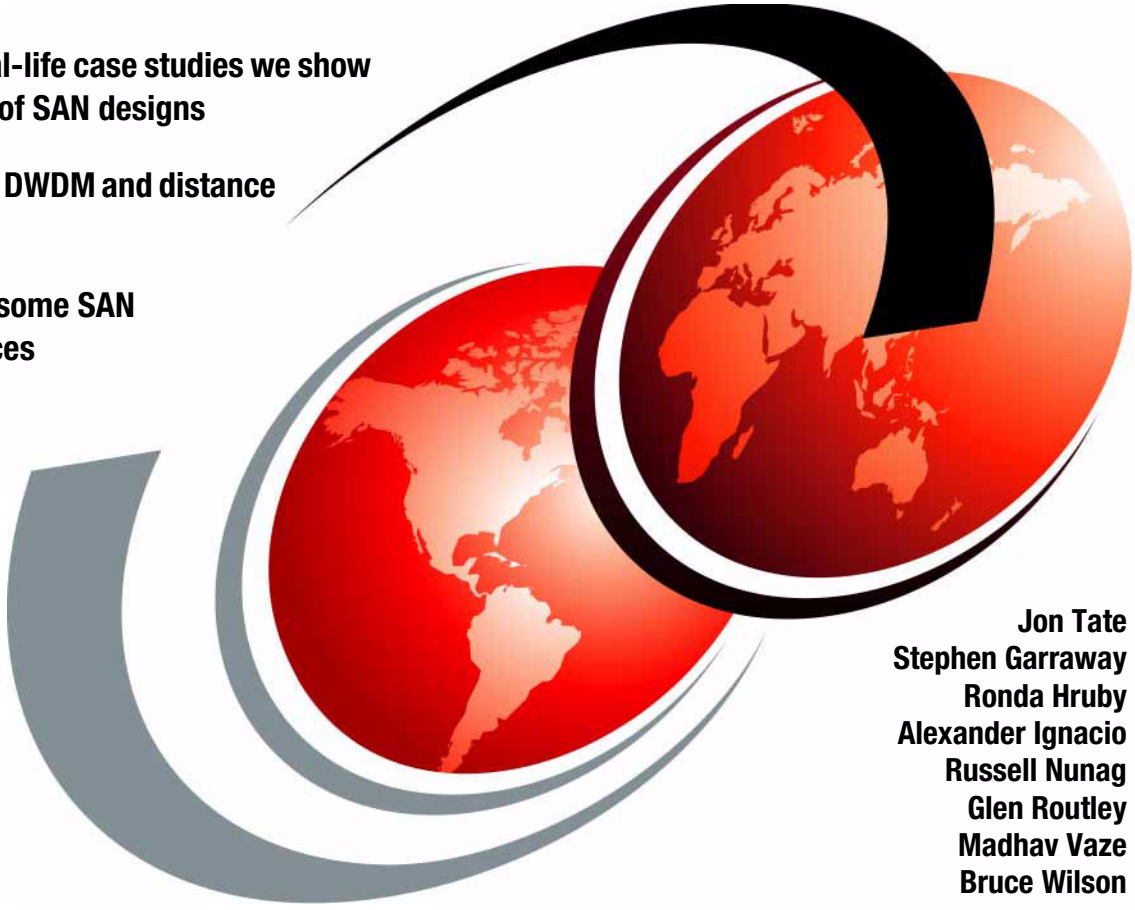
Draft Document for Review December 28, 2004 5:59 pm SG24-6384-00

IBM TotalStorage: SAN Product, Design and Optimization Guide

- Utilizing real-life case studies we show a selection of SAN designs
- Understand DWDM and distance solutions
- Introduces some SAN best practices



Jon Tate
Stephen Garraway
Ronda Hruby
Alexander Ignacio
Russell Nunag
Glen Routley
Madhav Vaze
Bruce Wilson

Redbooks

ibm.com/redbooks

International Technical Support Organization

IBM TotalStorage SAN

December 2004

Note: Before using this information and the product it supports, read the information in “Notices” on page xliii.

First Edition (December 2004)

This edition applies to the products described within.

This document created or updated on December 28, 2004.

© Copyright International Business Machines Corporation 2004. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Figures xxvii

Tables xxxvii

Examples xli

Notices xliii
Trademarks xliv

Preface xlv
The team that wrote this redbook xlv
Become a published author xlvii
Comments welcome xlvii

Chapter 1. Introduction 1
1.1 Beyond disaster recovery 2
1.1.1 Whose responsibility is it? 3
1.1.2 The Internet brings increased risks 4
1.1.3 Planning for business continuity 5
1.2 Using a SAN for business continuance 6
1.2.1 SANs and business continuance 7
1.3 SAN business benefits 8
1.3.1 Storage consolidation and sharing of resources 8
1.3.2 Data sharing 10
1.3.3 Non-disruptive scalability for growth 11
1.3.4 Improved backup and recovery 11
1.3.5 High performance 13
1.3.6 High availability server clustering 13
1.3.7 Improved disaster tolerance 14
1.3.8 Allow selection of “best of breed” storage 14
1.3.9 Ease of data migration 14
1.3.10 Reduced total costs of ownership 15
1.3.11 Storage resources match e-business enterprise needs 15

Chapter 2. SAN fabric components 17
2.1 Fibre Channel technology components 18
2.1.1 Fibre Channel transmission rates 19
2.2 Fiber optic interconnects 19
2.2.1 Small Form Factor Pluggable Media 20

© Copyright IBM Corp. 2004. All rights reserved.

2.2.2 Gigabit Interface Converters . . . 23
2.2.3 Gigabit Link Modules . . . 26
2.2.4 Media Interface Adapters . . . 27
2.2.5 1x9 transceivers . . . 27
2.2.6 Fibre Channel adapter cable . . . 28
2.2.7 Host Bus Adapters . . . 28
2.2.8 Bridges and SAN Data Gateways . . . 30
2.2.9 Arbitrated Loop hubs . . . 30
2.2.10 Switched hubs . . . 31
2.2.11 Switches . . . 31
2.2.12 Core switches . . . 33
2.2.13 Directors . . . 34
2.2.14 Director features . . . 35
2.2.15 Storage considered as legacy . . . 37
2.2.16 SerDes . . . 37
2.2.17 Backplane and blades . . . 37
2.2.18 Test gear . . . 39

Chapter 3. SAN features . . . 43
3.1 Fabric implementation . . . 44
3.1.1 Blocking . . . 45
3.1.2 Ports . . . 46
3.1.3 SAN topologies . . . 48
3.1.4 Point-to-point . . . 49
3.1.5 Arbitrated loop . . . 50
3.1.6 Switched fabric . . . 57
3.1.7 Inter Switch Links . . . 60
3.1.8 Adding new devices . . . 66
3.2 Classes of service . . . 67
3.2.1 Class 1 . . . 68
3.2.2 Class 2 . . . 68
3.2.3 Class 3 . . . 68
3.2.4 Class 4 . . . 69
3.2.5 Class 5 . . . 69
3.2.6 Class 6 . . . 69
3.2.7 Class F . . . 70
3.2.8 Communication . . . 70
3.3 Buffers . . . 70
3.4 Addressing . . . 74
3.4.1 World Wide Name . . . 74
3.4.2 WWN and WWPN . . . 75
3.4.3 24-bit Port address . . . 77
3.4.4 Loop address . . . 79

3.4.5 FICON addressing 80
3.5 Fabric services 85
3.5.1 Management services 85
3.5.2 Time services 85
3.5.3 Name services 85
3.5.4 Login services 85
3.5.5 Registered State Change Notification 86
3.6 Logins 86
3.6.1 Fabric login 86
3.6.2 LPort login 87
3.6.3 Process login 87
3.7 Path routing mechanisms 87
3.7.1 Spanning tree 87
3.7.2 Fabric Shortest Path First 89
3.7.3 What is FSPF? 89
3.7.4 How does FSPF work? 92
3.7.5 How does FSPF help? 92
3.7.6 What happens when there is more than one shortest path? 92
3.7.7 Can FSPF cause any problems? 94
3.7.8 FC-PH-2 and speed 96
3.7.9 1 Gb/s, 2 Gb/s and beyond 98
3.7.10 FC-PH, FC-PH-2, and FC-PH-3 99
3.7.11 Layers 101
3.8 Zoning 104
3.8.1 Hardware zoning 106
3.8.2 Software zoning 109
3.9 Trunking 112
3.9.1 Frame filtering 114
3.9.2 Oversubscription 114
3.9.3 Congestion 115
3.9.4 Information units 115
3.9.5 The movement of data 115
3.9.6 Data encoding 116
3.10 Ordered set, frames, sequences, and exchanges 119
3.10.1 Ordered set 120
3.10.2 Frames 121
3.10.3 Sequences 121
3.10.4 Exchanges 121
3.10.5 Frames 122
3.10.6 “In order” and “out of order” 124
3.10.7 Latency 124
3.10.8 Heterogeneousness 125
3.10.9 Open Fiber Control: OFC or Non-OFC 125

3.11 Fibre Channel Arbitrated Loop (FC-AL) . . . 126
3.11.1 Loop protocols . . . 126
3.11.2 Fairness algorithm . . . 130
3.11.3 Loop addressing . . . 130
3.11.4 Private devices on NL_Ports . . . 131
3.12 Factors and considerations . . . 134
3.12.1 Limits . . . 134
3.12.2 Security . . . 135
3.12.3 Interoperability . . . 136
3.13 Standards . . . 136
3.13.1 SAN industry associations and organizations . . . 137
3.13.2 Storage Networking Industry Association . . . 139
3.13.3 Fibre Channel Industry Association . . . 139
3.13.4 The SCSI Trade Association . . . 140
3.13.5 InfiniBand (SM) Trade Association . . . 140
3.13.6 National Storage Industry Consortium . . . 140
3.13.7 Internet Engineering Task Force . . . 141
3.13.8 American National Standards Institute . . . 141
3.13.9 INCITS . . . 141
3.13.10 List of evolved Fibre Channel standards . . . 141
3.14 SAN software management standards . . . 145
3.14.1 Application management . . . 147
3.14.2 Data management . . . 148
3.14.3 Resource management . . . 149
3.14.4 Network management . . . 149
3.14.5 Device Management . . . 151
3.14.6 Fabric management methods . . . 153
3.14.7 Common methods . . . 153
3.14.8 The SNIA Shared Storage Model . . . 164
3.14.9 Long distance links . . . 165
3.14.10 Backup windows . . . 165
3.14.11 Restore/disaster recovery time . . . 166
3.15 zSeries and S/390 . . . 166
3.15.1 pSeries . . . 167
3.15.2 xSeries . . . 167
3.15.3 iSeries . . . 168
3.16 Security . . . 168
3.16.1 Control types . . . 169
3.16.2 Vulnerabilities . . . 169
3.16.3 Fibre Channel security . . . 171
3.16.4 Security mechanisms . . . 172
3.16.5 IP security . . . 174
3.16.6 Virtualization . . . 174

3.17 Solutions . . . 176
3.18 Emerging technologies . . . 179
3.19 iSCSI . . . 179
3.20 iFCP . . . 180
3.21 FCIP . . . 180

Chapter 4. Extending fiber over distance with DWDM . . . 183
4.1 Why we need fiber saving devices . . . 184
4.1.1 Storage Area Network . . . 184
4.1.2 Metropolitan Area Network . . . 185
4.1.3 Wide Area Networks . . . 185
4.2 Fiber extending device concepts . . . 186
4.2.1 Channel extenders . . . 186
4.2.2 Multiplexers . . . 187
4.2.3 Multiplexers and demultiplexers . . . 192
4.2.4 Optical add/drop multiplexers . . . 193
4.2.5 Optical amplifiers . . . 195
4.2.6 Regenerative repeaters . . . 195
4.3 DWDM topologies . . . 195
4.3.1 Point-to-point . . . 196
4.3.2 Linear . . . 196
4.3.3 Ring . . . 197
4.4 Factors that will affect distance . . . 200
4.4.1 Light or link budget . . . 201
4.4.2 Buffer credits . . . 202
4.4.3 Quality of fiber . . . 203
4.4.4 Cable types . . . 203
4.4.5 Droop . . . 205
4.4.6 Latency . . . 206
4.4.7 Sizing . . . 206
4.4.8 Hops . . . 206
4.4.9 Physical location of repeaters . . . 207
4.4.10 Standards . . . 207
4.4.11 Terminology . . . 207
4.4.12 Protocol definitions . . . 209

Chapter 5. SAN disciplines . . . 211
5.1 Floor plan . . . 212
5.1.1 SAN inventory . . . 212
5.1.2 Cable types and cable routing . . . 213
5.1.3 Planning considerations and recommendations . . . 218
5.1.4 Structured cabling . . . 219
5.1.5 Data center fiber cabling options . . . 220

5.1.6 Cabinets . . . 222
5.1.7 Phone sockets . . . 224
5.1.8 Environmental considerations . . . 224
5.1.9 Location . . . 224
5.1.10 Sequence for design . . . 225
5.2 Naming conventions . . . 227
5.2.1 Servers . . . 227
5.2.2 Cabinets . . . 227
5.2.3 SAN fabric components . . . 227
5.2.4 Cables . . . 229
5.2.5 Zones . . . 229
5.3 Documentation . . . 230
5.4 Power-on sequence . . . 231
5.5 Security . . . 231
5.5.1 General . . . 231
5.5.2 Physical access . . . 232
5.5.3 Remote access . . . 233
5.6 Education . . . 234
5.6.1 SAN administrators . . . 235
5.6.2 Skills . . . 235
5.6.3 Certification . . . 235

Chapter 6. Host Bus Adapters . . . 239
6.1 Selection criterion . . . 240
6.1.1 IBM supported HBAs . . . 240
6.1.2 Special features . . . 240
6.1.3 Quantity of servers . . . 240
6.1.4 HBA parameter settings . . . 241

Chapter 7. SAN design considerations . . . 243
7.1 What do you want to achieve with a SAN? . . . 244
7.2 Existing resources needs and planned growth . . . 247
7.2.1 Collecting the data about existing resources . . . 247
7.2.2 Planning for future needs . . . 248
7.2.3 Platforms and storage . . . 249
7.3 Select the core design for your environment . . . 250
7.3.1 Selecting the topology . . . 250
7.3.2 Scalability . . . 251
7.3.3 Performance . . . 251
7.3.4 Redundancy and resiliency . . . 253
7.4 Host connectivity and Host Bus Adapters . . . 256
7.4.1 Selection criterion . . . 257
7.4.2 Multipathing software . . . 258

7.4.3 Storage sizing . . . 260
7.4.4 Management software . . . 261
7.5 Director class or switch technology . . . 261
7.6 General considerations . . . 280
7.6.1 Ports and ASICs . . . 280
7.6.2 Class F . . . 280
7.6.3 Domain IDs . . . 280
7.6.4 Zoning . . . 281
7.6.5 Physical infrastructure and distance . . . 281
7.7 Interoperability issues in the design . . . 282
7.7.1 Certification and support . . . 284
7.7.2 OEM/IBM mixes . . . 284
7.8 Pilot and test the design . . . 285

Chapter 8. IBM TotalStorage SAN b-type family . . . 287
8.1 Product description . . . 288
8.1.1 2005-B32 . . . 288
8.1.2 2005-H08 and 2005-H16 . . . 289
8.1.3 2109-F16 and 2109-F32 . . . 291
8.1.4 2109-M12 . . . 294
8.1.5 IBM TotalStorage SAN Switch M14 . . . 300
8.2 Features . . . 306
8.3 Advanced Security . . . 307
8.3.1 Fabric configuration servers . . . 309
8.3.2 Management access controls . . . 309
8.3.3 Device connection controls . . . 310
8.3.4 Switch connection controls . . . 310
8.3.5 Fibre Channel Authentication Protocol . . . 310
8.4 ISL . . . 311
8.4.1 ISLs without trunking . . . 312
8.4.2 ISLs with trunking . . . 313
8.4.3 Switch count . . . 315
8.4.4 Distributed fabrics . . . 315
8.5 FICON . . . 317
8.6 Fabric management . . . 318
8.6.1 Web Tools . . . 319
8.6.2 Advanced Performance Monitoring . . . 321
8.6.3 Fabric Watch . . . 322
8.6.4 Fabric Manager . . . 325
8.6.5 SCSI Enclosure Services . . . 327
8.7 Zoning . . . 328
8.7.1 Preparing to use zoning . . . 330
8.7.2 Increasing availability . . . 331

8.7.3 Advanced zone terminology . . . 331
8.7.4 Zoning types . . . 333
8.7.5 Zone configuration . . . 335
8.7.6 Zoning administration . . . 336
8.7.7 QuickLoop . . . 337
8.8 Switch interoperability . . . 339
8.9 Feature codes . . . 340

Chapter 9. IBM TotalStorage SAN n-type family and replication appliance. 345
9.1 CNT UltraNet family product description . . . 347
9.1.1 CNT Ultranet Edge Storage Router . . . 347
9.2 CNT Ultranet Replication Appliance . . . 349
9.3 IBM TotalStorage SAN256N director . . . 354
9.3.1 Ultranet Multi-service Director Components . . . 356
9.3.2 Ultranet Multi-service director supported attachments . . . 363
9.3.3 SAN256N Multiple Services . . . 363
9.4 FC/9000 Director product description . . . 365
9.4.1 FC/9000-64 Fibre Channel Director . . . 366
9.4.2 FC/9000-128 Fibre Channel Director . . . 367
9.4.3 FC/9000-256 Fibre Channel Director . . . 368
9.4.4 FC/9000 director supported attachments . . . 369
9.4.5 FC/9000 supported port types . . . 370
9.4.6 Availability . . . 370
9.4.7 Scalable capacity CNT director high availability option . . . 371
9.5 FC/9000 components . . . 372
9.5.1 FC/9000 support the following components: . . . 372
9.5.2 CNT director: 1 Gb/s and 128 port capable . . . 379
9.5.3 Upgrade Path/Extensible Core Architecture . . . 385
9.5.4 Cabinet . . . 386
9.6 Features . . . 386
9.6.1 Ultranet Multi-Service Director High Availability . . . 386
9.6.2 FC/9000 availability options . . . 387
9.6.3 Logical Domains . . . 388
9.6.4 Trace and mirror ports . . . 389
9.6.5 Management software inVSN Enterprise Manager . . . 390
9.6.6 inVSN Storage Network Manager . . . 392
9.6.7 FICON . . . 392
9.6.8 Intermix mode (FICON and FCP) . . . 394
9.6.9 Getting started with the FC/9000 . . . 395
9.6.10 Setup validation . . . 396
9.6.11 Dial home settings . . . 397
9.6.12 inVSN Enterprise Manager installation . . . 397

9.6.13 Installing . . . 398
9.6.14 Management software . . . 398
9.6.15 inVSN Enterprise Manager database . . . 401
9.6.16 .Auditing and event logging . . . 402
9.6.17 Using the inVSN event log . . . 404
9.7 Zoning . . . 405
9.8 Zoning . . . 406
9.8.1 Understanding CNT name server zoning . . . 406
9.8.2 Understanding hardware enforced WWN zoning . . . 409
9.8.3 Understanding CNT broadcast zoning . . . 410
9.8.4 Managing by device WWN . . . 410
9.8.5 Cascading of directors . . . 410
9.9 Configuration guidelines . . . 413
9.9.1 Port positioning for performance . . . 414
9.9.2 Port configuration flexibility . . . 415
9.9.3 Hop count . . . 416
9.9.4 FC/9000 port tuning options . . . 417
9.9.5 Interoperability and legacy device toleration . . . 418
9.10 Private and public device support . . . 418
9.10.1 24-bit and 8-bit addressing . . . 419
9.10.2 Private loop . . . 419
9.10.3 Public loop . . . 419
9.10.4 Translative loop mode . . . 420
9.10.5 Firmware level . . . 421
9.10.6 Trivial File Transfer Protocol . . . 421
9.11 Power-up process . . . 422
9.12 Interoperability . . . 424
9.12.1 Standards . . . 424
9.13 Feature codes . . . 425
9.13.1 Ultranet Multi-Service Director . . . 425
9.13.2 FC/9000 Fibre Channel Director . . . 425

Chapter 10. IBM TotalStorage SAN m-type family . . . 429
10.1 Product description . . . 431
10.1.1 Machine type and model number changes . . . 431
10.1.2 McDATA Sphereon 4300 Fabric Switch . . . 431
10.1.3 McDATA Sphereon 4500 Fabric Switch . . . 433
10.1.4 McDATA Sphereon 3232 Fabric Switch . . . 435
10.1.5 McDATA Intrepid 6140 Director . . . 438
10.1.6 McDATA Intrepid 6064 Director . . . 446
10.2 Fabric planning . . . 453
10.2.1 Dual fabrics and directors . . . 453
10.2.2 Server to storage ratio . . . 453

10.2.3 ISLs . . . 454
10.2.4 Load balancing . . . 454
10.2.5 Principal switch selection . . . 455
10.2.6 Special considerations . . . 457
10.2.7 Open Fabric . . . 458
10.2.8 Supported devices, servers and HBAs . . . 458
10.3 Features of directors and switches . . . 459
10.3.1 Element Manager . . . 459
10.3.2 FICON Management Server . . . 459
10.3.3 Full Volatility Option . . . 459
10.3.4 Open Systems Management Server . . . 460
10.3.5 Open Trunking . . . 460
10.3.6 Preferred Path . . . 463
10.3.7 SANtegrity Binding . . . 463
10.3.8 Feature activation . . . 464
10.4 FICON support . . . 465
10.5 Fabric management . . . 465
10.5.1 In-band management . . . 465
10.5.2 Out-of-band management . . . 466
10.5.3 EFC Server . . . 467
10.5.4 EFC Manager . . . 472
10.5.5 Troubleshooting . . . 474
10.5.6 SANpilot interface . . . 475
10.5.7 Command line interface . . . 476
10.5.8 SNMP . . . 477
10.6 Zoning . . . 477
10.6.1 Configuring zones . . . 478
10.6.2 Zoning and LUN masking . . . 480
10.6.3 Persistent binding . . . 481
10.6.4 Blocking a port . . . 481
10.6.5 Merging fabrics . . . 481
10.7 Performance . . . 482
10.8 Security . . . 483
10.8.1 Restricting access to those that need it . . . 483
10.8.2 Controlling access at the switch . . . 484
10.8.3 SANtegrity Security Suite . . . 484
10.9 SANavigator . . . 484
10.10 Licensing . . . 485
10.10.1 Warranties . . . 485
10.11 IBM feature codes . . . 485

Chapter 11. Cisco switches and directors . . . 489
11.1 Product description . . . 490

11.1.1 MDS 9120 and 9140 Multilayer Switches 490
11.1.2 MDS 9216 Multilayer Switch 491
11.1.3 MDS 9506 Multilayer Director 494
11.1.4 MDS 9509 Multilayer Director 495
11.2 Features 498
11.2.1 Supported attachments 498
11.2.2 Port addressing and port modes 499
11.2.3 Fibre Channel IDs and Persistent FC_ID 499
11.2.4 Supported port types 500
11.3 Supervisor module 503
11.3.1 Control and management 503
11.3.2 Optional modules 505
11.4 MDS 9000 SAN-OS 510
11.4.1 Licensed feature packages 511
11.4.2 PortChannels 514
11.4.3 Trunking 516
11.4.4 Quality of Service 517
11.4.5 Congestion control 517
11.4.6 IP services 519
11.4.7 Call home 519
11.4.8 Virtual SAN 519
11.5 Security management 527
11.5.1 Switch access security 527
11.5.2 User authentication 527
11.6 Troubleshooting features 530
11.6.1 Troubleshooting with Fabric Manager 530
11.6.2 Monitoring network traffic using SPAN 531
11.6.3 Monitoring traffic using Fibre Channel analyzers 536
11.7 FICON 539
11.8 Fabric management 540
11.8.1 Cisco MDS 9000 Fabric Manager 540
11.8.2 In-band management and out-of-band management 541
11.8.3 Using the setup routine 542
11.8.4 Controlling administrator access with users and roles 543
11.8.5 Accessing Cisco Fabric Manager 543
11.8.6 Connecting to a supervisor module 544
11.9 Zoning 544
11.9.1 Zone features 545
11.9.2 Zone membership 546
11.9.3 Configuring a zone 546
11.9.4 Zone enforcement 546
11.9.5 Zone sets 547
11.9.6 Default zone 547

11.10 Switch interoperability mode 548
11.10.1 Interoperability matrix 549
11.11 IBM Feature Codes 549

Chapter 12. SAN event data gathering tips 553
12.1 Overview 554
12.2 Hosts 554
12.2.1 AIX 554
12.2.2 HP-UX 555
12.2.3 Linux 556
12.2.4 . . . . . Microsoft Windows557
12.2.5 Novell Netware 558
12.2.6 SUN Solaris 558
12.3 Switches 559
12.3.1 SAN Switch 2031/2032 (McDATA) 559
12.3.2 SAN Switch 2062 (Cisco) 560
12.3.3 SAN Switch 2109 (Brocade) 561
12.3.4 SAN Switch 2042 (CNT) 562
12.4 Storage 562
12.4.1 IBM TotalStorage DS Family disk subsystem 562
12.4.2 IBM TotalStorage Enterprise Storage Server 563
12.4.3 3583 Tape Library and SDGM 564

Chapter 13. General solutions 567
13.1 Objectives of SAN implementation 568
13.2 Servers and host bus adapters 568
13.2.1 Path and dual-redundant HBA 569
13.2.2 Multiple paths 569
13.3 Software 570
13.4 Storage 570
13.5 Fabric 572
13.5.1 The fabric-is-a-switch approach 572
13.5.2 The fabric-is-a-network approach 574
13.6 High level fabric design 574
13.7 Definitions 578
13.7.1 Port formulas 580
13.8 Our solutions 581

Chapter 14. IBM TotalStorage SAN b-type family solutions 583
14.1 Performance solutions 584
14.2 Availability solutions 588
14.2.1 Single fabric 589
14.2.2 Dual fabric 592
14.3 Clustering solutions 594

14.3.1 Two-node clustering 594
14.3.2 Multi-node clustering 597
14.4 Secure solutions 600
14.5 Loop solutions 602

Chapter 15. IBM TotalStorage SAN m-type family solutions 607
15.1 Performance solutions 608
15.2 Availability solutions 612
15.2.1 Dual fabric 613
15.2.2 Dual sites 615
15.3 Clustering solutions 618
15.4 Secure solutions 622
15.5 Loop solutions 626

Chapter 16. IBM TotalStorage SAN n-type family solutions 631
16.1 Performance solutions 632
16.2 Availability solutions 636
16.2.1 Dual fabric 636
16.2.2 Dual fabrics over distance 640
16.3 Clustering solutions 644
16.4 Secure solutions 649
16.4.1 Logical Domain security solution 649
16.4.2 Soft zoning security solution 652
16.4.3 Heterogeneous zoning solution 654

Chapter 17. Cisco solutions 659
17.1 Performance solutions 660
17.2 Availability solutions 665
17.2.1 Dual fabric 665
17.2.2 Dual sites 668
17.3 Clustering solutions 672
17.3.1 Two-node clustering 673
17.3.2 Multi-node clustering 675
17.4 Secure solutions 678
17.4.1 Zoning security solution 678
17.5 Loop solutions 682
17.5.1 Utilizing the translative loop port 682

Chapter 18. Case studies 685
18.1 Case Study 1: Company One 686
18.1.1 Company profile 686
18.1.2 High-level business requirement(s) 686
18.1.3 Current infrastructure 686
18.1.4 Detailed requirements 686

18.1.5 Analysis (ports and throughput) . . . 687
18.2 Case Study 2: Company Two . . . 689
18.2.1 Company profile . . . 689
18.2.2 High-level business requirement(s) . . . 689
18.2.3 Current infrastructure . . . 689
18.2.4 Detailed requirements . . . 691
18.2.5 Analysis (ports and throughput) . . . 692
18.3 Case Study 3: Company Three . . . 697
18.3.1 Company profile . . . 697
18.3.2 High-level business requirement(s) . . . 697
18.3.3 Current infrastructure . . . 698
18.3.4 Detailed requirements . . . 698
18.3.5 Analysis (ports and throughput) . . . 699
18.4 Case Study 4: Company Four . . . 700
18.4.1 Company profile . . . 700
18.4.2 High-level business requirement(s) . . . 700
18.4.3 Current infrastructure . . . 701
18.4.4 Detailed requirements . . . 702
18.4.5 Analysis (ports and throughput) . . . 703
18.5 Case Study 5: Company Five . . . 705
18.5.1 Company profile . . . 705
18.5.2 High-level business requirement(s) . . . 705
18.5.3 Current infrastructure . . . 706
18.5.4 Detailed requirements . . . 707
18.5.5 Analysis (ports and throughput) . . . 708
18.6 Case Study 6: Company Six . . . 710
18.6.1 Company profile . . . 710
18.6.2 High-level business requirement(s) . . . 710
18.6.3 Current infrastructure . . . 710
18.6.4 Detailed requirements . . . 711
18.6.5 Analysis (ports and throughput) . . . 712

Chapter 19. IBM TotalStorage SAN b-type family solutions . . . 715
19.1 Case Study 1: Company One . . . 716
19.1.1 Switch design . . . 716
19.1.2 Performance . . . 723
19.1.3 Availability . . . 723
19.1.4 Security . . . 723
19.1.5 Distance . . . 724
19.1.6 Scalability . . . 724
19.1.7 “What if” failure scenarios . . . 724
19.1.8 Manageability and management software . . . 725
19.1.9 Core switch design . . . 726

19.2 Case Study 2: Company Two 729
19.2.1 Design 729
19.2.2 Performance 732
19.2.3 Availability 735
19.2.4 Security 735
19.2.5 Distance 736
19.2.6 Scalability 737
19.2.7 “What if” failure scenarios 737
19.2.8 Manageability and management software 737
19.3 Case Study 3: Company Three 740
19.3.1 Design 740
19.3.2 Performance 742
19.3.3 Availability 742
19.3.4 Security 743
19.3.5 Distance 743
19.3.6 Scalability 743
19.3.7 “What if” failure scenarios 743
19.3.8 Manageability and management software 744
19.4 .Case Study 4 - Company Four 744
19.4.1 Design 744
19.4.2 Performance 746
19.4.3 Availability 746
19.4.4 Security 746
19.4.5 Distance 747
19.4.6 Scalability 747
19.4.7 “What if” failure scenarios 747
19.4.8 Manageability and management software 748
19.5 Case Study 5: Company Five 748
19.5.1 Design 748
19.5.2 Performance 749
19.5.3 Availability 750
19.5.4 Security 750
19.5.5 Distance 751
19.5.6 Scalability 751
19.5.7 “What if” failure scenarios 751
19.5.8 Manageability and management software 752
19.6 Case Study 6: Company Six 752
19.6.1 Design 752
19.6.2 Performance 755
19.6.3 Availability 755
19.6.4 Security 755
19.6.5 Distance 756
19.6.6 Scalability 756

19.6.7 “What if” failure scenarios . . . 756
19.6.8 Manageability and management software . . . 757

Chapter 20. IBM TotalStorage SAN m-type family solutions . . . 759
20.1 Case Study 1: Company One . . . 760
20.1.1 Design using Directors . . . 760
20.1.2 Performance . . . 763
20.1.3 Availability . . . 764
20.1.4 Security . . . 764
20.1.5 Distance . . . 764
20.1.6 Scalability . . . 764
20.1.7 “What if” failure scenarios . . . 765
20.1.8 Manageability and management software . . . 765
20.1.9 Design using switches . . . 766
20.1.10 Performance . . . 770
20.1.11 Availability . . . 770
20.1.12 Security . . . 771
20.1.13 Distance . . . 771
20.1.14 Scalability . . . 771
20.1.15 “What if” failure scenarios . . . 771
20.1.16 Manageability and management software . . . 772
20.2 Case Study 2: Company Two . . . 773
20.2.1 Design . . . 773
20.2.2 Performance . . . 776
20.2.3 Availability . . . 777
20.2.4 Security . . . 777
20.2.5 Distance . . . 778
20.2.6 Scalability . . . 778
20.2.7 “What if” failure scenarios . . . 778
20.2.8 Manageability and management software . . . 779
20.3 Case Study 3: Company Three . . . 780
20.3.1 Design . . . 780
20.3.2 Performance . . . 782
20.3.3 Availability . . . 782
20.3.4 Security . . . 782
20.3.5 Distance . . . 783
20.3.6 Scalability . . . 783
20.3.7 “What if” failure scenarios . . . 783
20.3.8 Manageability and management software . . . 784
20.4 Case Study 4 - Company Four . . . 785
20.4.1 Design . . . 785
20.4.2 Performance . . . 787
20.4.3 Availability . . . 787

20.4.4 Security . . . 787
20.4.5 Distance . . . 788
20.4.6 Scalability . . . 788
20.4.7 “What if” failure scenarios . . . 788
20.4.8 Manageability and management software . . . 788
20.5 Case Study 5 - Company Five . . . 790
20.5.1 Design . . . 790
20.5.2 Performance . . . 791
20.5.3 Availability . . . 791
20.5.4 Security . . . 792
20.5.5 Distance . . . 792
20.5.6 Scalability . . . 792
20.5.7 “What if” failure scenarios . . . 792
20.5.8 Manageability and management software . . . 793
20.6 Case Study 6: Company Six . . . 794
20.6.1 Design . . . 794
20.6.2 Performance . . . 797
20.6.3 Availability . . . 797
20.6.4 Security . . . 798
20.6.5 Distance . . . 798
20.6.6 Scalability . . . 798
20.6.7 “What if” failure scenarios . . . 798
20.6.8 Manageability and management software . . . 799

Chapter 21. IBM TotalStorage SAN n-type family solutions . . . 801
21.1 Case Study 1: Company One . . . 802
21.1.1 Design . . . 802
21.1.2 Performance . . . 808
21.1.3 Availability . . . 808
21.1.4 Security . . . 808
21.1.5 Distance . . . 809
21.1.6 Scalability . . . 809
21.1.7 “What if” failure scenarios . . . 809
21.1.8 Manageability and management software . . . 810
21.2 Case Study 2: Company Two . . . 810
21.2.1 Design . . . 810
21.2.2 Performance . . . 815
21.2.3 Availability . . . 815
21.2.4 Security . . . 815
21.2.5 Distance . . . 816
21.2.6 Scalability . . . 816
21.2.7 “What if” failure scenarios . . . 817
21.2.8 Manageability and management software . . . 817

21.3 Case Study 3: Company Three 817
21.3.1 Design 817
21.3.2 Performance 819
21.3.3 Availability 819
21.3.4 Security 819
21.3.5 Distance 819
21.3.6 Scalability 820
21.3.7 “What if” failure scenarios. 820
21.3.8 Manageability and management software 820
21.4 Case Study 4: Company Four 821
21.4.1 Design 821
21.4.2 Performance 824
21.4.3 Availability 825
21.4.4 Security 825
21.4.5 Distance 825
21.4.6 Scalability 825
21.4.7 “What if” failure scenarios 825
21.4.8 Manageability and management software 826
21.5 Case Study 5: Company Five. 826
21.5.1 Design 826
21.5.2 Performance 828
21.5.3 Availability 828
21.5.4 Security 828
21.5.5 Distance 829
21.5.6 Scalability 829
21.5.7 “What if” failure scenarios 829
21.5.8 Manageability and management software 830
21.6 Case Study 6: Company Six 830
21.6.1 Design 830
21.6.2 Performance 833
21.6.3 Availability 833
21.6.4 Security 833
21.6.5 Distance 834
21.6.6 Scalability 834
21.6.7 “What if” failure scenarios 835
21.6.8 Manageability and management software 835

Chapter 22. Cisco solutions 837
22.1 Case Study 1: Company One 838
22.1.1 Design using directors. 838
22.1.2 Performance 842
22.1.3 Availability 842
22.1.4 Security 842

22.1.5 Distance . . . 843
22.1.6 Scalability . . . 843
22.1.7 “What if” failure scenarios . . . 843
22.1.8 Manageability and management software . . . 844
22.1.9 Design using switches . . . 845
22.1.10 Performance . . . 849
22.1.11 Availability . . . 849
22.1.12 Security . . . 850
22.1.13 Distance . . . 850
22.1.14 Scalability . . . 850
22.1.15 “What if” failure scenarios . . . 850
22.1.16 Manageability and management software . . . 851
22.2 Case Study 2: Company Two . . . 852
22.2.1 Design . . . 852
22.2.2 Performance . . . 855
22.2.3 Availability . . . 856
22.2.4 Security . . . 856
22.2.5 Distance . . . 857
22.2.6 Scalability . . . 857
22.2.7 “What if” failure scenarios . . . 857
22.2.8 Manageability and management software . . . 858
22.3 Case Study 3: Company Three . . . 858
22.3.1 Design . . . 858
22.3.2 Performance . . . 860
22.3.3 Availability . . . 860
22.3.4 Security . . . 860
22.3.5 Distance . . . 861
22.3.6 Scalability . . . 861
22.3.7 “What if” failure scenarios . . . 861
22.3.8 Manageability and management software . . . 861
22.4 Case Study 4 - Company Four . . . 862
22.4.1 Design . . . 863
22.4.2 Performance . . . 864
22.4.3 Availability . . . 865
22.4.4 Security . . . 865
22.4.5 Distance . . . 865
22.4.6 Scalability . . . 865
22.4.7 “What if” failure scenarios . . . 865
22.4.8 Manageability and management software . . . 866
22.5 Case Study 5 - Company Five . . . 867
22.5.1 Design . . . 867
22.5.2 Performance . . . 869
22.5.3 Availability . . . 869

22.5.4 Security . . . 869
22.5.5 Distance . . . 870
22.5.6 Scalability . . . 870
22.5.7 “What if” failure scenarios . . . 870
22.5.8 Manageability and management software . . . 870
22.6 Case Study 6: Company Six . . . 871
22.6.1 Design . . . 871
22.6.2 Performance . . . 875
22.6.3 Availability . . . 875
22.6.4 Security . . . 876
22.6.5 Distance . . . 876
22.6.6 Scalability . . . 876
22.6.7 “What if” failure scenarios . . . 876
22.6.8 Manageability and management software . . . 877

Chapter 23. IBM TotalStorage SAN b-type family distance solutions . . 879
23.1 High-level availability objectives . . . 881
23.2 Storage consolidation with remote disk . . . 881
23.3 Two sites up to 10 km apart . . . 886
23.4 Two sites up to 100 km apart . . . 891
23.5 Two sites - Point-to-point DWDM solution . . . 893
23.6 Two sites: Point-to-point DWDM with ESS PPRC . . . 897
23.7 Multiple site ring DWDM solution . . . 898
23.8 Two sites: Channel extender and WAN extension . . . 901
23.9 Two sites with FCIP connection . . . 903
23.10 Remote tape vaulting . . . 905
23.11 Remote tape vaulting with disaster tolerance . . . 908

Chapter 24. IBM TotalStorage SAN m-type family distance solutions . . 909
24.1 High-level availability objectives . . . 910
24.2 Storage consolidation with remote disk . . . 911
24.3 Two sites up to 10 km apart . . . 915
24.4 Two sites up to 100 km apart . . . 920
24.5 Two sites - Point-to-point DWDM solution . . . 922
24.6 Two sites: Point-to-point DWDM with ESS PPRC . . . 926
24.7 Multiple site ring DWDM solution . . . 927
24.8 Two sites: Channel extender and WAN extension . . . 930
24.9 Two sites with iFCP connection . . . 932
24.10 Remote tape vaulting . . . 934
24.11 Remote tape vaulting with disaster tolerance . . . 937

Chapter 25. IBM TotalStorage SAN n-type family distance solutions . . 939
25.1 High-level availability objectives . . . 941
25.2 Storage consolidation with remote disk . . . 941

25.3 Two sites up to 10 km apart . . . 945
25.4 Two sites up to 100 km apart . . . 950
25.5 Two sites - Point-to-point DWDM solution . . . 952
25.6 Two sites: Point-to-point DWDM with ESS PPRC . . . 956
25.7 Multiple site ring DWDM solution . . . 957
25.8 Two sites: Channel extender and WAN extension . . . 960
25.9 Two sites with FCIP connection . . . 962
25.10 Remote tape vaulting . . . 964
25.11 Remote tape vaulting with disaster tolerance . . . 967

Chapter 26. Cisco distance solutions . . . 969
26.1 High-level availability objectives . . . 970
26.2 Storage consolidation with remote disk . . . 971
26.3 Two sites up to 10 km apart . . . 975
26.4 Two sites up to 100 km apart . . . 980
26.5 Two sites - Point-to-point DWDM solution . . . 982
26.6 Two sites: Point-to-point DWDM with ESS PPRC . . . 986
26.7 Multiple site ring DWDM solution . . . 987
26.8 Two sites: Channel extender and WAN extension . . . 990
26.9 Two sites with FCIP connection . . . 992
26.10 Remote tape vaulting . . . 994
26.11 Remote tape vaulting with disaster tolerance . . . 997

Chapter 27. SAN best practices . . . 999
27.1 Scaling . . . 1000
27.1.1 How to scale easily . . . 1000
27.1.2 How to avoid downtime . . . 1000
27.1.3 Adding a switch/director . . . 1001
27.1.4 Adding ISLs . . . 1002
27.1.5 Performance monitoring and reporting . . . 1002
27.2 Know your workloads . . . 1002
27.3 Port placement . . . 1003
27.4 WWNs . . . 1006
27.5 Tools . . . 1007
27.6 Documentation . . . 1009
27.7 Configurations . . . 1009
27.8 Common practice faults . . . 1010
27.9 Zoning . . . 1010
27.10 Powering a SAN up or down . . . 1013

Glossary . . . 1015

Related publications . . . 1039
IBM Redbooks . . . 1039

Other resources 1039
Referenced Web sites 1040
How to get IBM Redbooks 1041
IBM Redbooks collections 1041

Index 1043

Figures

1-1 Business outage causes . . . 6
1-2 Storage consolidation . . . 9
1-3 Logical storage consolidation . . . 10
1-4 Loading the IP network . . . 12
1-5 SAN total storage solutions . . . 16
2-1 SFP Hot Pluggable optical transciever . . . 20
2-2 Small Form Fixed pin-through-hole Transceiver . . . 21
2-3 SFF hot-pluggable transceiver (SFP) with LC connector fiber cable . . . 22
2-4 Dual SC fiber-optic plug connector . . . 25
2-5 Gigabit Interface Converter . . . 25
2-6 Gigabit Link Module . . . 26
2-7 Media Interface Adapter . . . 27
2-8 1x9 transceivers . . . 27
2-9 Fibre Channel adapter cable . . . 28
2-10 Host Bus Adapter . . . 29
2-11 Fibre Channel Switch Environment . . . 33
2-12 Meshed topology switched fabric . . . 36
2-13 A diagram of a backplane and blades architecture . . . 38
2-14 Connecting an FC analyzer . . . 41
3-1 Cascading directors . . . 44
3-2 Non-blocking and blocking switching . . . 45
3-3 Fibre Channel port types . . . 48
3-4 Point-to-point . . . 49
3-5 Arbitrated loop . . . 50
3-6 Private loop implementation . . . 53
3-7 Public loop implementation . . . 54
3-8 Sample switched fabric configuration . . . 58
3-9 Cascading in a switched fabric . . . 60
3-10 Parallel ISLs - low traffic . . . 61
3-11 Parallel ISLs - high traffic . . . 61
3-12 ISL Trunking . . . 62
3-13 Four-switch fabric . . . 64
3-14 Adjacent FC devices . . . 72
3-15 World Wide Name addressing scheme . . . 75
3-16 WWN and WWPN . . . 76
3-17 WWN and WWPN entries in a name server table . . . 77
3-18 Fabric port address . . . 78
3-19 Ficon port addressing . . . 81

© Copyright IBM Corp. 2004. All rights reserved.

3-20 FICON single switch - switched point-to-point link address 82
3-21 ficon addressing for cascaded directors 83
3-22 Two cascaded director ficon addressing 84
3-23 Fabric shortest path first 90
3-24 FSPF calculates the route taking the least hops 91
3-25 Other possible paths 91
3-26 FSPF and round robin 93
3-27 FSPF can lead to oversubscription and congestion. 94
3-28 Hops and their cost, speed 95
3-29 Mixing 2 Gb/s and 1 Gb/s 99
3-30 Fibre Channel layers 102
3-31 Zoning 105
3-32 An example of zoning 106
3-33 Zoning based on the switch port-number. 107
3-34 Hardware zoning 108
3-35 Zoning based on the devices WWN. 110
3-36 Trunking 113
3-37 8b/10b encoding logic 117
3-38 Private loop implementation. 129
3-39 Public loop implementation 129
3-40 Arbitrated loop address translation 132
3-41 Groups involved in setting Storage Networking Standards 137
3-42 SAN management hierarchy 147
3-43 Common Interface Model for SAN management 148
3-44 Typical SAN environment 150
3-45 Device management elements 151
3-46 MIB tree 157
3-47 Vulnerabilities. 170
4-1 Time Division Multiplexer concepts 188
4-2 Wave Division Multiplexer concepts 189
4-3 Generic example of DWDM architecture 191
4-4 Multiplexer to demultiplexer 193
4-5 Both multiplexer and demultiplexer 193
4-6 Light dropped and added 194
4-7 Example of OADM using dielectric filter. 195
4-8 Point-to-point topology 196
4-9 Linear topology between three locations 196
4-10 Ring topology using two DWDM and two OADM. 197
4-11 Ring topology with three DWDM 198
4-12 DWDM module showing east and west 198
4-13 East and west must have same wavelengths within the same band . . 199
4-14 Light propagation through fiber 203
4-15 Light propagation in single-mode fiber. 204

4-16 Light propagation in multi-mode fiber . . . 204
4-17 ESCON droop example . . . 205
5-1 Mode differences through the fiber-optic cable . . . 214
5-2 Messy cabling, no cabinet, and no cable labels . . . 223
5-3 Naming convention development . . . 228
7-1 Single fabric non resilient . . . 253
7-2 Single fabric resilient . . . 254
7-3 Redundant fabric non resilient . . . 255
7-4 Redundant fabric resilient . . . 256
7-5 Multiple paths to the same LUN . . . 258
7-6 Multipath in single fabric SAN . . . 259
7-7 Director class or switch dilemma . . . 262
7-8 External managing of director class product . . . 264
7-9 Ports on different blades . . . 265
7-10 Routes in director class product . . . 265
7-11 Blade zoning . . . 266
7-12 Director class product versus full meshed switch fabric . . . 268
7-13 Director class 64 ports against 64 ports switch fabric . . . 269
7-14 Adding tapes to the director class SAN . . . 270
7-15 Switches with loop support . . . 271
7-16 ISL between two redundant fabrics . . . 272
7-17 Two director class products with FC-AL support solution . . . 273
7-18 Two director class products without FC-AL support . . . 274
7-19 Edge switch with only one connection . . . 275
7-20 One director class product without FC-AL support . . . 276
7-21 Two switch solution with FC-AL support . . . 277
7-22 Two switch solution without FC-AL support . . . 278
7-23 Single switch solution with FC-AL support . . . 279
8-1 IBM TotalStorage SAN Switch H08 . . . 290
8-2 IBM TotalStorage SAN Switch H16 . . . 290
8-3 IBM SAN Fibre Channel Switch 2109-F16 . . . 291
8-4 IBM SAN Fibre Channel Switch 2109-F32 . . . 293
8-5 2109-M12 . . . 295
8-6 Port side of the 2109-M12 . . . 296
8-7 CP (left) and Fibre Channel Module (right) . . . 297
8-8 Physical port numbering . . . 298
8-9 Physical port location to area numbering cross reference . . . 299
8-10 IBM TotalStorage SAN Switch M14 . . . 301
8-11 Port side of M14 . . . 304
8-12 Port card . . . 305
8-13 Parallel ISLs without trunking . . . 312
8-14 2109 ISL trunking . . . 314
8-15 Extended Fabrics feature using dark fiber and DWDM . . . 316

8-16 Remote Switch feature using ATM 317
8-17 SES management 328
8-18 IBM SAN Switch zoning 330
8-19 Overlapping zones 333
9-1 CNT Ultranet Multi-services Director Heterogeneous Extended Fabric346
9-2 CNT FC/9000 director based heterogeneous storage networking. . . . 347
9-3 Ultranet Edge Storage Router extending the fabric across WAN 348
9-4 CNT Ultranet Edge Storage Router 349
9-5 CNT Ultranet Replication Appliance 350
9-6 Ultranet Replication Appliance with front cover open 352
9-7 CNT Ultranet Multi-services Director Heterogeneous Extended Fabric354
9-8 CNT Ultranet Multi-service Director Model N16 355
9-9 CNT SAN256N TFIO Module 357
9-10 CNT SAN256N Control Module (TCM) 358
9-11 CNT SAN256N TSW Switch Module 359
9-12 CNT SAN256N Power Supply Assembly 360
9-13 CNT SAN256N Upper fan assembly 361
9-14 CNT SAN256N side fan assembly. 362
9-15 CNT SAN256N Air Baffle 363
9-16 SAN256N design for high availability 365
9-17 CNT FC/9000-64 Fibre Channel Director 367
9-18 CNT FC/9000-128 Fibre Channel Director 368
9-19 CNT FC/9000-256 Fibre Channel Director 369
9-20 FC/9000 director scalability 371
9-21 CNT director XFIO2 module 374
9-22 8-port XFIO2 module 375
9-23 FWI module 376
9-24 FCME card 377
9-25 Director - power supply assembly 378
9-26 CNT director - fan assembly 378
9-27 CNT FC/9000 director - module locations 380
9-28 8 port FIO module 381
9-29 CNT director FIO module 382
9-30 FSW module 383
9-31 FCM module 384
9-32 SAN256N Logical Domains 388
9-33 Typical CNT director communication setup 396
9-34 inVSN primary option screen 398
9-35 Display of inVSN available options 400
9-36 inVSN: Using CNT Audit-Trail to monitor user activities 403
9-37 inVSN: Accessing the CNT event log 404
9-38 inVSN Enterprise Manager screenshot of multiple switch connected tn a fabric405

9-39 CNT SAN256N logical domains and name server zoning 409
9-40 One CNT fabric consisting of two cascaded directors 412
9-41 Port design form. 413
9-42 Description of an adjacent port 414
9-43 Remote site distance solution 415
9-44 Single fabric can consist of 8 directors with maximum hop count of 3. 416
9-45 Port display screen 417
9-46 CNT wait cycle process 423
10-1 McDATA Sphereon 4300 Fabric Switch 432
10-2 McDATA Sphereon 4500 Fabric Switch 434
10-3 McDATA Sphereon 3232 Fabric Switch switch 435
10-4 McDATA Intrepid 6140 Director 438
10-5 ED-6140 port map (front) 440
10-6 ED-6140 port map (rear) 440
10-7 McDATA Intrepid 6140 Director hardware (front view) 443
10-8 McDATA Intrepid 6140 Director hardware (rear view) 444
10-9 McDATA Intrepid 6064 Director 446
10-10 ED-6064 port map 447
10-11 McDATA Intrepid 6064 Director hardware (front view) 449
10-12 McDATA Intrepid 6064 Director hardware (rear view) 451
10-13 McDATA Open Trunking 462
10-14 LCD panel on front of McDATA 1U EFC Management Server 468
10-15 Rear view of McDATA 1U EFC Management Server 468
10-16 EFC Server public intranet (one ethernet connection) 470
10-17 EFC Server private network (two ethernet connections) 471
10-18 EFCM 8.0 main window 473
11-1 MDS 9120 Multilayer Switch (IBM 2061-020) 490
11-2 MDS 9140 Multilayer Switch (IBM 2061-040) 491
11-3 MDS 9216 Multilayer Switch (IBM 2062-D01) 491
11-4 Cisco MDS 9216 Multilayer Fabric Switch layout 492
11-5 MDS 9506 Multilayer Director (IBM 2062-D04) 494
11-6 MDS 9509 Multilayer Director (IBM 2062-D07) 495
11-7 Cisco MDS 9509 Multilayer Director layout 497
11-8 Cisco MDS 9000 family port types 502
11-9 MDS 9500 Series supervisor module 504
11-10 16 port switching module 506
11-11 32 port switching module 507
11-12 Cisco MDS 9000 14+2 Multi-Protocol Services Module 507
11-13 8-port IP Services Module 508
11-14 Caching Services Module 509
11-15 Cisco MDS9000 Port Analyzer Adapter -2. 509
11-16 PortChannels and ISLs on the Cisco MDS 9000 family of switches . . 515
11-17 Trunking and PortChanneling 516

11-18 Forward Congestion Control . . . 518
11-19 Traditional SAN . . . 521
11-20 Virtual SAN . . . 522
11-21 Inter VSAN Routing . . . 526
11-22 Security with local authentication . . . 528
11-23 Security with RADIUS server . . . 529
11-24 SPAN destination ports . . . 532
11-25 SD_Port for incoming traffic (ingress direction) . . . 533
11-26 SD_Port for outgoing traffic (egress direction) . . . 533
11-27 Fibre Channel analyzer without SPAN . . . 537
11-28 Fibre Channel analyzer using SPAN . . . 538
11-29 Using a single SD_Port to monitor traffic . . . 539
11-30 Out-of-band management connection . . . 541
11-31 In-band management connection . . . 542
11-32 Zoning overview . . . 545
12-1 Storage Manager View Event Log icon . . . 563
13-1 Fabric in a director . . . 573
13-2 Two examples of switch cascading . . . 575
13-3 Ring design . . . 575
13-4 Meshed network design . . . 576
13-5 Host-tier and storage-tier . . . 577
13-6 Tier to tier . . . 577
13-7 Core-edge design . . . 578
14-1 High performance design . . . 584
14-2 Expanding the SAN fabric via E_Ports . . . 587
14-3 2109-M14 core-edge solution . . . 589
14-4 High availability dual enterprise SAN fabric . . . 592
14-5 Microsoft cluster with dual switch with redundant fabric . . . 595
14-6 Datacenter Server 2000 for MSCS . . . 598
14-7 Secure SAN . . . 601
14-8 QuickLoop . . . 603
15-1 High performance design . . . 608
15-2 Expanded SAN fabric via E_Ports . . . 611
15-3 Redundant fabrics . . . 613
15-4 Dual sites . . . 615
15-5 Single director clustering solution . . . 618
15-6 Secure solution . . . 622
15-7 Tape attachment using Sphereon 2026-224 switches . . . 626
15-8 Tape zoning . . . 627
16-1 High performance design . . . 632
16-2 Dual fabric design . . . 637
16-3 Dual fabric design over distance . . . 640
16-4 High availability solution using CNT Broadcast feature . . . 644

16-5 Solution using logical domains on the CNT(INRANGE) director 649
16-6 Protecting your data from both human error and sabotage 653
16-7 A heterogeneous solution connected to a single SAN fabric 655
17-1 High performance design 660
17-2 Expanding the SAN fabric via E_Ports 663
17-3 Traditional dual fabric design (without VSANs) 666
17-4 Dual fabric design with VSANs 667
17-5 Traditional across site dual fabric design 669
17-6 Across site Dual fabric design using VSANs 670
17-7 Microsoft cluster with redundant fabric 673
17-8 Datacenter Server 2000 for MSCS 676
17-9 Protecting your data from both human error and sabotage 680
17-10 Utilizing the Cisco MDS 9000 TL_Port 682
18-1 Case Study 2: Server Schematic 691
18-2 Different ISL for data access and for data replication 695
18-3 Case Study 3: Server Schematic 698
18-4 Case Study 4: Server schematic 702
18-5 Case Study 5: Server Schematic 707
18-6 Case Study 6: Server Schematic 711
19-1 Core SAN design 716
19-2 Adding additional two switches to the SAN 718
19-3 SAN after first year of expansion 719
19-4 SAN design after three years of operation 720
19-5 Core edge design 722
19-6 Management network 725
19-7 Core switch design 726
19-8 Expanding to 40 servers using core switch technology 728
19-9 Getwell site initial design 729
19-10 Feelinbad site initial design 731
19-11 Adding additional storage ports for non-SGI servers 733
19-12 Trunking in Getwell Center 735
19-13 Core switch design for Getwell site 738
19-14 Core switch design for Feelinbad site 739
19-15 Proposed design for Company Three 741
19-16 Initial design using IBM TotalStorage SAN Switch M14 745
19-17 Proposed design for Company Five 748
19-18 Proposed design for Primary site 752
19-19 Proposed design for Secondary site 753
19-20 DWDM connection between sites 754
20-1 Core SAN design using a McDATA ED-6064 Director 760
20-2 Fully redundant McDATA ED-6064 Director solution 762
20-3 McDATA ED-6064 solution with all potential servers 762
20-4 Management network 766

20-5 Initial design using McDATA 2027-232 Switches 767
20-6 Final design to accommodate all potential servers 769
20-7 Management network 773
20-8 Getwell SAN design using McDATA ED-6064 Director and 2027-232 switches774
20-9 Feelinbad SAN design using McDATA ES-3032 switches 775
20-10 Management network 780
20-11 Initial design using McDATA 2027-232 Switches 781
20-12 Management network 784
20-13 Initial design using McDATA ED-6064 Switches. 786
20-14 Management network 789
20-15 The proposed design for Company Five 790
20-16 Management network 793
20-17 The proposed design for the Primary site 794
20-18 The proposed design for the Secondary site 795
20-19 The complete solution 796
20-20 Management network 800
21-1 Single director solution. 802
21-2 The 20-server solution 804
21-3 The 40 server, single director solution 806
21-4 The 40-server, dual director solution 807
21-5 Management of a CNT SAN solution. 810
21-6 The Getwell configuration 812
21-7 The Feelinbad configuration 813
21-8 The proposed design for Company Three 818
21-9 The proposed solution for the East site 822
21-10 The proposed solution for the West site 823
21-11 The combined solution. 824
21-12 The proposed design for Company Five 827
21-13 The proposed design for the Primary site 830
21-14 The proposed design for the Secondary site 831
21-15 The complete solution 832
22-1 Core SAN design using a Cisco MDS 9506 Director 838
22-2 Fully redundant Cisco MDS 9506 Director solution. 840
22-3 Cisco MDS 9506 solution with all potential servers 841
22-4 Management network 845
22-5 Initial design using Cisco MDS 9140 Switches 846
22-6 Final design to accommodate all potential servers 848
22-7 Management network 852
22-8 Getwell SAN design using Cisco MDS 9506 Director 853
22-9 Feelinbad SAN design using Cisco MDS 9506 directors. 854
22-10 Initial design using Cisco MDS 9216 Switches 859
22-11 Management network 862

22-12 Initial design using Cisco MDS 9506 director. 863
22-13 Management network 867
22-14 The proposed design for Company Five 868
22-15 Management network 871
22-16 The proposed design for the Primary site 872
22-17 The proposed design for the Secondary site 873
22-18 The complete solution 874
23-1 Remote disk consolidation 883
23-2 SAN distance extension up to 10 km 887
23-3 SAN distance extension up to 100 km with repeaters 892
23-4 Point-to-point DWDM solution - two sites 894
23-5 Multiple site: Ring topology DWDM solution 899
23-6 SAN extension over (ATM) WAN using channel extenders 902
23-7 Remote access via FCIP router over TCP/IP 904
23-8 Remote tape vaulting 906
23-9 Remote tape vaulting with disaster tolerance 908
24-1 Remote disk consolidation 912
24-2 SAN distance extension up to 10 km 916
24-3 SAN distance extension up to 100 km with repeaters 921
24-4 Point-to-point DWDM solution - two sites 923
24-5 Multiple site: Ring topology DWDM solution 928
24-6 SAN extension over (ATM) WAN using channel extenders 931
24-7 Remote access via iFCP switch over TCP/IP 933
24-8 Remote tape vaulting 935
24-9 Remote tape vaulting with disaster tolerance 937
25-1 Remote disk consolidation 942
25-2 SAN distance extension up to 10 km 946
25-3 SAN distance extension up to 100 km with repeaters 951
25-4 Point-to-point DWDM solution - two sites 953
25-5 Multiple site: Ring topology DWDM solution 958
25-6 SAN extension over (ATM) WAN using channel extenders 961
25-7 Remote access via FCIP router over TCP/IP 963
25-8 Remote tape vaulting 965
25-9 Remote tape vaulting with disaster tolerance 967
26-1 Remote disk consolidation 972
26-2 SAN distance extension up to 10 km 976
26-3 SAN distance extension up to 100 km with repeaters 981
26-4 Point-to-point DWDM solution - two sites 983
26-5 Multiple site: Ring topology DWDM solution 988
26-6 SAN extension over (ATM) WAN using channel extenders 991
26-7 Remote access via FCIP switch over TCP/IP 993
26-8 Remote tape vaulting 995
26-9 Remote tape vaulting with disaster tolerance 997

27-1 Connecting high I/O servers to the core switches 1003
27-2 The CNT port population guidelines 1005

Tables

2-1 Distances using SFP-based fiber optics 22
2-2 Distance using 1 Gb/s GBIC based fiber optics. 24
2-3 Distance using 2 Gb/s GBIC based fiber-optics. 24
3-1 WWN company identifiers 76
3-2 Application bytes per second on 2 Gb/s link using 2048-byte payload . 97
3-3 10 bit encodings of K28.5 119
3-4 Transmission words in a frame 122
3-5 The frame header 122
3-6 Name server entries. 133
3-7 Phantom addresses 133
3-8 Copper and Fibre Channel limits 134
3-9 T11 projects 142
3-10 Comparison of management access methods. 154
3-11 Comparison of Management Access Method 155
5-1 Comparison between structured and non-structured cables 217
5-2 Checklist for proposed layout. 226
8-1 Comparison of management access method 318
8-2 Different zone types 335
10-1 Machine type and model number changes 431
11-1 Physical specifications for the MDS 9216 493
11-2 Physical specifications for the MDS 9509 498
11-3 Distance specifications for SFP optics. 503
11-4 Cisco MDS9000 Port Analyzer Adapter truncate modes. 510
11-5 Fabric Manager standard and Server comparison. 513
11-6 VSAN versus zoning 523
11-7 Interoperability mode changes. 548
15-1 FC-SW capable tape drives 630
18-1 Case Study 2: Server and storage inventory 690
19-1 Number of used ports 716
19-2 Number of used ports 719
19-3 Number of used ports 721
19-4 Number of used ports 722
19-5 Number of used ports 727
19-6 Number of used ports 728
19-7 Number of used ports 729
19-8 Number of used ports 731
19-9 Number of used ports 738
19-10 Number of used ports 739

© Copyright IBM Corp. 2004. All rights reserved.

19-11 Number of used ports 741
19-12 Ports: East site 746
19-13 Ports: West Site 746
19-14 Number of used ports 749
19-15 Number of used ports 753
19-16 Number of used ports 754
20-1 Year 1 ports: 10 servers. 763
20-2 Year 2 ports: 20 servers. 763
20-3 Year 3 ports: 40 servers. 763
20-4 Year 1 ports: 10 servers. 770
20-5 Year 2 ports: 20 servers. 770
20-6 Year 3 ports: 40 servers. 770
20-7 Initial ports 775
20-8 Initial ports 776
20-9 Port usage 782
20-10 Ports: East site 787
20-11 Ports: West Site 787
20-12 Initial ports 791
20-13 Ports at Local site 797
20-14 Ports at Remote site 797
21-1 Port usage at Company One in the first year. 803
21-2 Port usage at Company One in the second year 805
21-3 Port usage at Company One in the third year, single director 806
21-4 Port usage at Company One in the third year, dual director 807
21-5 Summary of port requirements 811
21-6 Port usage at Getwell. 813
21-7 Port usage at Feelinbad. 814
21-8 Port usage at Company Three. 818
21-9 Port usage at the East site. 822
21-10 Port usage at the West Site 823
21-11 Port usage at Company Five 827
21-12 Port usage at the Primary site 832
21-13 Port usage at the Secondary site. 832
22-1 Year 1 ports: 10 servers. 841
22-2 Year 2 ports: 20 servers. 842
22-3 Year 3 ports: 40 servers. 842
22-4 Year 1 ports: 10 servers. 849
22-5 Year 2 ports: 20 servers. 849
22-6 Year 3 ports: 40 servers. 849
22-7 Initial ports 854
22-8 Initial ports 855
22-9 Port usage 860
22-10 Ports: East site 864

22-11 Ports: West Site 864
22-12 Initial ports 869
22-13 Ports at Local site 875
22-14 Ports at Remote site 875

Examples

Note to Author: This file needs to be removed from your .book before final publication. Delete the LOE file from your book in FrameMaker 6.0 using:
Open .book > select the file > Edit > Delete File from Book

12-1 setting the time to 15:31 on Feb 27 2005 561
12-2 Setting the date and time on the 3583 SDGM 564
12-3 564
12-4 564
13-1 Host-to-storage oversubscription 580

© Copyright IBM Corp. 2004. All rights reserved.

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:
This information contains sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

© Copyright IBM Corp. 2004. All rights reserved.

Trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

@server®	FICON®	Redbooks™
@server®	HACMP™	RS/6000®
Redbooks	Illustra™	S/360™
(logo) ™ibm.com®	Informix®	S/370™
iSeries™	IBM®	S/390®
pSeries®	Lotus®	Service Director™
xSeries®	Magstar®	Storage Tank™
z/Architecture™	Monday™	System/36™
z/OS®	MVS™	System/360™
zSeries®	Netfinity®	System/370™
AFS®	NetView®	System/38™
AIX®	NetVista™	System/390®
AS/400®	Notes®	SANergy®
CICS®	NUMA-Q®	Tivoli Enterprise™
DB2®	OS/390®	Tivoli Enterprise Console®
Enterprise Storage Server®	OS/400®	Tivoli®
Enterprise Systems Architecture/390®	Parallel Sysplex®	TotalStorage®
ESCON®	Perform™	Wave®
FlashCopy®	PowerPC®	WebSphere®
	PR/SM™	

The following terms are trademarks of other companies:

Intel, Intel Inside (logos), MMX, and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

SET, SET Secure Electronic Transaction, and the SET Logo are trademarks owned by SET Secure Electronic Transaction LLC.

Other company, product, and service names may be trademarks or service marks of others.

Preface

In this IBM Redbook, we revisit some of the core components and technologies that underpin a storage area network (SAN). We cover, to varying degrees, some of the latest additions to the IBM SAN portfolio, discuss general SAN design considerations, and build these considerations into a selection of real world case studies.

We have also consolidated material from other SAN redbooks to create a complete overview of the depth and breadth of the IBM TotalStorage SAN portfolio.

We realise that there are many ways to design a SAN and put all the components together. In our examples, we have incorporated the major considerations that need to be taken into account, but still left room to manoeuvre on the SAN field of play.

This redbook focuses on the SAN products that are generally considered to form the backbone of the SAN fabric today: switches and directors. With this backbone, developing it has prompted discrete approaches to the design of a SAN fabric. The bespoke vendor implementation of technology that is characteristic in the design footprint of switches and directors, means that we have an opportunity to answer challenges in different ways.

We will show examples where strength can be built in to our SAN using the network and the features of the components themselves. Our aim is to show that you can cut your SAN fabric according to your cloth.

The team that wrote this redbook

This redbook was produced by a team of specialists from around the world working at the International Technical Support Organization, San Jose Center.

Jon Tate is a Project Manager for IBM TotalStorage SAN Solutions at the International Technical Support Organization, San Jose Center. Before joining the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2 support for IBM storage products. Jon has 19 years of experience in storage software and management, services, and support, and is both an IBM Certified IT Specialist, and an IBM SAN Certified Specialist.

Stephen Garraway is an IT specialist.

© Copyright IBM Corp. 2004. All rights reserved.

Ronda Hruby is an IT specialist.

Alexander Ignacio is an IT specialist.

Glen Routley is an IT specialist.

Russell Nunag is an IT specialist.

Madhav Vaze is an IT specialist.

Bruce Wilson is an IT specialist.

Thanks to the following people for their contributions to this project:

Tom Cady
Deanna Polm
Sangam Racherla
Sokkieng Wang
International Technical Support Organization, San Jose Center

Jim Banask
Cal Blombaum
William Champion
Scott Drummond
Parker Grannis
Edith Kropf
Pam Lukes
Victoria Perris
Michael Starling
Diana Tseng
Ernie Williamson
Michelle Wright
IBM Storage Systems Group

Jim Baldyga
Brian Steffler
Brocade Communications Systems

Dave Burchwell
CNT Technologies Corporation

Brent Anderson
McDATA Corporation

Tom and Jenny Chang

Garden Inn Hotel, Los Gatos, California

Become a published author

Join us for a two- to six-week residency program! Help write an IBM Redbook dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You'll team with IBM technical professionals, Business Partners and/or customers.

Your efforts will help increase product acceptance and customer satisfaction. As a bonus, you'll develop a network of contacts in IBM development labs, and increase your productivity and marketability.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our Redbooks to be as helpful as possible. Send us your comments about this or other Redbooks in one of the following ways:

- Use the online **Contact us** review redbook form found at:

  ibm.com/redbooks

- Send your comments in an Internet note to:

  redbook@us.ibm.com

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. QXXE Building 80-E2
  650 Harry Road
  San Jose, California 95120-6099

1

Introduction

Until recently, disaster planning for businesses focused on recovering centralized data centers following a catastrophe, either natural or man-made. While these measure remain important to disaster planning, the protection they provide is far from adequate for today's distributed computing environments.

The goal for companies today is to achieve a state of business continuity, where critical systems and networks are always available. To attain and sustain business continuity, companies must engineer availability, security, and reliability into every process from the outset.

In this chapter we consider the many benefits SAN has to offer in these areas.

© Copyright IBM Corp. 2004. All rights reserved.

1.1 Beyond disaster recovery

When disaster recovery emerged as a formal discipline and a commercial business in the 1980s, the focus was on protecting the data center — the heart of a company's heavily centralized IT structure. This model began to shift in the early 1990s to distributed computing and client/server technology.

At the same time, information technology became embedded in the fabric of virtually every aspect of a business. Computing was no longer something done in the background. Instead, critical business data could be found across the enterprise — on desktop PCs and departmental local area networks, as well as in the data center.

This evolution continues today. Key business initiatives such as enterprise resource planning (ERP), supply chain management, customer relationship management and e-business have all made continuous, ubiquitous access to information crucial to an organization. This means business can no longer function without information technology in the following areas:

- Data
- Software
- Hardware
- Networks
- Call centers
- Laptop computers

A company that sells products on the Web, for example, or supports customers with an around-the-clock call center, must be operational 24 hours a day, 7 days a week, or customers will go elsewhere. An enterprise that uses e-business to acquire and distribute parts and products is not only dependent on its own technology but that of its suppliers. As a result, protecting critical business processes, with all their complex interdependencies, has become as important as safeguarding data itself.

The goal for companies with no business tolerance for downtime is to achieve a state of business continuity, where critical systems and networks are continuously available, no matter what happens. This means thinking proactively: engineering availability, security, and reliability into business processes from the outset — not retrofitting a disaster recovery plan to accommodate ongoing business continuity requirements.

1.1.1 Whose responsibility is it?

Many senior executives and business managers consider business continuity the responsibility of the IT department. However, it is no longer sufficient or practical to vest the responsibility exclusively in one group. Web-based and distributed computing have made business processes too complex and decentralized. What's more, a company's reputation, customer base and, of course, revenue and profits are at stake. All executives, managers, and employees must therefore participate in the development, implementation, and ongoing support of continuity assessment and planning.

The same information technology driving new sources of competitive advantage has also created new expectations and vulnerabilities. On the Web, companies have the potential to deliver immediate satisfaction — or dissatisfaction — to millions of people. Within ERP and supply chain environments, organizations can reap the rewards of improved efficiencies, or feel the impact of a disruption anywhere within their integrated processes.

With serious business interruption now measured in minutes rather than hours, even success can bring about a business disaster. Web companies today worry more about their ability to handle unexpected peaks in customer traffic than about fires or floods — and for good reason. For example, an infrastructure that cannot accommodate a sudden 200 percent increase in Web site traffic generated by a successful advertising campaign can result in missed opportunities, reduced revenues, and a tarnished brand image. Because electronic transactions and communications take place so quickly, the amount of work and business lost in an hour far exceeds the toll of previous decades. According to reports, the financial impact of a major system outage can be enormous:

- US$6.5 million per hour in the case of a brokerage operation
- US$2.6 million per hour for a credit-card sales authorization system
- US$14,500 per hour in automated teller machine (ATM) fees if an ATM system is offline

Even what was once considered a minor problem — a faulty hard drive or a software glitch — can cause the same level of loss as a power outage or a flooded data center, if a critical business process is affected. For example, it has been calculated that the average financial loss per hour of disk array downtime stands at:

- US$29,301 in the securities industry
- US$26,761 for manufacturing
- US$17,093 for banking
- US$9,435 for transportation

More difficult to calculate are the intangible damages a company can suffer: lower morale and productivity, increased employee stress, delays in key project time lines, diverted resources, regulatory scrutiny, and a tainted public image. In this climate, executives responsible for company performance now find their personal reputations at risk. Routinely, companies that suffer online business disruptions for any reason make headlines the next day, with individuals singled out by the press. Moreover, corporate directors and officers can be liable for the consequences of business interruption or loss of business-critical information. Most large companies stipulate in their contracts that suppliers must deliver services or products under any circumstances. What's more, adequate protection of data may be required by law, particularly for a public company, financial institution, utility, health care organization, or government agency.

Together, these factors make business continuity the shared responsibility of an organization's entire senior management, from the CEO to line-of-business executives in charge of crucial business processes. Although IT remains central to the business continuity formula, IT management alone cannot determine which processes are critical to the business and how much the company should pay to protect those resources.

1.1.2 The Internet brings increased risks

A recent IBM survey of 226 business recovery corporate managers revealed that only eight percent of Internet businesses are prepared for a computer system disaster. Yet doing business online means exposing many business-critical applications to a host of new risks. While the Internet creates tremendous opportunity for competitive advantage, it can also give partners, suppliers, customers, employees and hackers increased access to corporate IT infrastructures. Unintentional or malicious acts can result in a major IT disruption. Moreover, operating a Web site generates organizational and system-related interdependencies that fall outside of a company's control from Internet Service Providers (ISP) and telecommunications carriers to the hundreds of millions of public network users.

Therefore, the greatest risk to a company's IT operations may no longer be a hurricane, a 100-year flood, a power outage, or even a burst pipe. Planning for continuity in an e-business environment must address vulnerability to network attacks, hacker intrusions, viruses, and spam, as well as ISP and telecommunication line failures.

1.1.3 Planning for business continuity

Few organizations have the need or the resources to assure business continuity equally for every functional area. Therefore, any company that has implemented a single business continuity strategy for the entire organization is likely under-prepared, or spending money unnecessarily. The key to business continuity lies in understanding your business, determining which processes are critical to staying in that business, and identifying all the elements crucial to those processes. Specialized skills and knowledge, physical facilities, training, and employee satisfaction, as well as information technology, should all be considered. By thoroughly analyzing these elements, you can accurately identify potential risks and make informed business decisions about accepting, mitigating or transferring those risks.

Once you have developed a program for assuring that critical processes will be available around the clock, you should assume that it will fail — and commit to keeping your program current with business and technology infrastructure changes. A fail-safe strategy assumes that no business continuity program can provide absolute protection from every type of damage — no matter how comprehensive your high-availability, redundancy, fault tolerance, clustering, and mirroring strategies.

Today, the disasters most likely to bring your business to a halt are the result of human error or malice: the employee who accidentally deletes a crucial block of data; the disgruntled ex-employee seeking revenge by introducing a debilitating virus; the thief who steals vital trade secrets from your mainframe; or the hacker who invades your network. According to a joint study by the U.S. Federal Bureau of Investigation and the Computer Security Institute, the number and severity of successful corporate hacks is increasing dramatically, particularly intrusions by company insiders. In one study, 250 Fortune 1000 companies reported losses totaling US$137 million in 1997 — an increase of 37 percent over the previous year.

Making an executive commitment to regularly testing, validating, and refreshing your business continuity program can protect your company against perhaps the greatest risk of all — complacency. In the current environment of rapid business and technology change, even the smallest alteration to a critical application or system within your enterprise or supply chain can cause an unanticipated failure, impacting your business continuity. Effective business protection planning addresses not only what you need today, but what you will need tomorrow and into the future.

1.2 Using a SAN for business continuance

Although some of the concepts that we detail purely apply to only the SAN environment, there are general considerations that need to be taken into account in any environment. Any company that is serious about business continuance will have considered and applied processes or procedures to take into account any of the eventualities that may occur, such as those listed in Figure 1-1.

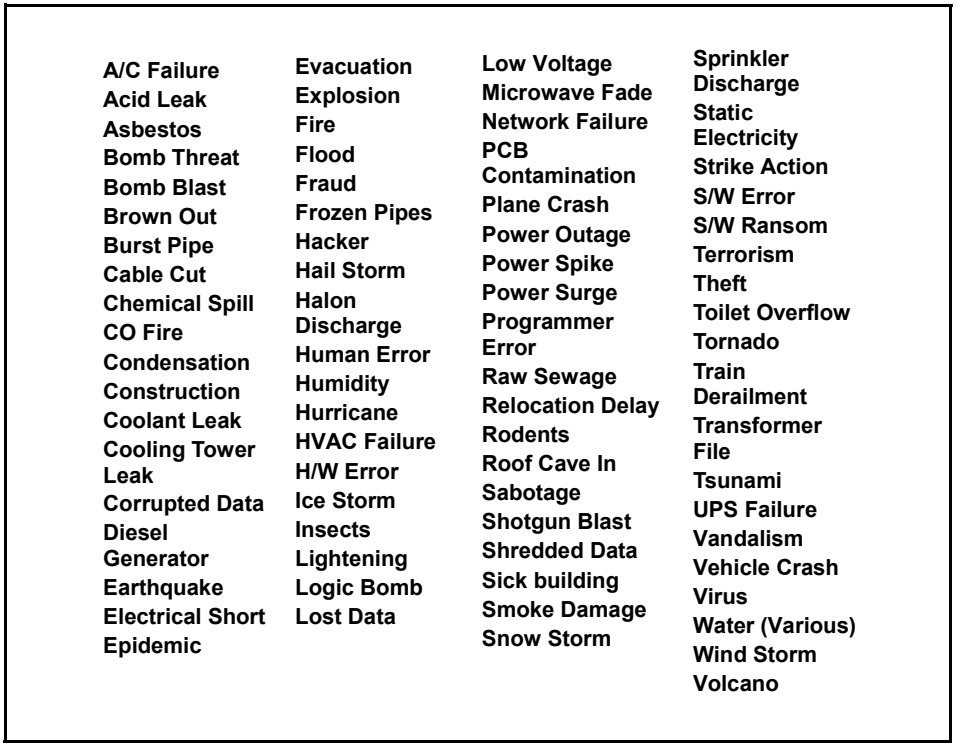


Figure 1-1 Business outage causes

Some of these problems are not necessarily common to all regions throughout the world, but they should be considered nonetheless, even if only to dismiss the eventuality that they may happen. Careful consideration will result in a deeper understanding of what is likely to cause a business outage, rather than just adopting an attitude that says “it will not happen to me”. After all, remember that the Titanic was unsinkable.

1.2.1 SANs and business continuance

So why would the risk increase if you were to implement a SAN in your environment? The short answer is that it may not increase the risk. It may expose you to more risk over a greater area, for example, the SCSI 25 m restriction means that a small bomb planted in the correct position would do quite nicely. If you are using a SAN for distance solutions, then it might be necessary to increase the size of the bomb, or plant many more of them, to cause the same effect.

What a SAN means is that you now are beginning to explore the potential for ensuring that your business can actually continue in the wake of a disaster. It may be able to do this by:

- Providing for greater operational distances
- Providing mirrored storage solutions for local disasters
- Providing failover support for local disasters
- Providing remote vaulting anywhere in the world
- Providing high availability file serving functionality
- Providing the ability to avoid space outage situations for higher availability

If we are to take the simple example of distance, what a SAN will allow you to do is to break the SCSI distance barrier. Does this in itself make you any safer? Of course it doesn't. Does it give you an opportunity to minimize the risk to your business. Of course it does.

It is up to you if you decide to use that to your advantage, or ignore it and the other benefits that it may bring to your business. One thing is certain though; if you don't exploit the SAN's potential to its fullest, other people may. Those other people might be your competitors. Does that worry you? If it doesn't, then you may want to stop reading right now, because this redbook is not for you! We are targeting those people that are concerned with unleashing the potential of their SAN, or are interested in seeing what a SAN can do.

But that is not all we will do. We will provide you with as much information as we can, that will cover the data center environment from floor to ceiling and the considerations that you should take to ensure minimal exposure to any outage.

As availability is linked to business continuance and recovery, we will also cover methods that can be employed to ensure that the data in your SAN is available to those that are authorized to access it, and protected from those that aren't.

1.3 SAN business benefits

Today's business environment creates many challenges for the enterprise IT planner. This is a true statement and relates to more than just business continuance, so perhaps now is a good time to look at whether deploying a SAN will solve more than just one problem. It may be an opportunity to look at where you are today and where you want to be in three years' time. Is it better to plan for migration to a SAN from the start, or try to implement one later after other solutions have been considered and possibly implemented? Are you sure that the equipment that you install today will still be usable three years later? Is there any use that you can make of it outside of business continuance? A journey of a thousand miles begins with one step.

In the topics that follow we will remind you of some of the business benefits that SANs can provide. We have identified some of the operational problems that a business faces today, and which could potentially be solved by a SAN implementation.

1.3.1 Storage consolidation and sharing of resources

By enabling storage capacity to be connected to servers at a greater distance, and by disconnecting storage resource management from individual hosts, a SAN enables disk storage capacity to be consolidated. The results can be lower overall costs through better utilization of the storage, lower management costs, increased flexibility, and increased control.

This can be achieved physically or logically, as we explain in the following sections.

Physical consolidation

Data from disparate storage subsystems can be combined on to large, enterprise class shared disk arrays, which may be located at some distance from the servers. The capacity of these disk arrays can be shared by multiple servers, and users may also benefit from the advanced functions typically offered with such subsystems. This may include RAID capabilities, remote mirroring, and instantaneous data replication functions, which might not be available with smaller, integrated disks. The array capacity may be partitioned, so that each server has an appropriate portion of the available gigabytes.

Physical consolidation of storage is shown in Figure 1-2.

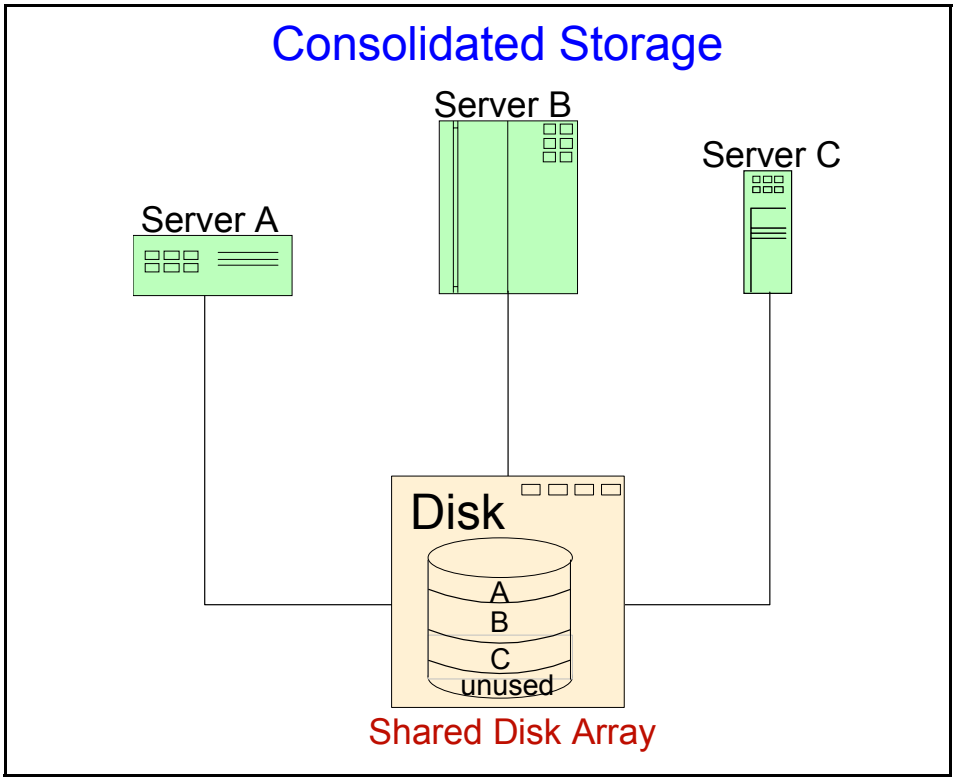


Figure 1-2 Storage consolidation

Available capacity can be dynamically allocated to any server requiring additional space. Capacity not required by a server application can be re-allocated to other servers. This avoids the inefficiency associated with free disk capacity attached to one server not being usable by other servers. Extra capacity may be added, in a non-disruptive manner.

Logical consolidation

It is possible to achieve shared resource benefits from the SAN, but without moving existing equipment. A SAN relationship can be established between a client and a group of storage devices that are not physically co-located (excluding devices which are internally attached to servers). A logical view of the combined disk resources may allow available capacity to be allocated and re-allocated between different applications running on distributed servers, to achieve better utilization. Consolidation is covered in greater depth in *IBM Storage Solutions for Server Consolidation*, SG24-5355.

In Figure 1-3 we show a logical consolidation of storage.

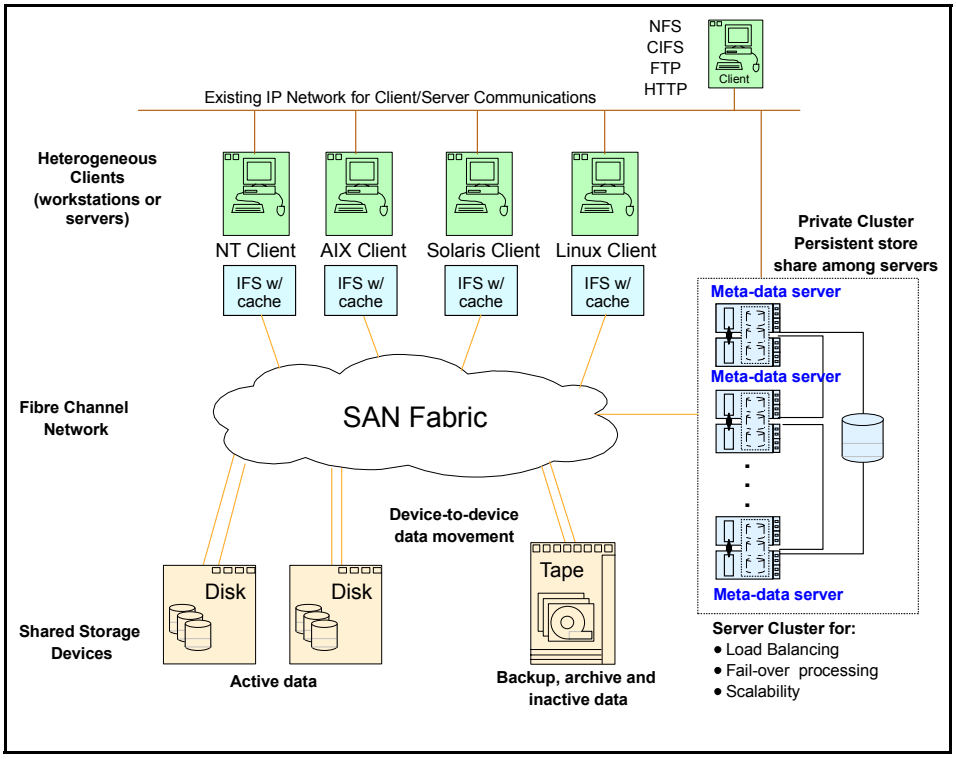


Figure 1-3 Logical storage consolidation

1.3.2 Data sharing

The term “data sharing” is used somewhat loosely by users and some vendors. It is sometimes interpreted to mean the replication of files or databases to enable two or more users, or applications, to concurrently use separate copies of the data. The applications concerned may operate on different host platforms. A SAN may ease the creation of such duplicated copies of data using facilities such as remote mirroring.

Data sharing may also be used to describe multiple users accessing a single copy of a file. This could be called “true data sharing”. In a homogeneous server environment, with appropriate application software controls, multiple servers may access a single copy of data stored on a consolidated storage subsystem.

If attached servers are heterogeneous platforms (for example, with a mix of UNIX® and Windows® NT), sharing of data between such unlike operating system environments is complex. This is due to differences in file systems, data

formats, and encoding structures. IBM, however, uniquely offers a true data sharing capability, with concurrent update, for selected heterogeneous server environments, using the Tivoli® SANergy™ File Sharing solution.

The SAN advantage in enabling enhanced data sharing may reduce the need to hold multiple copies of the same file or database. This reduces duplication of hardware costs to store such copies. It also enhances the ability to implement cross enterprise applications, such as e-business, which may be inhibited when multiple data copies are stored.

1.3.3 Non-disruptive scalability for growth

There is an explosion in the quantity of data stored by the majority of organizations. This is fueled by the implementation of applications, such as e-business, e-mail, Business Intelligence, Data Warehouse, and Enterprise Resource Planning. Some industry analysts, such as IDC and Gartner Group, estimate that electronically stored data is doubling every year. In the case of e-business applications, opening the business to the Internet, there have been reports of data growing by more than 10 times annually. This is a nightmare for planners, as it is increasingly difficult to predict storage requirements.

A finite amount of disk storage can be connected physically to an individual server due to adapter, cabling, and distance limitations. With a SAN, new capacity can be added as required, without disrupting ongoing operations. SANs enable disk storage to be scaled independently of servers.

1.3.4 Improved backup and recovery

With data doubling every year, what effect does this have on the backup window? Backup to tape, and recovery, are operations which are problematic in the parallel SCSI or LAN based environments. For disk subsystems attached to specific servers, two options exist for tape backup. Either it must be done onto a server attached tape subsystem, or by moving data across the LAN.

Tape pooling

Providing tape drives to each server is costly, and also involves the added administrative overhead of scheduling the tasks, and managing the tape media. SANs allow for greater connectivity of tape drives and tape libraries, especially at greater distances. Tape pooling is the ability for more than one server to logically share tape drives within an automated library. This can be achieved by software management, using tools, such as Tivoli Storage Manager; or with tape libraries with outboard management, such as the IBM 3494.

LAN-free and server-free data movement

Backup using the LAN moves the administration to centralized tape drives or automated tape libraries. However, at the same time, the LAN experiences very high traffic volume during the backup or recovery operations, and this can be extremely disruptive to normal application access to the network. Although backups can be scheduled during non-peak periods, this may not allow sufficient time. Also, it may not be practical in an enterprise which operates in multiple time zones.

We illustrate loading the IP network in Figure 1-4.

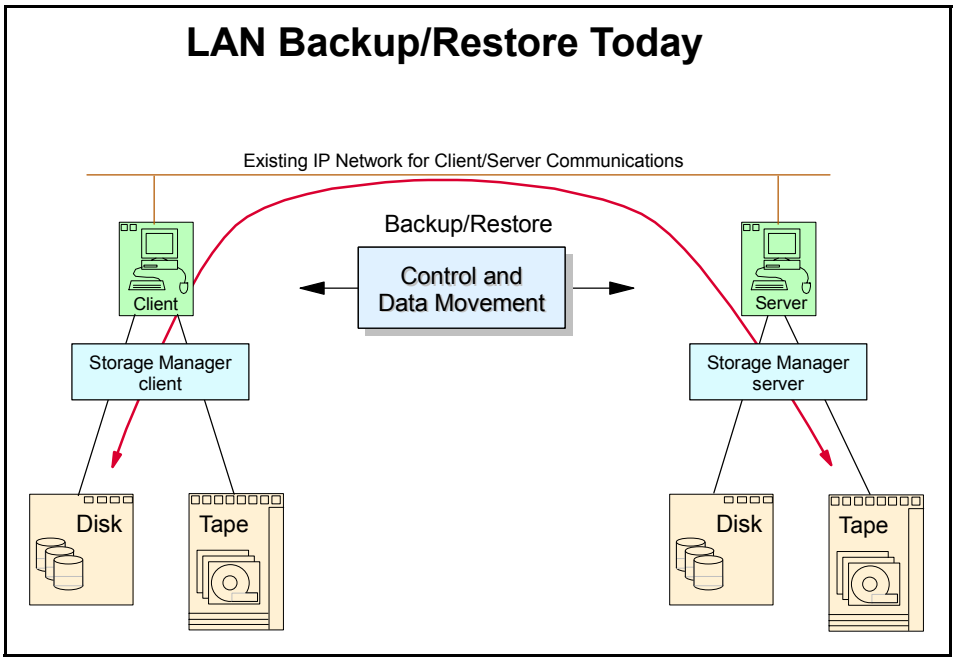


Figure 1-4 Loading the IP network

SAN provides the solution, by enabling the elimination of backup and recovery data movement across the LAN. Fibre Channel's high bandwidth and multi-path switched fabric capabilities enables multiple servers to stream backup data concurrently to high speed tape drives. This frees the LAN for other application traffic. The IBM Tivoli software solution for LAN-free backup offers the capability for clients to move data directly to tape using the SAN. A future enhancement to be provided by IBM Tivoli will allow data to be read directly from disk to tape (and tape to disk), bypassing the server. This solution is known as server-free backup.

1.3.5 High performance

Applications benefit from the more efficient transport mechanism of Fibre Channel. Currently, Fibre Channel transfers data at 200 MB/s, several times faster than typical SCSI capabilities, and many times faster than standard LAN data transfers. Future implementations of Fibre Channel at 400 and 800 MB/s have been defined, offering the promise of even greater performance benefits in the future. Indeed, prototypes of storage components which meet the 2 Gigabit transport specification are already in existence.

The elimination of conflicts on LANs, by removing storage data transfers from the LAN to the SAN, may also significantly improve application performance on servers.

1.3.6 High availability server clustering

Reliable and continuous access to information is an essential prerequisite in any business. As applications have shifted from robust mainframes to the less reliable client/file server environment, so have server and software vendors developed high availability solutions to address the exposure. These are based on clusters of servers. A cluster is a group of independent computers managed as a single system for higher availability, easier manageability, and greater scalability. Server system components are interconnected using specialized cluster interconnects, or open clustering technologies, such as Fibre Channel - Virtual Interface mapping.

Complex software is required to manage the failover of any component of the hardware, the network, or the application. SCSI cabling tends to limit clusters to no more than two servers. A Fibre Channel SAN allows clusters to scale to 4, 8, 16, and even to 100 or more servers, as required, to provide very large shared data configurations, including redundant pathing, RAID protection, and so on. Storage can be shared, and can be easily switched from one server to another. Just as storage capacity can be scaled non-disruptively in a SAN, so can the number of servers in a cluster be increased or decreased dynamically, without impacting the storage environment.

1.3.7 Improved disaster tolerance

Advanced disk arrays, such as IBM Enterprise Storage Server™ (ESS), provide sophisticated functions, like Peer-to-Peer Remote Copy services, to address the need for secure and rapid recovery of data in the event of a disaster. Failures may be due to natural occurrences, such as fire, flood, or earthquake; or to human error. A SAN implementation allows multiple open servers to benefit from this type of disaster protection, and the servers may even be located some distance (up to 10 km) from the disk array which holds the primary copy of the data. The secondary site, holding the mirror image of the data, may be located up to a further 100 km from the primary site.

IBM has also announced Peer-to-Peer Copy capability for its Virtual Tape Server (VTS). This will allow VTS users to maintain local and remote copies of virtual tape volumes, improving data availability by eliminating all single points of failure.

1.3.8 Allow selection of “best of breed” storage

Internal storage, purchased as a feature of the associated server, is often relatively costly. A SAN implementation enables storage purchase decisions to be made independently of the server. Buyers are free to choose the best of breed solution to meet their performance, function, and cost needs. Large capacity external disk arrays may provide an extensive selection of advanced functions. For instance, the ESS includes cross platform functions, such as high performance RAID 5, Peer-to-Peer Remote Copy, Flash Copy, and functions specific to S/390®, such as Parallel Access Volumes (PAV), Multiple Allegiance, and I/O Priority Queuing. This makes it an ideal SAN attached solution to consolidate enterprise data.

Client/server backup solutions often include attachment of low capacity tape drives, or small automated tape subsystems, to individual PCs and departmental servers. This introduces a significant administrative overhead as users, or departmental storage administrators, often have to control the backup and recovery processes manually. A SAN allows the alternative strategy of sharing fewer, highly reliable, powerful tape solutions, such as the IBM Magstar® family of drives and automated libraries, between multiple users and departments.

1.3.9 Ease of data migration

Data can be moved non-disruptively from one storage subsystem to another using a SAN, without server intervention. This may greatly ease the migration of data associated with the introduction of new technology, and the retirement of old devices.

1.3.10 Reduced total costs of ownership

Expenditure on storage today is estimated to be in the region of 50% of a typical IT hardware budget. Some industry analysts expect this to grow to as much as 75% by the end of the year 2002. IT managers are becoming increasingly focused on controlling these growing costs.

Consistent, centralized management

As we have shown, consolidation of storage can reduce wasteful fragmentation of storage attached to multiple servers. It also enables a single, consistent data and storage resource management solution to be implemented, such as IBM StorWatch™ tools, combined with software such as Tivoli Storage Network Manager, Tivoli Storage Manager, and Tivoli SAN Manager, which can reduce costs of software and human resources for storage management.

Reduced hardware costs

By moving data to SAN attached storage subsystems, the servers themselves may no longer need to be configured with native storage. In addition, the introduction of LAN-free and server-free data transfers largely eliminate the use of server cycles to manage housekeeping tasks, such as backup and recovery, and archive and recall. The configuration of what might be termed “thin servers” therefore might be possible, and this could result in significant hardware cost savings to offset against costs of installing the SAN fabric.

1.3.11 Storage resources match e-business enterprise needs

By eliminating islands of information, typical of the client/server model of computing, and introducing an integrated storage infrastructure, SAN solutions match the strategic needs of today’s e-business.

We show this in Figure 1-5.

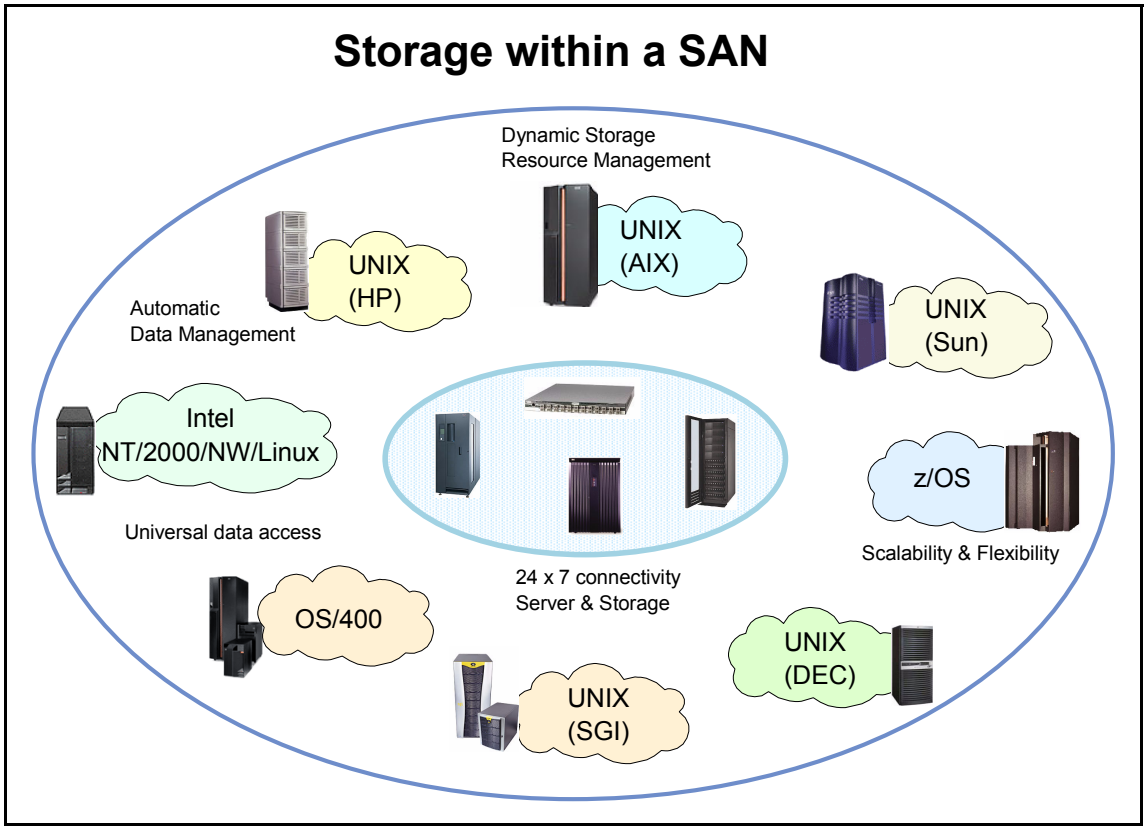


Figure 1-5 SAN total storage solutions

A well designed, well thought out SAN can bring many benefits, and not only those related to business continuance. Utilizing the storage network will be key to the storage and successful retrieval of data in the future, and the days of server centric storage are rapidly becoming a distant memory.

2

SAN fabric components

In this chapter we describe the Fibre Channel products that are used and are likely to be encountered in an IBM Enterprise SAN implementation. This does not mean that you cannot implement other SAN compatible products, including those from other vendors, but the interoperability agreement must be clearly documented and agreed upon.

Fibre Channel is an open standard communications and transport protocol as defined by ANSI (Committee X3T11) and operates over copper and fiber optic cabling at distances of up to 10 kilometers. IBM's implementation is in fiber optic cabling and will be referred to as Fibre Channel cabling, or FC cabling, in this redbook.

We will start off by covering items of hardware that may be used to build a networked storage solution. The whole purpose of a SAN is to interconnect hosts/servers and storage. There are also the components (and their subcomponents) that make up the SAN itself.

Fibre or Fiber?: Fibre Channel was originally designed to support fiber optic cabling only. When copper support was added, the committee decided to keep the name in principle but change its spelling from fiber to fibre. When referring to specific cabling, the correct American English spelling of fiber should be used.

© Copyright IBM Corp. 2004. All rights reserved.

2.1 Fibre Channel technology components

Today's hardware designers are required to balance requirements for performance and special features against steadily increasing pressure on design-cycle time, flexibility, and cost. Rapid advances in technology, library design, and design automation have made application specific integrated circuit (ASIC) technology an increasingly viable choice for many applications.

The fabric electronics utilize personalized ASICs and its predefined set of elements, such as logic functions, I/O circuits, memory arrays and backplane to create specialized fabric interface components.

The switch and director electronics consist of a system board and multiple port cards respectively, incorporating the Fibre Channel port interfaces, switching mechanism, embedded switch control processor, and support logic for the embedded processor logic.

An ASIC provides services to Fibre Channel ports that may be used to connect to external N_Ports (such as an F_Port or FL_Port), external loop devices (such as an FL_Port), or to other switches (such as an E_Port). The ASIC contains the Fibre Channel interface logic, message/buffer queuing logic, and receive buffer memory for the on-chip ports, as well as other support logic.

IBM produces custom chip cores designed for consumer, networking and storage, customer applications. Recently IBM expanded its industry-leading family of ASIC technologies from high-end systems to consumer applications.

IBM has three new 130 nanometer (nm) High Speed SERDES (HSS) cores for its Cu-11 application specific integrated circuit (ASIC) family. SERDES, which stands for serial-deserializer, is essentially a circuit technique for rapid transmission of large amounts of structured data. The new cores include:a 6.4Gbps core for switched backplane upgrades; a Fibre Channel Standard core for the newest generation interfaces for storage applications at 4.25 Gbps; and a core for next generation I/O interconnect based on the emerging PCI Express standard.

This new set of IBM HSS cores supports more than a dozen high-speed interface standards and a number of emerging technologies — including a fully compliant PCI Express core, which provides the next generation of processor and component interconnect with an array of interoperable industry platforms. IBM HSS cores also offer full compliance with industry standards and are validated using procedures and methodologies specifically required by those standards.

2.1.1 Fibre Channel transmission rates

The current set of vendor offerings for switches, host bus adapters, and storage devices is standardized on 2.125 Gb/s components. Most hardware can auto-negotiate to 1Gb/s if necessary to support older hardware.

Host bus adapter (HBA) and switch vendors are expected to start offering 4Gbps Fibre Channel products in late 2004 or early 2005. Therefore 4.25 Gbps is an interim speed before the industry takes on the major transition to 10 Gbps Fibre Channel. 4 Gbps Fibre Channel utilizes the same technology as 2 Gbps but is twice as fast, and is a transition technology that gives users more speed at almost the same prices as 2 Gbps products.

End-user adoption of 4 Gbps Fibre Channel will start slowly with users that need more bandwidth. The applications for the higher speed links will be between switches for Inter Switch Links (ISL). Another application that makes 4 Gbps attractive will be in backup. Backup windows are shrinking, and theoretically, 4 Gbps is twice as fast as 2 Gbps. This implies half the elapsed time for a tape backup. Real-time streaming video applications could also benefit from the faster speed.

2.2 Fiber optic interconnects

In Fibre Channel technology, frames are moved from source to destination using gigabit transport, which is a requirement to achieve fast transfer rates. To communicate with gigabit transport, both sides have to support this type of communication. This can be accomplished by installing this feature into the device or by using specially designed interfaces which can convert other communication transport into gigabit transport. Gigabit transport can be used in a copper or fiber optic infrastructure. We recommend that you consider using a fiber optic implementation if you need to avoid the distance limitation of SCSI, or are likely to in the future.

Nowadays, fibre optic implementations are common due to the ease of use and flexibility that it offers. Copper implementations are primarily used within a rack or a frame for many short interconnects.

The interfaces that are used to convert the internal communication transport of gigabit transport are as follows:

- Small Form Factor Pluggable Media(SFP)
- Gigabit Interface Converters (GBIC)
- Gigabit Link Modules (GLM)
- 1x9 transceivers
- Media Interface Adapters (MIA)

We provide a brief description of the types of cables and connectors, and their functions in the following topics.

2.2.1 Small Form Factor Pluggable Media

The most common fibre channel interconnect component in use today is the small form factor pluggable media as shown in Figure 2-1 on page 20. This component is hot pluggable on the I/O module or the HBA, and the cable is also hot-pluggable. The SFP optical trancievers can use short or long-wavelength lasers.

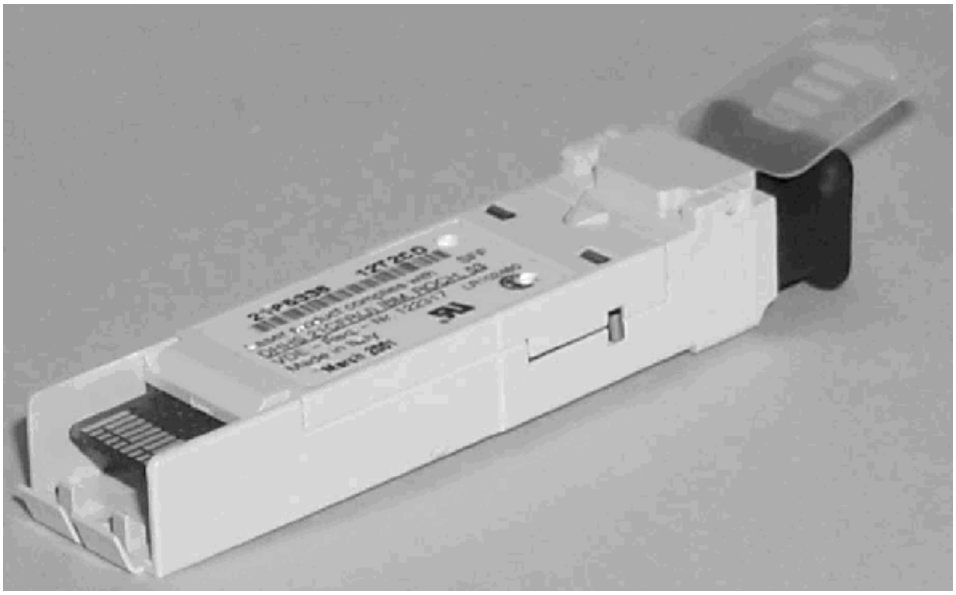


Figure 2-1 SFP Hot Pluggable optical transciever

Another version of the tranceiver are called Small Form Fixed optical transceivers, and are mounted on the I/O module or the HBA via pin-through-hole technology as shown in Figure 2-2 on page 21.The IBM 1063 Mb/s up to 2125 Mb/s Small Form Fixed (SFF) serial optical converters are the an example of these. The transceivers, which are designed for increased densities, performance, and reduced power, are well-suited for Gigabit Ethernet, Fibre Channel, and 1394b applications.

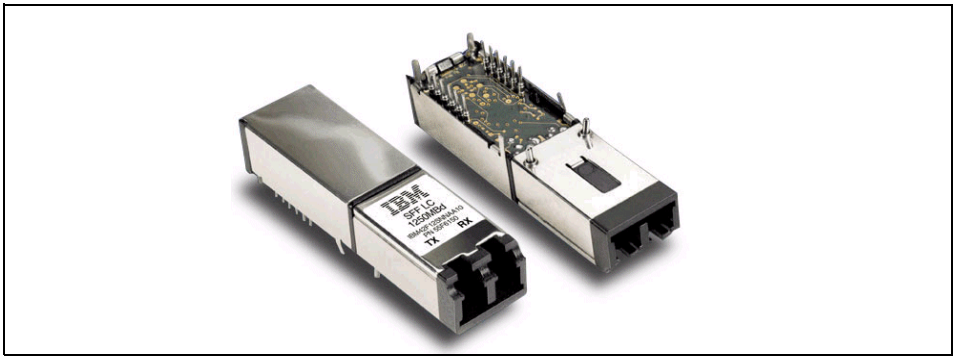


Figure 2-2 Small Form Fixed pin-through-hole Transceiver

The small dimensions of the SFP optical transceivers are ideal in switches and other products where many transceivers have to be configured in a small space. Using these SFP devices, manufacturers can increase the density of transceivers on a board, compared with what was possible using previous optical transceiver technologies. This device is flexible, self-configuring for 100 MB/s or 200 MB/s transmission rates for current or future speeds, providing a seamless transition. Its enhanced design features include frequency agility, reduced power consumption, and lower cost transmission.

The SFP serial optical transceivers are integrated fiber-optic transceivers providing a high-speed serial electrical interface for connecting processors, switches, and peripherals through an optical fiber cable. In the Gigabit Ethernet environment, for example, these transceivers can be used in local area network (LAN) switches or hubs, as well as in interconnecting processors. In SANs, they can be used for transmitting data between peripheral devices and processors.

The SFP optical transceivers use short wavelength and long wavelength lasers that enable cost-effective data transmission over optical fibers at distances of 500 m up to 10 km. They are designed to connect easily to a system card through an industry-standard connector. Multi-mode optical fiber cables, terminated with industry-standard LC connectors, can be used as illustrated in Figure 2-3.

Figure 2-3 SFF hot-pluggable transceiver (SFP) with LC connector fiber cable

The distances that can be achieved using the SFP short wavelength and long wavelength are listed in Table 2-1.

Table 2-1 Distances using SFP-based fiber optics

Type of fiber	SWL	LWL
9/125 µm Optical Fiber	n/a	• Distance for 1.250 Gb/s version: 2 m - 5 km • Distance for 2.125 Gb/s version: 2 m - 10 km • Distance for 4.125Gb/s: 2m - 10km • Distance for10 Gb/s: 2m - 10km
50/125 µm Optical Fiber	• Distance for 1.0625 Gb/s version: 2 - 500 m • Distance for 1.250 Gb/s version: 2 - 550 m • Distance for 2.125 Gb/s version: 2 - 300 m • Distance for 4.250 Gb/s: 2- 150m • Distance for 10G.bs: 2-55m	2 - 550 m

Type of fiber	SWL	LWL
62.5/125 µm Optical Fiber	• Distance for 1.0625 Gb/s version: 2 - 300 m • Distance for 1.250 Gb/s version: 2 - 275 m • Distance for 2.125 Gb/s version: 2 - 150 m	2 - 550 m

The distances shown are not necessarily the supported distances, and this will have to be verified with the fiber transport installer.

There are also some especially high-powered laser versions of SFPs to provide extended distance. For example, the CNT SAN256N and FC/9000 directors offer f/c 2235 Long Wave 2 Gbps SFPs that provide up to 35 Km links, or f/c 2280 which extends the link to 80 km. These must be used within CNT switches because dedicated port buffer credits are coordinated with these SFPs to help users extend FCP and FICON Inter Switch Links (ISLs) up to 35km or 80km distance, in campus environments or metropolitan area networks.

The McDATA 6140 director offers special universal port modules f/c 3235 with higher powered lasers to extend long-wave distance capabilities to 35km. The IBM TotalStorage 2109-M14 also offers f/c 2380 extended distance long-wave SFPtransceivers for up to 80 km links. Cisco also offers special SFPs that can extend the distance to 35 kms..

SFP: The Small Form Factor Hot-Pluggable module is known as SFP. The IBM SFF optical transceiver is offered with an LC connector due to its robustness, increased transmission performance, better EMI, virtually no optical or electrical cross talk, and less jitter, with the primary objective being to ensure reliability, serviceability, and availability.

An Small Form Fixed is not hot-pluggable, whereas an SFP is.

2.2.2 Gigabit Interface Converters

IBM offers laser-based, hot-pluggable, data communications transceivers for a wide range of networking applications requiring high data rates. The transceivers, which are designed for ease of configuration and replacement, are well-suited for Gigabit Ethernet, Fibre Channel, and 1394b applications. Current GBICs are capable of 2125 Mb/s transmission rates. Older technologies offered 1063 Mb/s and 1250 Mb/s speeds. GBICs are available in both short wavelength and long wavelength versions, providing configuration flexibility. Users can easily add a GBIC in the field to accommodate a new configuration requirement or replace an existing device to allow for increased availability.

GBICs are integrated fiber-optic transceivers providing a high-speed serial electrical interface for connecting processors, switches, and peripherals through an optical fiber cable. In the Gigabit Ethernet environment, for example, these transceivers can be used in local area network (LAN) switches or hubs, as well as in interconnecting processors. In SANs they can be used for transmitting data between peripheral devices and processors.

The GBICs use lasers that enable cost-effective data transmission over optical fibers at distances of up to 10 km. These compact, hot-pluggable, field-replaceable modules are designed to connect easily to a system card through an industry-standard connector. Single-mode or multi-mode optical fiber cables, terminated with industry-standard SC connectors, can be used.

Most GBICS available now are 2125 Mb/s and 2500 Mb/s GBICs (referred to below as 2 Gb/s GBIC). We have included below the distance capabilities for each of the different types of GBICs.

There are two types of connections on the service side, namely, DB-9 and HSSDC. The fiber-optic type has two modes of operation. The difference between them is in the laser wave length. The two modes are:

- Short wavelength mode: SWL
- Long wavelength mode: LWL

The distances that can be achieved using both the 1 Gb/s-based GBICs are shown in Table 2-2.

Table 2-2 Distance using 1 Gb/s GBIC based fiber optics

Type of fiber	SWL	LWL
9/125 µm Optical Fiber	n/a	10 km
50/125 µm Optical Fiber	2 - 550 m	2 - 550 m
62.5/125 µm Optical Fiber	2 - 300 m	2 - 550 m

The distances that can be achieved using both the 2 Gb/s based GBICs are shown in Table 2-3.

Table 2-3 Distance using 2 Gb/s GBIC based fiber-optics

Type of fiber	SWL	LWL
9/125 µm Optical Fiber	n/a	10 km
50/125 µm Optical Fiber	2 - 300 m	2 - 550 m
62.5/125 µm Optical Fiber	2 - 150 m	2 - 550 m

The standard dual SC plug is used to connect to the fiber optic cable. This is shown in Figure 2-4.

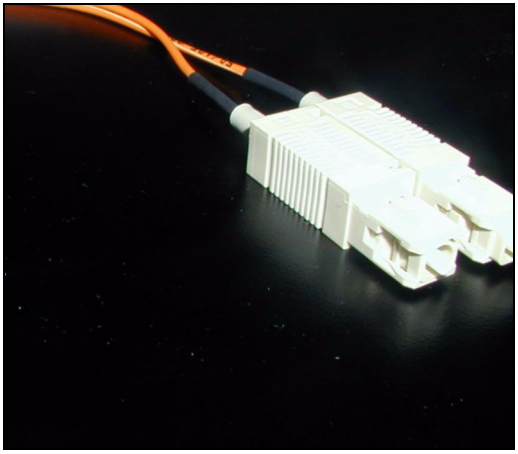


Figure 2-4 Dual SC fiber-optic plug connector

GBICs are usually hot-pluggable, easy to configure and replace. On the optical side they use low-loss, SC type, push-pull, optical connectors. They are mainly used in hubs, switches, directors, and gateways.

Shortwave (or multi-mode) GBICs are usually color coded beige with a black exposed surface; and longwave (or single-mode) GBICs are usually color coded blue with blue exposed surfaces.

The transfer rates typically range from 1063 Mb/s,1250 Mb/s, 2125 Mb/s, or 2500 Mb/s. A GBIC is shown in Figure 2-5.

Figure 2-5 Gigabit Interface Converter

The selection of a GBIC for SAN interconnection is just as important a consideration as choosing a hub or a switch, and should not be overlooked or taken lightly.

2.2.3 Gigabit Link Modules

Gigabit Link Modules (GLMs) — sometimes referred to as Gigabaud Link Modules) — were used in early Fibre Channel applications. GLMs are a low cost alternative to GBICs, but they sacrifice the ease of use and hot-pluggable installation and replacement characteristics that GBICs offer. This means that you need to power down the device for maintenance, replacement, or repair.

GLMs enable computer manufacturers to integrate low-cost, high-speed fiber optic communications into devices. They use the same fiber optic for the transport of optical signal as GBICs. GLMs also use two types of lasers, SWL and LWL, to transport the information across the fiber optic channel. The transfer rates that are available are 266 Mb/s and 1063 Mb/s.

The 266 Mb/s and 1063 Mb/s GLM cards support continuous, full-duplex communication. The GLM converts encoded data that has been serialized into pulses of laser light for transmission into the optical fiber. A GLM at a second optical link, running at the same speed as the sending GLM, receives these pulses, along with the requisite synchronous clocking signals.

With 1063 Mb/s you can achieve the distances listed in Table 2-2, “Distance using 1 Gb/s GBIC based fiber optics” on page 24.

A GLM is shown in Figure 2-6.

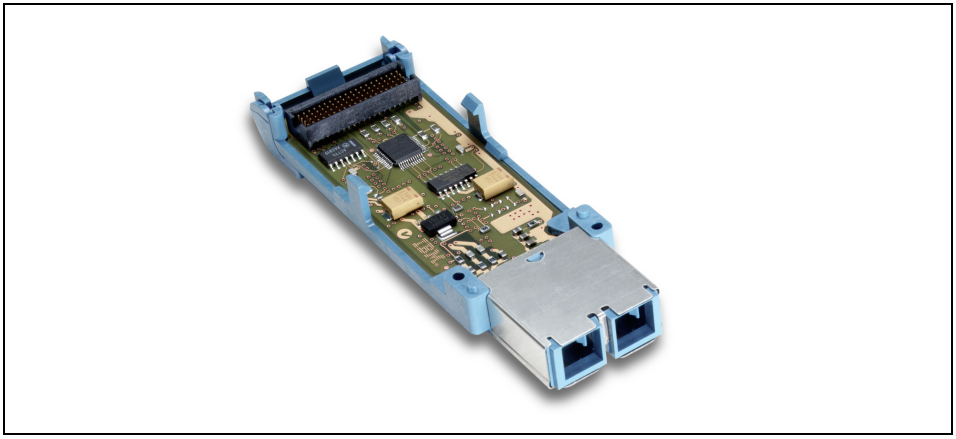


Figure 2-6 Gigabit Link Module

2.2.4 Media Interface Adapters

Media Interface Adapters (MIAs) can be used to facilitate conversion between optical and copper interface connections. Typically, MIAs are attached to host bus adapters, but they can also be used with switches and hubs. If a hub or switch only supports copper or optical connections, MIAs can be used to convert the signal to the appropriate media type, copper or optical.

An MIA is shown in Figure 2-7.

Figure 2-7 Media Interface Adapter

2.2.5 1x9 transceivers

Some switch manufacturers prefer to use 1x9 transceivers for providing SC connection to their devices. 1x9 transceivers (Figure 2-8) have some advantages over GBICs:

- Easier to cool
- Better air flow
- More reliable (2.5 times that of a GBIC)

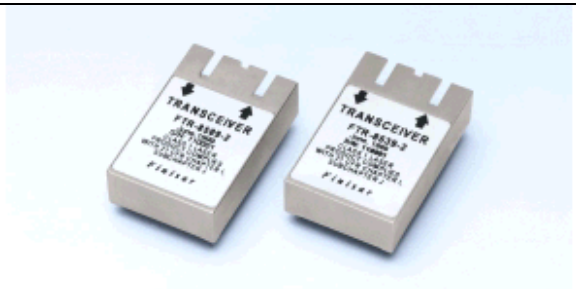


Figure 2-8 1x9 transceivers

2.2.6 Fibre Channel adapter cable

The LC-SC adapter cable attaches to the end of an LC-LC cable to support SC device connections. A combination of one LC/LC fiber cable and one LC/SC adapter cable is required for each connection. This is used to connect from some of the older 1 Gb/s devices to a 2 Gb/s capable and LC interface-based SAN.

Shown in Figure 2-9 is a Fibre Channel adapter cable.

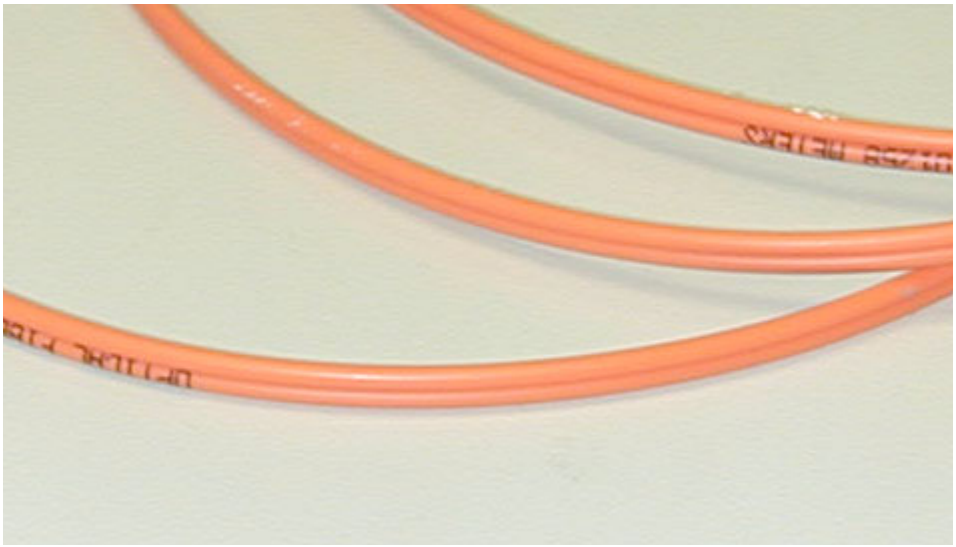


Figure 2-9 Fibre Channel adapter cable

2.2.7 Host Bus Adapters

The device that acts as the interface between the fabric of a SAN and either a host or a storage device is a Host Bus Adapter (HBA). In the case of storage devices, they are often just referred to as Host Adapters.

The HBA connects to the bus of the host or storage system. It has some means of connecting to the cable or fiber leading to the SAN. Some devices offer more than one Fibre Channel connection. At the time of writing, single and dual ported offerings were available. The function of the HBA is to convert the parallel electrical signals from the bus into a serial signal to pass to the SAN.

Some typical HBAs are shown in Figure 2-10.

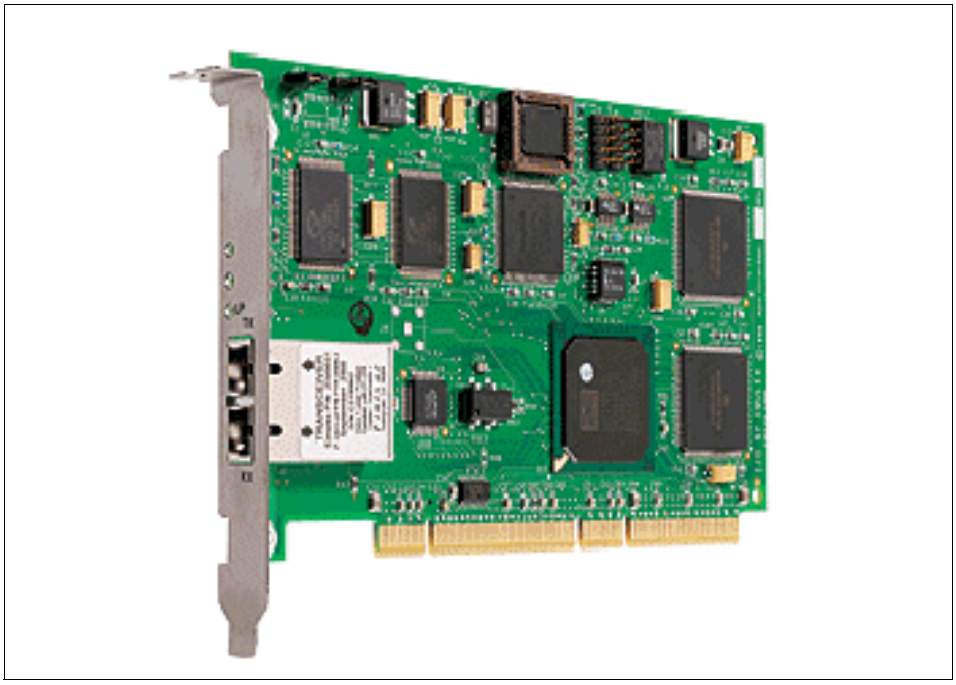


Figure 2-10 Host Bus Adapter

Various cables may be supported by the HBAs, for example:

- **Glass fiber**
  - Single mode
  - Multi mode
- **Copper**
  - Twisted pair
  - Co-axial.

There are several manufacturers of HBAs and an important consideration when planning a SAN, is the choice of HBAs. Some HBAs may have interoperability problems with some other FIbre Channel components.

A server or storage device may have one HBA or it may have many. Depending upon the particular configuration of the SAN, if there are more than one, they might all be identical, or they could be of different types.

The adapters in storage arrays are usually determined by the manufacturer. Factors influencing the choice of HBAs in servers are dealt with in Chapter 4, “SAN design considerations” on page 221.

2.2.8 Bridges and SAN Data Gateways

A bridge is a device which converts signals and data from one form to another. In the specific case of a SAN, a bridge is a unit converting between Fibre Channel and legacy storage protocols such as:

- SCSI
- SSA

This allows the architect of the SAN to protect the investment made in legacy equipment. Another benefit is that some of the limitations of SCSI can be overcome.

Those limitations include:

- 25 m cable length
  - Fibre Channel allows SCSI devices connected to bridges to be hundreds of meters away.
- 16 device maximum addressability

The offerings from IBM are called SAN Data Gateways or SAN Data Gateway Routers, but the generic terms of *bridge* or *router* are commonly used.

Depending on the particular bridge, it may be possible to have only a single Fibre Channel port, whereas some will support multiple ports. It is usual for bridges to have multiple legacy connections or busses.

2.2.9 Arbitrated Loop hubs

There are various ways to connect Fibre Channel devices together using different topologies. If the selected topology is Arbitrated Loop (FC-AL), then the devices can either be cabled in a loop by using discreet cabling, or an Arbitrated Loop hub can be used.

In FC-AL all devices on the loop share the bandwidth. The total number of devices which may participate in the loop is 126. For practical reasons, however, the number tends to be limited to no more than 10 and 15.

Due to the limitations of FC-AL, it is not normal to build a SAN just around hubs. Dependant upon various factors covered in Chapter 4, “SAN design considerations” on page 221, it is possible to attach a hub to a switched fabric. This allows devices which do not support the switched topology to be utilized in a large SAN.

Hubs are typically used in a SAN to attach devices or servers which do not support switched fabrics only FC-AL. They may be either unmanaged or managed.

Unmanaged hubs

The Fibre Channel hub serves as a cable concentrator and as a means to configure the Arbitrated Loop based on the connections it detects. When any of the hub's interfaces senses no cable connected, that interface shuts down and the hub port is bypassed as part of the Arbitrated Loop configuration.

The hub allows devices to be hot plugged into or out of an existing loop. The hub will ensure that the integrity of the loop is maintained at all times. I/Os may be lost, however, if an active connection is removed without taking some action at the OS and/or application level.

Managed hubs

Managed hubs offer all the features of unmanaged hubs, but in addition, offer the ability to manage them remotely. This is done by built in firmware, usually communicating to a management system over an ethernet network.

2.2.10 Switched hubs

There is also a device called a switched hub. In this case each of the attached devices is in its own Arbitrated Loop. These loops are then internally connected by a switched fabric.

A switched hub is useful to connect several FC-AL devices together but to allow them to communicate at full Fibre Channel bandwidth rather than them all sharing the bandwidth.

It is usual for switched hubs to be managed rather than unmanaged.

2.2.11 Switches

Switches allow Fibre Channel devices to be connected together, implementing a switched fabric topology between them. Unlike in an Arbitrated Loop hub, where all connected devices share the bandwidth, in a switch, all devices operate, theoretically at full Fibre Channel bandwidth. This is because the switch creates a direct communication path between any two ports which are exchanging data. The switch intelligently routes frames from the initiator to the responder.

Figure 2-11 on page 33 shows an example of a fabric switched environment.

It is possible to connect switches together in cascades and meshes using Inter Switch links or ISLs. It should be noted that devices from different manufacturers may not inter-operate fully.

As well as implementing this switched fabric, the switch also provides a variety of fabric services and features such as:

- Name services
- Fabric control
- Time services
- Automatic discovery and registration of host and storage devices
- Rerouting of frames, if possible, in the event of a port problem
- Storage Services (virtualization, replication, extended distances)

Features which are usually implemented in Fibre Channel switches include:

- Telnet and/or RS-232 interface for management
- HTTP server for Web-based management
- MIB for SNMP monitoring
- Hot swappable, redundant power supplies and cooling devices
- Online replaceable GBICs/interfaces
- Zoning
- Trunking
- Fibre Channel Protocol (FCP)
- Ficon
- Ficon CUP
- iSCSI
- FCIP

Multiple protocol support is offered by most swtich vendors as optional features, including both open systems (FCP) and mainframe (FICON) protocols intermix. Newer technologies like iSCSI and FCIP are gradually being introduced and included into products.

It is common to refer to switches as either *core* switches or *edge* switches depending on where they are located in the SAN. If the switch forms, or is part of the SAN backbone, then it is a core switch. If it is mainly used to connect to hosts or storage then it is called an edge switch. There are certain cases where it is appropriate for storage, servers or both to be connected directly to core switches

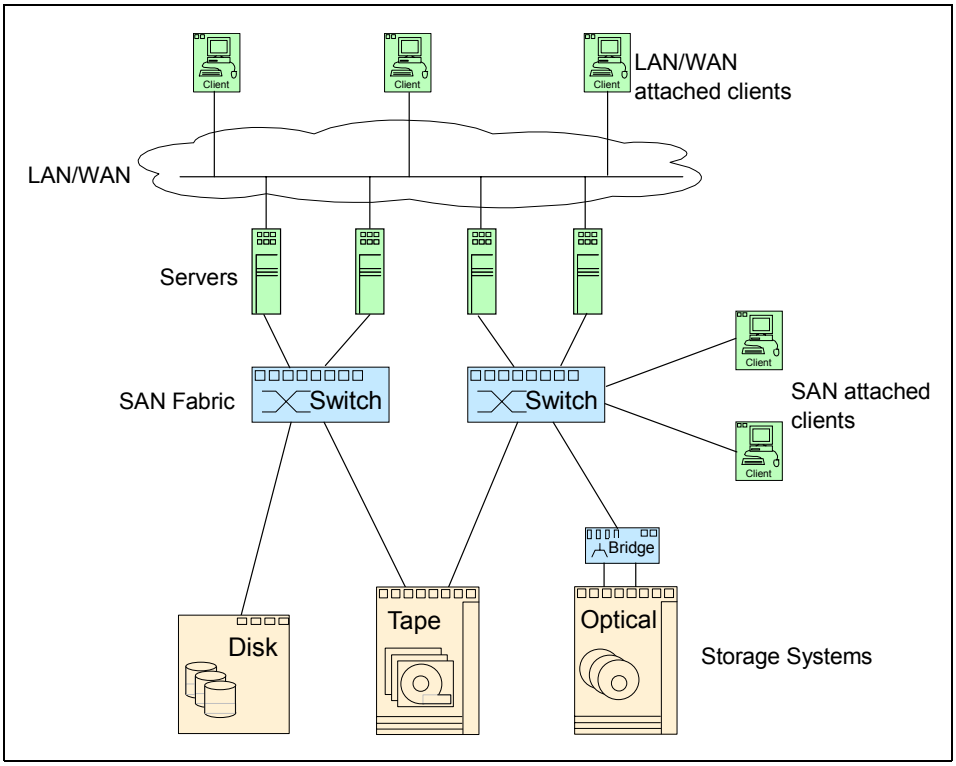


Figure 2-11 Fibre Channel Switch Environment

2.2.12 Core switches

A device which is *designed* as a core switch could be deployed as an edge switch. Commonly, however, devices used as core switches are designed with more resilience and more features than may be needed for edge devices. This is because a core switch will carry data for many edge switches.

Some additional features that *may* be implemented in core switches include:

- Enhanced security features
- Frame filtering, see 3.9.1, “Frame filtering” on page 114
- Backplane and blade based design for ease of expansion
- Potentially 99.999% uptime
- Non disruptive upgrade of firmware
- Redundant components

The demarcation between core switches and directors becomes quite vague at the top end of core switches and the low end of directors. As little as a year ago, the boundary between switches and directors could simply have been seen as the maximum number of ports in a chassis. Now, the boundary is changing and the port counts are getting closer and other factors are the differentiators. This is similar to the way that the boundaries between open systems and mainframes have changed over the last, say, 20 years.

2.2.13 Directors

A director can be thought of as a core switch:

- They implement a switched fabric between multiple ports.
- They have resilience built in by redundancy.
- Field replaceable units can be hot swapped.
- They can be connected to other directors and switches using Inter Switch links.
- They offer additional functions besides switching with plug-in modules to provide storage services like virtualization, replication, and WAN connections..
- Normally they have a non-blocking architecture. Please see 3.1.1, “Blocking” on page 45.

These are some of the significant differences.

Port capacity

First, directors tend to be larger than switches. For example, directors currently on the market can support between eight and 256 ports, whereas the biggest switch available (at the time of writing) is 32 port.

MTBF

Director manufacturers are able to guarantee 99.999% uptime on their devices, whereas switches will tend to be specified at 99.9% uptime. Put another way, this is the difference between about five minutes and about nine hours a year downtime. Of course, the particular guarantees depend on manufacturer and will vary from contract to contract.

Latency

The latency between ports on a director will tend to be significantly lower than the latency between ports on switches that are connected using Inter Switch links. It should be noted that the actual latency between particular pairs of director ports can vary, depending on the architecture of the unit.

Firmware updates

The firmware on a director can be updated on line, whereas when the firmware on a switch is updated, a reboot is generally required. This means that there will need to be an interruption to the service with a switch. In the case of a director, however, although their may be a slight delay for a few Fibre Channel frames, I/O can continue without the application being stopped or timing out.

2.2.14 Director features

It is expected that a director *will* have all of the good features which a switch *might* have.

Switch frame buffering

An important criterion in selecting a switch is the number of frames that can be buffered on the port. During periods of high activity and frame movement, the switch may not be able to transmit a frame to its intended destination. This is true if two ports are sending data to the same destination. Given this situation, but depending on the class of service, the switch may sacrifice the frames it is not able to process. Not only does frame buffering reduce this likelihood, it also enhances performance.

Domain number routing decision

Another great performance improvement can be realized in the way in which the 24-bit port address is built. Because the address is divided into domain, area, and port, it is possible to make the routing decision on a single byte. As one example of this, if the domain number of the destination address indicates that the frame is intended for a different switch, the routing process can forward the frame to the appropriate interconnection without the need to process the entire 24-bit address and the associated overhead.

Data path in switched fabric

A complex switched fabric can be created by interconnecting Fibre Channel switches. Switch-to-switch connections are performed by E_Port connections. This mean that if you want to interconnect switches, they need to support E_Ports. Switches may also support multiple E_Port connections to expand the bandwidth.

In a switched fabric, a “cut-through” switching mechanism is used. This is not unique to switched fabrics and it is also used in Ethernet switches. The function is to speed packet routing from port to port.

When a frame enters the switch, cut-through logic examines only the link level destination ID of the frame. Based on the destination ID, a routing decision is

made, and the frame is switched to the appropriate port by internal routing logic contained in the switch. It is this cut-through which increases performance by reducing the time required to make a routing decision. The reason for this is that the destination ID resides in the first four bytes of the frame header, and this allows the cut-through to be accomplished quickly. A routing decision can be made at the instant the frame enters the switch, without interpretation of anything other than the four bytes.

In such a configuration with interconnected switches, known as a meshed topology, multiple paths from one N_Port to another can exist.

An example of a meshed topology is shown in Figure 2-12.

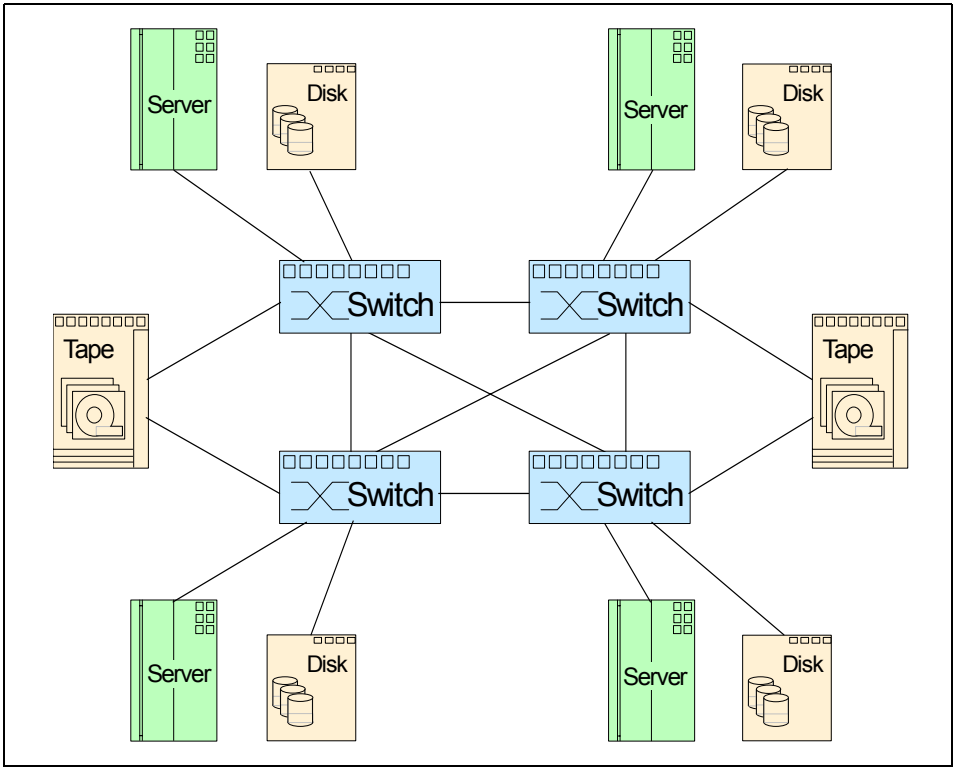


Figure 2-12 Meshed topology switched fabric

Storage services

Many switch vendors are expanding the functionality of the directors and switches by offering modules which plug into a switch slot to offer advanced functions like storage virtualization, storage replication, or SAN extension over extremely long distance.

One example of this is the IBM SAN Volume Controller storage virtualization software running on the CISCO MDS9000 Caching Services Module. This offers storage virtualization within the Cisco switch.

2.2.15 Storage considered as legacy

In the context of a SAN, legacy equipment consists of devices that do not inherently support Fibre Channel. As an example, SCSI disk arrays and tape drives and SSA devices may be considered to be legacy equipment.

In order to protect your investment, it is often a requirement that legacy equipment gets reused or perhaps we should say continues to be used after the implementation of the SAN. After all, a company may have many terabytes of data stored on SCSI or SSA disks.

So, these SCSI and SSA devices need to be connected into the SAN, how are we going to do it?

Well, a bridge, router or gateway device is used to convert between these protocols. These have Fibre Channel connectivity offering connections to the legacy equipment at the same time.

2.2.16 SerDes

The communication over a fiber, whether optical or copper, is serial. Computer busses on the other hand, use parallel busses. This means that Fibre Channel devices need to be able to convert between the two. For this, they use a serializer/deserializer which is commonly referred to as a SerDes.

2.2.17 Backplane and blades

Rather than having a single printed circuit assembly containing all the components in a device, sometimes the design used is that of a *backplane* and *blades*. For example, directors and large core switches are usually implemented using this technology.

The backplane is a circuit board with multiple connectors into which other cards may be plugged. These other cards are usually referred to as blades, but other terms could be used.

If the backplane is in the centre of the unit with blades being plugged in at the back and the front, then it would usually be referred to as a midplane.

We show a backplane and blades architecture diagram in Figure 2-13 on page 38.

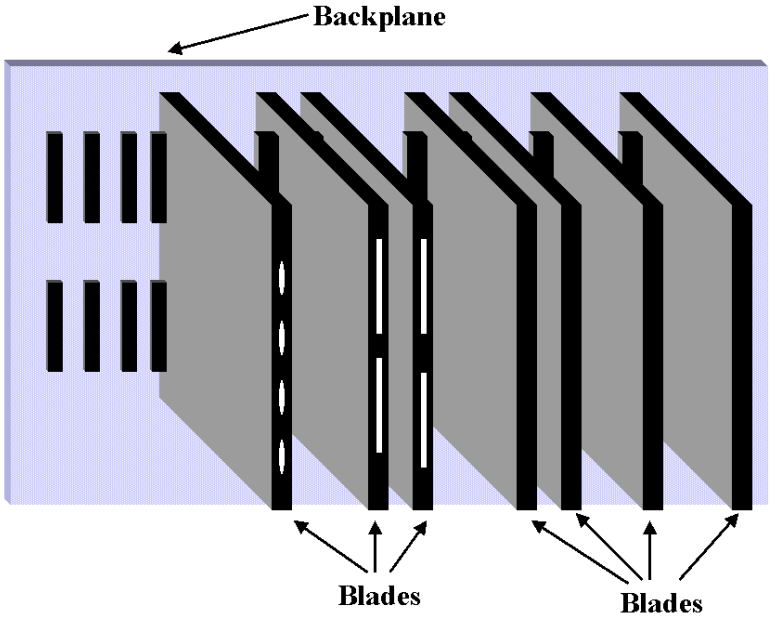


Figure 2-13 A diagram of a backplane and blades architecture

If a backplane has components such as transistors or integrated circuits, then it is an *active* backplane. If it has no components at all, or just passive components such as resistors and capacitors then it is a *passive* backplane.

Some major benefits which may be possible using this design:

- On the fly upgrades by adding extra blades giving additional ports
- On the fly implementation of other functionality, for example new protocol support by adding blades with different functionality
- Potential to have different levels of firmware on different blades
- Passive backplane: Leading to very high level of reliability of the unit as a whole, faults can be isolated to a blade. This is especially true if the backplane has no components, but is just a circuit board with sockets and conductor tracks.

Attention: A product may be described by its manufacturer as *director* class, but if it has a single backplane then this becomes a single point of failure.

It is common practise for mainframe sites to implement duplicate ESCON directors and such companies may consider it necessary to utilize duplicate Fibre Channel directors in a SAN.

At the time of writing, there were no directors or switches available which had a redundant backplane.

2.2.18 Test gear

There are a few different pieces of test gear that might be seen in the realms of Fiber Optics.

GO/NOGO testers

These are simple devices which allow the user to prove that light is passing through the cable. Commonly a laser source which is attached to one end of the fiber. If light reaches the other end, then the fiber is continuous. This is a useful way to quickly identify the two ends of a particular fibre in a bundle, which is routed out of sight. The emerging light can be detected if the loose end of the fiber is placed near a sheet of writing paper.

The laser may well be much higher power than those in use by Fibre Channel devices, for example, they may be Class 3 lasers.

Attention: Class 3 lasers are dangerous and there is a risk to the eyes. Do not look into lasers in test equipment.

Light sources and attenuation meters

GO/NOGO testers do not prove that the quality of the fiber is high enough for reliable communication. Specialized light sources and attenuation or power meters can be used to validate short distance cables.

The exact method of using the test equipment will depend on the test gear itself. The result is that the tester will be able to determine the attenuation along the fiber (usually measured in dB (decibels).

This test is considerably more time consuming than the GO/NOGO test.

The equipment needs to be regularly calibrated by an authorized agency in order to be sure that the results are accurate.

Optical Time Domain Reflectometer

An Optical Time Domain Reflectometer (OTDR) is an expensive piece of test gear. The cost can be in the tens of thousands of US dollars.

It is used to investigate the quality of long fiber optic cables, maybe as long as hundreds or even thousands of kilometers.

The OTDR sends out a pulse of light along the fibre and looks for reflections back. There will be a reflection:

- At the point where the fiber is plugged into the OTDR
- At the end of the fiber or a break
- Any splices in the fiber
- Sharp bends
- Damage to the fiber

When long distance fibers are laid, it is usual practice for them to be tested using an OTDR.

The device creates a trace of where the backscatter or reflections take place. This is either displayed on a screen or printed (or both). The trace shows time or distance (directly proportional) on the horizontal axis and power on the vertical axis.

Fibre Channel analyzer

In much the same way as there are network analyzers for looking at traffic going over a network, ethernet for example, there are similar devices for Fibre Channel.

Ethernet analyzers are quite common these days, however their Fibre Channel counterparts are less common and there are a few points to be made about them:

- Ethernet analyzers (twisted pair) can be connected to the network without disruption, and can analyze all data on the network. Fibre Channel analyzers are placed in a Fibre Channel link. They monitor frames going through the link not all data on the fabric. The insertion of the analyzer is disruptive. It is connected by unplugging the link, and inserting the analyzer into the link. This is shown in Figure 2-14.

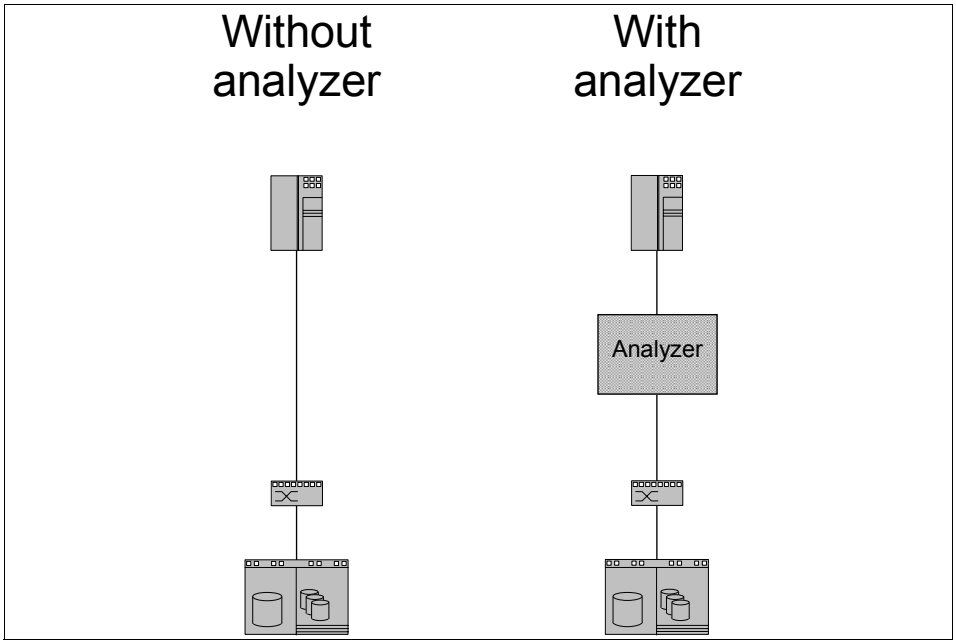


Figure 2-14 Connecting an FC analyzer

- Manufacturers are integrating FC analyzers into their directors. They utilize mirror ports that can be configured to repeat another port's transmit or recieve stream so an external analyser can be attached. Also, they have built-in capture buffers that can capture the FC data or state machines. This trace can be looked at without an external analyzer.
- The final thing is the cost. Fibre Channel analyzers are a great deal more expensive than ethernet analyzers.

For these reasons, it is unlikely that end users will have a Fibre Channel analyzer.

Additional trace facilities

Many switch vendors are providing integrated trace capability within the switch. This makes fabric debug much easier and less expensive.

One example of this is the CISCO SPAN port design. The SPAN feature is specific to switches in the Cisco MDS 9000 family. It monitors network traffic that passes through a Fibre Channel interface. This monitoring is done using a standard Fibre Channel analyzer (or a similar switch probe) that is attached to a SPAN port called SD_Port. SD_Ports do not receive frames, they merely transmit a copy of the source traffic. The SPAN feature is non-intrusive and does

not affect switching of network traffic for any SPAN source ports. The Cisco MDS 9000 Family Port Analyzer Adapter can be used to enable advanced debug and performance analysis of MDS 9000 fabrics. The Cisco MDS 9000 Family Port Analyzer Adapter encapsulates the Fibre Channel frames coming from the MDS 9000 SPAN port and delivers them to standard ethernet frames over a 1000base-T ethernet interface. The frames can then be analyzed using the free ethereal network protocol analyzer software. This allows cost-effective monitoring of the Fibre Channel traffic.

Another example of the newer intelligent trace facilities being offered is from the CNT Ultranet Multi-service Director. The SAN256N offers mirror ports which can be traced via a standard FC analyzer. The data is also captured within an internal buffer in the switch, and can be transferred to a PC for analysis with the Finisar GTX viewer. Multiple levels of trace can be captured, ie at the state machine level versus the Fibre Channel I/O.

3

SAN features

In this chapter we discuss some terminology and concepts that are derived from the Fibre Channel standards and frequently found in SAN device specifications and installations.

We also overview some common features and characteristics of the SAN environment, such as distance, applications, and the different platforms that can benefit from a SAN implementation.

© Copyright IBM Corp. 2004. All rights reserved.

3.1 Fabric implementation

We can build a SAN with a single switch and attached devices. However, as our fabric expands, we will eventually run out of ports. One possible solution is to move to a bigger switch or director, and another solution is to interconnect switches together to build a larger fabric. Another reason that we may need to interconnect switches or directors is to cover longer distances, for example, a building-to-building interconnection for backup and disaster recovery.

Note: It is not unusual to see directors referred to as switches. This is a statement as to the architecture that is employed *within* the director. That is to say, the director adheres to the Fibre Channel Switched Fabric (FC-SW) standard and employs the same switching protocol as a switch. There is no Fibre Channel “Director” Fabric standard! In this redbook, where something does not apply to both switches and directors equally, we make this distinction clear.

The diagram in Figure 3-1 shows two cascaded directors located at two different sites that can be up to 10 km apart. In this way all four servers can connect to both ESS devices.

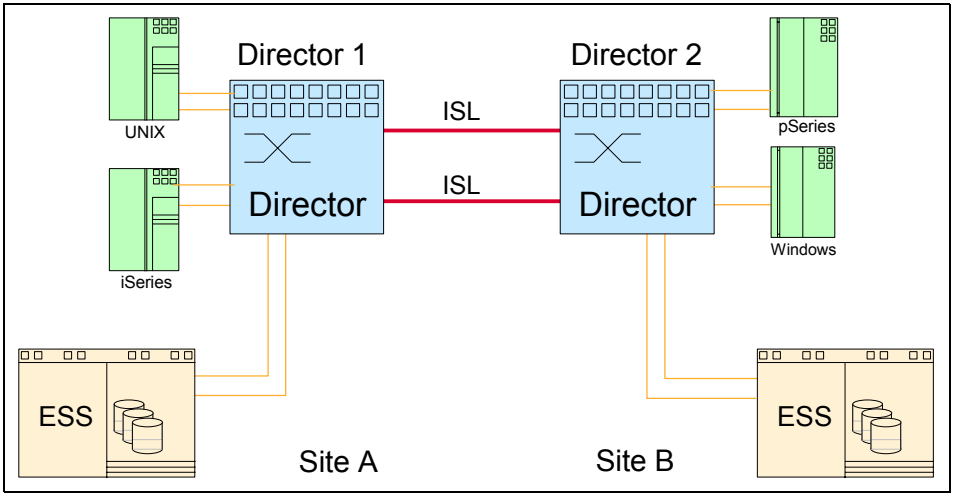


Figure 3-1 Cascading directors

3.1.1 Blocking

To support highly performing fabrics, the fabric components (switches or directors) must be able to move data around without any impact to other ports, targets, or initiators that are on the same fabric. If the internal structure of a switch or director cannot do so without impact, we end up with blocking.

As the fabric components do not typically read the data that they are transmitting or transferring, this means that as data is being received, data is being transmitted too. What this means is that because the potential can be as much as 1000 MB/s bandwidth for each direction of the communication, a fabric component will need to be able to support this. So that data does not get delayed within the SAN fabric component itself, switches, directors, and hubs may employ a non-blocking switching architecture. Non-blocking switches provide for multiple connections travelling through the internal components of the switch concurrently.

Blocking means that the fibre channel data does not get to the destination. This is opposed to congestion, where data will still be delivered, albeit with a delay. Switches and directors may employ a non-blocking switching architecture. Non-blocking switches and directors are the Ferraris on the SAN racetrack.

\We illustrate this concept in Figure 3-2.

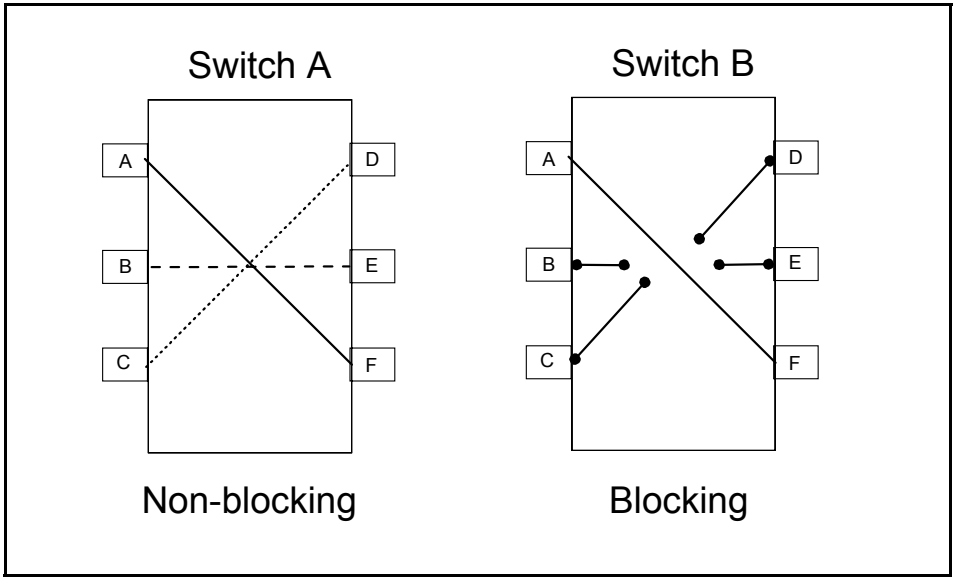


Figure 3-2 Non-blocking and blocking switching

In this example, non-blocking Switch A, port A speaks to port F, B speaks to E, and C speaks to D without any form of suspension of communication or delay; that is to say, the communication is not blocked. In the blocking Switch B, while port A is speaking to F, all other communication has been stopped or blocked.

3.1.2 Ports

The basic building block of fibre channel is the port. The following lists the various types of fibre channel port types and their purpose in switches, servers, and storage.

These are the types of Fibre Channel ports that are likely to be encountered:

- **E_Port:** This is an expansion port. A port is designated an E_Port when it is used as an inter switch expansion port (ISL) to connect to the E_Port of another switch, to enlarge the switch fabric.
- **F_Port:** This is a fabric port that is not loop capable. It is used to connect an N_Port point-to-point to a switch.
- **FL_Port:** This is a fabric port that is loop capable. It is used to connect NL_Ports to the switch in a public loop configuration.
- **G_Port:** This is a generic port that can operate as either an E_Port or an F_Port. A port is defined as a G_Port after it is connected but has not received response to *loop* initialization or has not yet completed the *link* initialization procedure with the adjacent Fiber Channel device.
- **L_Port:** This is a loop capable node or switch port.
- **U_Port:** This is a universal port. A more generic switch port than a G_Port. It can operate as either an E_Port, F_Port, or FL_Port. A port is defined as a U_Port when it is not connected or has not yet assumed a specific function in the fabric.
- **N_Port:** This is a node port that is not loop capable. It is used to connect an equipment port to the fabric.
- **NL_Port:** This is a node port that is loop capable. It is used to connect an equipment port to the fabric in a loop configuration through an L_Port or FL_Port.
- **MTx_Port:** CNT port used as a mirror for viewing the transmit stream of the port to be diagnosed.
- **MRx_Port:** CNT port used as a mirror for viewing the recieve stream of the port to be diagnosed.
- **SD_Port:** Cisco SPAN port used for mirroring another port for diagnostic purposes.

- **T_Port:** This was used previously by CNT as a mechanism of connecting directors together. This has been largely replaced by the E_Port
- **TL_Port:** This is a private to public bridging of switches or directors (the Cisco Systems MDS 9000 Family also uses translative loop ports).

The ports of a switch that connect to the devices N_Ports are called F_Ports. Coupling switches together introduces a new kind of connection, switch to switch. The port at which frames pass between switches within the fabric is called an E_Port.

A switch port will typically support one or more of the following Port Modes:

- F_Port (defined in the FC-PH standard)
- FL_Port (arbitrated loop connection defined in the FC-AL standard)
- E_Port (defined in the FC-SW standard).

A switch that only provides F_Ports and FL_Ports forms a non-expandable fabric. In order to be part of an expandable fabric the switch must have at least one port capable of E_Port operation.

A switch port that has the capability to support more than one port mode attempts to configure itself first as an FL_Port, then as an E_Port and finally as an F_Port, depending on which of the three modes are supported and the port it is connected to.

Switch ports that support both F_Port and E_Port modes are called G_Ports.

Figure 3-3 on page 48 represents some of the most commonly encountered Fibre Channel port types.

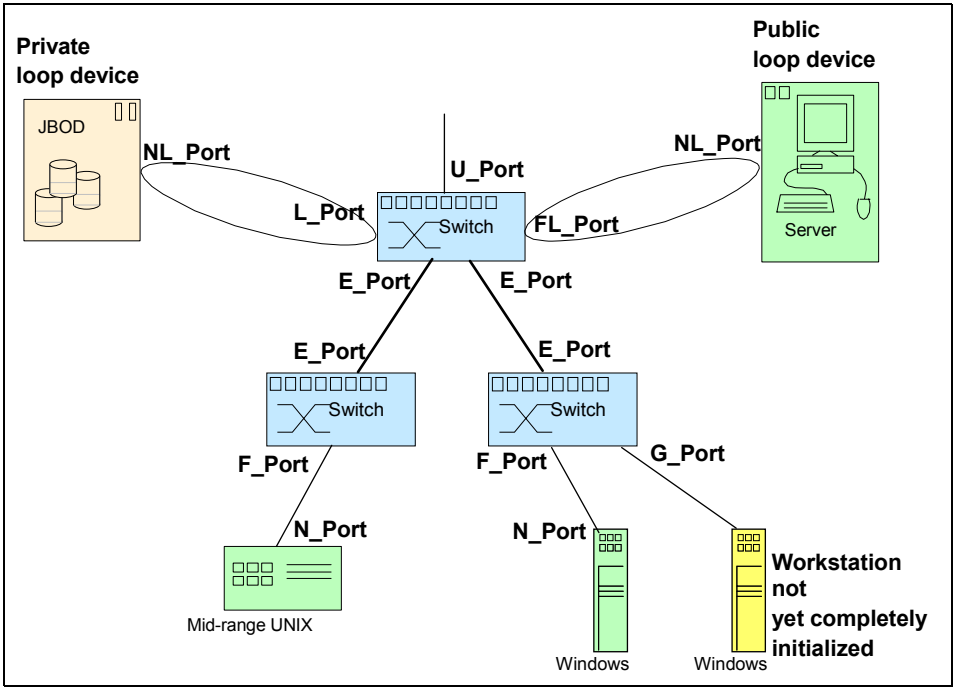


Figure 3-3 Fibre Channel port types.

3.1.3 SAN topologies

Fibre Channel provides three distinct interconnection topologies. By having more than one interconnection option available, a particular application can choose the topology that is best suited to its requirements. The three Fibre Channel topologies are:

- Point-to-point
- Arbitrated loop
- Switched fabric

We discuss these in greater detail in the topics that follow.

3.1.4 Point-to-point

A point-to-point connection is the simplest topology. It is used when there are exactly two nodes, and future expansion is not predicted. There is no sharing of the media, which allows the devices to use the total bandwidth of the link. A simple link initialization is needed before communications can begin.

Fibre Channel is a full duplex protocol, which means both paths transmit data simultaneously. Fibre Channel connections based on the 1 Gb standard are able to transmit at 100 MB/s and receive at 100 MB/s simultaneously. For Fibre Channel connections based on the 2 Gb standard, they can transmit at 200 MB/s and receive at 200 MB/s simultaneously. This will extend to 4.25 Gb/s and 10 Gb/s technologies as well.

Illustrated in Figure 3-4 is a simple point-to-point connection.

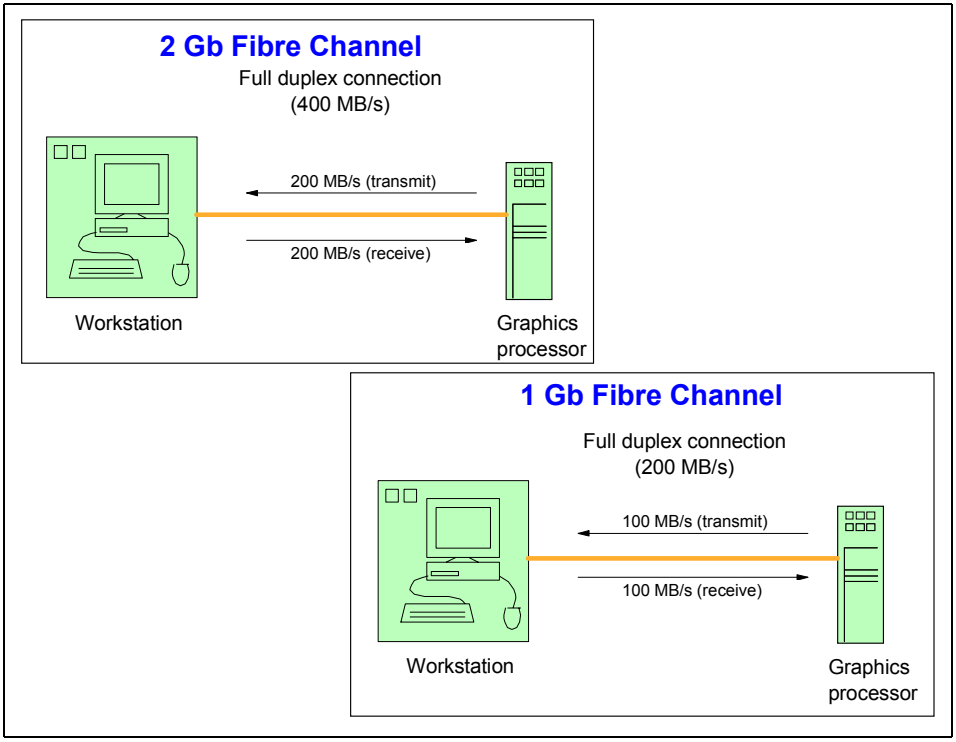


Figure 3-4 Point-to-point

3.1.5 Arbitrated loop

The second topology is Fibre Channel Arbitrated Loop (FC-AL). FC-AL is more useful for storage applications. It is a loop of up to 126 nodes (NL_Ports) that is managed as a shared bus. Traffic flows in one direction, carrying data frames and primitives around the loop with a total bandwidth of 200 MB/s (or 100 MB/s for a loop based on 1 Gb/s technology).

Using arbitration protocol, a single connection is established between a sender and a receiver, and a data frame is transferred around the loop. When the communication comes to an end between the two connected ports, the loop becomes available for arbitration and a new connection may be established. Loops can be configured with hubs to make connection management easier. A distance of up to 10 km is supported by the Fibre Channel standard for both of these configurations. However, latency on the arbitrated loop configuration is affected by the loop size.

A simple loop, configured using a hub, is shown in Figure 3-5.

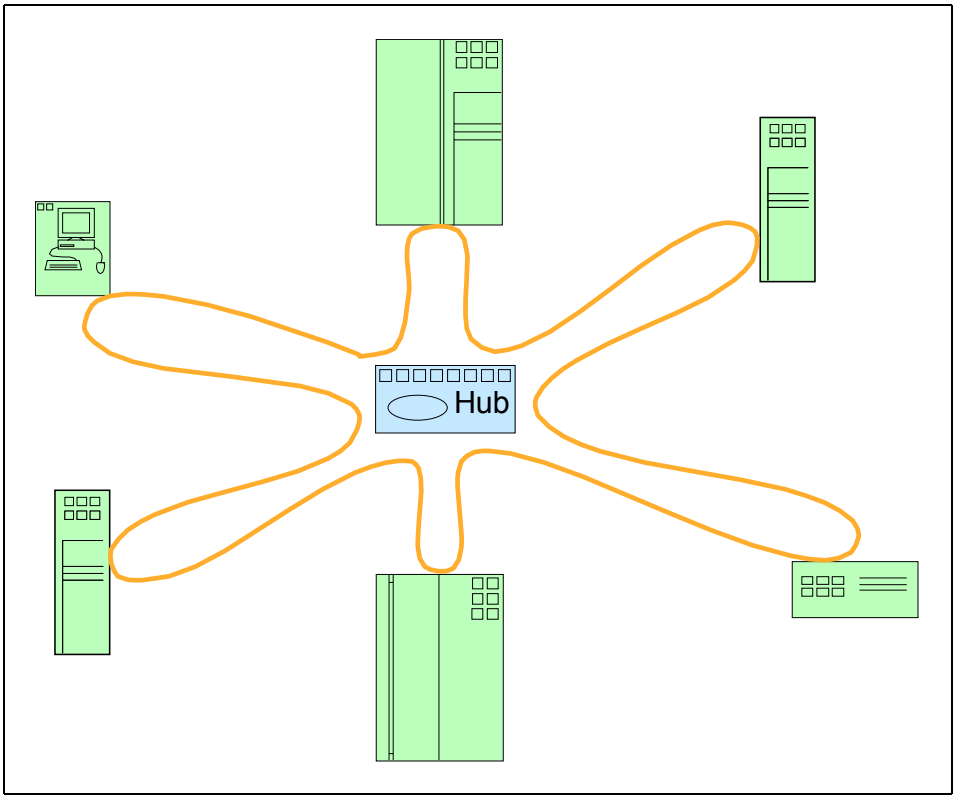


Figure 3-5 Arbitrated loop

Loop protocols

To support the shared behavior of the arbitrated loop, a number of loop-specific protocols are used. These protocols are used to:

- Initialize the loop and assign addresses.
- Arbitrate for access to the loop.
- Open a loop circuit with another port in the loop.
- Close a loop circuit when two ports have completed their current use of the loop.
- Implement the access fairness mechanism to ensure that each port has an opportunity to access the loop.

Loop initialization

Loop initialization is a necessary process for the introduction of new participants on to the loop. Whenever a loop port is powered on or initialized, it executes the loop initialization primitive (LIP) to perform loop initialization. Optionally, loop initialization may build a positional map of all the ports on the loop. The positional map provides a count of the number of ports on the loop, their addresses and their position relative to the loop initialization master.

Following loop initialization, the loop enters a stable monitoring mode and begins with normal activity. An entire loop initialization sequence may take only a few milliseconds, depending on the number of NL_Ports attached to the loop. Loop initialization may be started by a number of causes. One of the most likely reasons for loop initialization is the introduction of a new device. For instance, an active device may be moved from one hub port to another hub port, or a device that has been powered on could re-enter the loop.

A variety of ordered sets have been defined to take into account the conditions that an NL_Port may sense as it starts the initialization process. These ordered sets are sent continuously while a particular condition or state exists. As part of the initialization process, loop initialization primitive sequences (referred to collectively as LIPs) are issued. As an example, an NL_Port must issue at least three identical ordered sets to start initialization. An ordered set transmission word always begins with the special character K28.5.

Once these identical ordered sets have been sent, and as each downstream device receives the LIP stream, devices enter a state known as open-init. This causes the suspension of any current operation and enables the device for the loop initialization procedure. LIPs are forwarded around the loop until all NL_Ports are in an open-init condition.

At this point, the NL_Ports need to be managed. In contrast to a Token-Ring, the Arbitrated Loop has no permanent master to manage the topology.

Therefore, loop initialization provides a selection process to determine which device will be the temporary loop master. After the loop master is chosen it assumes the responsibility for directing or managing the rest of the initialization procedure. The loop master also has the responsibility for closing the loop and returning it to normal operation.

Selecting the loop master is carried out by a subroutine known as the Loop Initialization Select Master (LISM) procedure. A loop device can be considered for temporary master by continuously issuing LISM frames that contain a port type identifier and a 64-bit World-Wide Name. For FL_Ports the identifier is x'00' and for NL_Ports it is x'EF'.

When a downstream port receives a LISM frame from a upstream partner, the device will check the port type identifier. If the identifier indicates an NL_Port, the downstream device will compare the WWN in the LISM frame to its own. The WWN with the lowest numeric value has priority. If the received frame's WWN indicates a higher priority, that is to say it has a lower numeric value, the device stops its LISM broadcast and starts transmitting the received LISM. Had the received frame been of a lower priority, the receiver would have thrown it away and continued broadcasting its own.

At some stage in proceedings, a node will receive its own LISM frame, which indicates that it has the highest priority, and succession to the throne of temporary loop master has taken place. This node will then issue a special ordered set to indicate to the others that a temporary master has been selected.

Hub cascading

Since an arbitrated loop hub supplies a limited number of ports, building larger loops may require linking another hub. This is called hub cascading. A server with an FC-AL, shortwave, host bus adapter can connect to an FC-AL hub 500 m away. Each port on the hub can connect to an FC-AL device up to 500 m away. Cascaded hubs use one port on each hub for the hub-to-hub connection and this increases the potential distance between nodes in the loop by an additional 500 m. In this topology the overall distance is 1500 m. Both hubs can support other FC-AL devices at their physical locations. Stated distances assume a 50 micron multi-mode cable.

Loops

There are two different kinds of loops: private and public.

- **Private loop:** The private loop does not connect with a fabric, only to other private nodes using attachment points called NL_Ports. A private loop is enclosed and known only to itself. In Figure 3-6 we show a private loop.

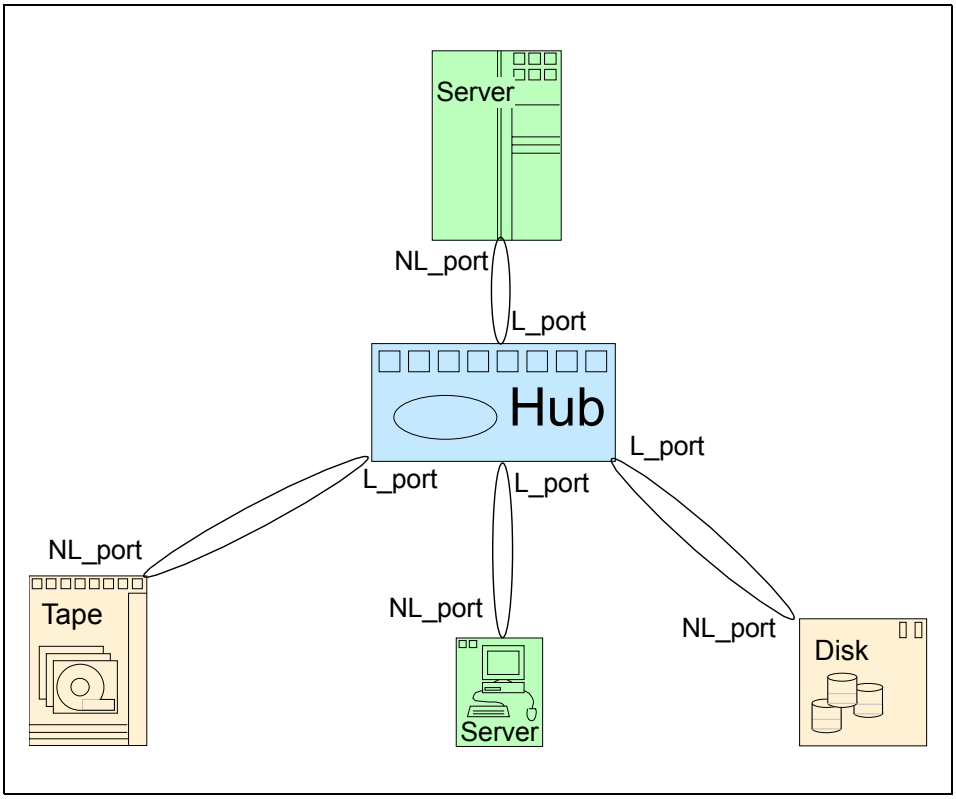


Figure 3-6 Private loop implementation

- **Public loop:** A public loop requires a fabric and has at least one FL_Port connection to a fabric. A public loop extends the reach of the loop topology by attaching the loop to a fabric. Figure 3-7 shows a public loop.

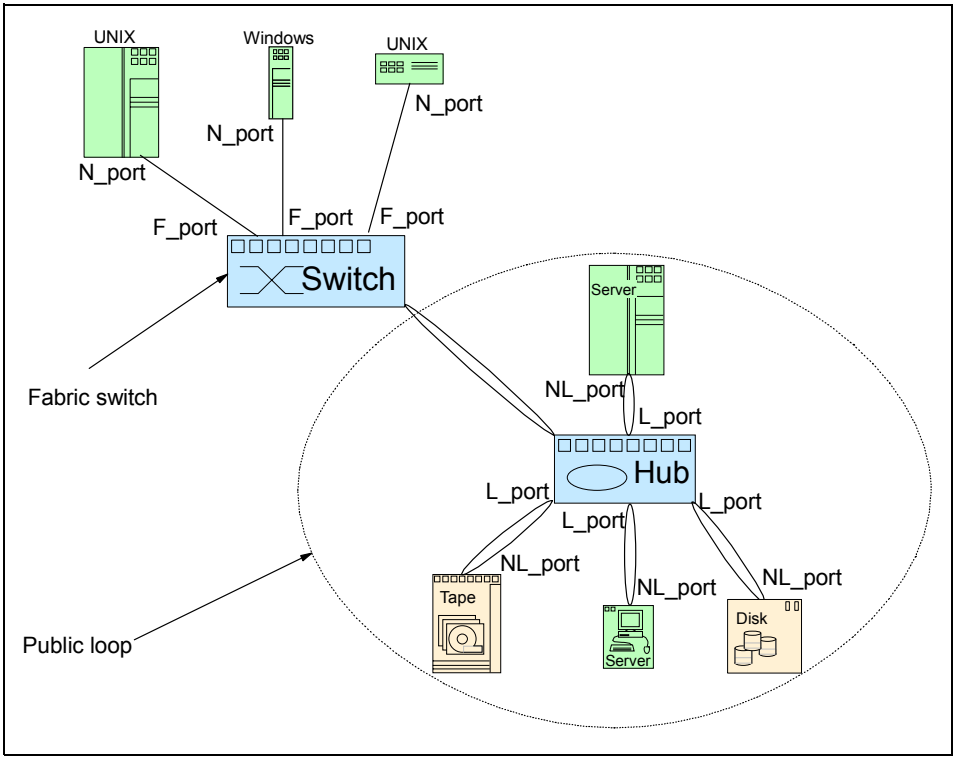


Figure 3-7 Public loop implementation

Arbitration

When a loop port wants to gain access to the loop, it has to arbitrate. When the port wins arbitration, it can open a loop circuit with another port on the loop; a function similar to selecting a device on a bus interface. Once the loop circuit has been opened, the two ports can send and receive frames between each other. This is known as loop tenancy.

If more than one node on the loop is arbitrating at the same time, the node with the lower Arbitrated Loop Physical Address (AL_PA) gains control of the loop. Upon gaining control of the loop, the node then establishes a point-to-point transmission with another node using the full bandwidth of the media. When a node has finished transmitting its data, it is not required to give up control of the loop. This is a channel characteristic of Fibre Channel. However, there is a fairness algorithm, which states that a device cannot regain control of the loop until the other nodes have had a chance to control the loop.

Loop addressing

An NL_Port, like an N_Port, has a 24-bit port address. If no switch connection exists, the two upper bytes of this port address are zeroes (x'00 00') and referred to as a private loop. The devices on the loop have no connection with the outside world. If the loop is attached to a fabric and an NL_Port supports a fabric login, the upper two bytes are assigned a positive value by the switch. We call this mode a public loop.

As fabric-capable NL_Ports are members of both a local loop and a greater fabric community, a 24-bit address is needed as an identifier in the network. In the case of public loop assignment, the value of the upper two bytes represents the loop identifier, and this will be common to all NL_Ports on the same loop that performed login to the fabric.

In both public and private arbitrated loops, the last byte of the 24-bit port address refers to the arbitrated loop physical address (AL_PA). The AL_PA is acquired during initialization of the loop and may, in the case of fabric-capable loop devices, be modified by the switch during login.

The total number of the AL_PAs available for arbitrated loop addressing is 127. This number is based on the requirements of 8b/10b running disparity between frames.

As a frame terminates with an end-of-frame character (EOF), this will force the current running disparity negative. In the Fibre Channel standard each transmission word between the end of one frame and the beginning of another frame should also leave the running disparity negative. If all 256 possible 8-bit bytes are sent to the 8b/10b encoder, 134 emerge with neutral disparity characters. Of these 134, seven are reserved for use by Fibre Channel. The 127 neutral disparity characters left have been assigned as AL_PAs. Put another way, the 127 AL_PA limit is simply the maximum number, minus reserved values, of neutral disparity addresses that can be assigned for use by the loop. This does not imply that we recommend this amount, or load, for a 200MB/s shared transport, but only that it is possible.

Arbitrated loop will assign priority to AL_PAs, based on numeric value. The lower the numeric value, the higher the priority is. For example, an AL_PA of x'01' has a much better position to gain arbitration over devices that have a lower priority or higher numeric value. At the top of the hierarchy it is not unusual to find servers, but at the lower end you would expect to find disk arrays.

It is the arbitrated loop initialization that ensures each attached device is assigned a unique AL_PA. The possibility for address conflicts only arises when two separated loops are joined together without initialization.

Closing a loop circuit

When two ports in a loop circuit complete their frame transmission, they may close the loop circuit to allow other ports to use the loop. The point at which the loop circuit is closed depends on the higher-level protocol, the operation in progress, and the design of the loop ports.

Supported devices

An arbitrated loop may support a variety of devices, such as:

- Individual Fibre Channel disk drives
- JBOD
- Fibre Channel RAID
- Native Fibre Channel tape sub-systems
- Fibre Channel to SCSI bridges

Broadcast

Arbitrated loop, in contrast to Ethernet, is a non-broadcast transport. When an NL_Port has successfully won the right to arbitration, it will open a target for frame transmission. Any subsequent loop devices in the path between the two will see the frames and forward them on to the next node in the loop.

It is this non-broadcast nature of arbitrated loop, by removing frame handling overhead from some of the loop, which enhances performance.

Distance

As stated before, arbitrated loop is a closed-ring topology. The total distance requirements being determined by the distance between the nodes. At gigabit speeds, signals propagate through fiber-optic media at five nanoseconds per meter and through copper media at four nanoseconds per meter. This is the delay factor.

Calculating the total propagation delay incurred by the loop's circumference is achieved by multiplying the length — both transmit and receive — of copper and fiber-optic cabling deployed by the appropriate delay factor. For example, a single 10 km link to an NL_Port would cause a 50 microsecond (10 km x 5 nanoseconds delay factor) propagation delay in each direction and 100 microseconds in total. This equates to 1 MB/s of bandwidth used to satisfy the link.

Bandwidth

For optical interconnects for SANs, the bandwidth requirements are greatly influenced by the capabilities of:

- The system buses

- Network switches
- The interface adapters that interface with them
- Traffic locality

The exact bandwidth required is somewhat dependent on implementation, but is currently in the range of 100 to 1000 MB/s. Determining bandwidth requirements is difficult, and there is no exact science that can take into account the unpredictability of sporadic bursts of data, for example. Planning bandwidth based on peak requirements could be wasteful. Designing for sustained bandwidth requirements, with the addition of safety margins, may be less wasteful.

3.1.6 Switched fabric

The third and most useful topology used in SAN implementations is Fibre Channel Switched Fabric (FC-SW). It applies to directors that support the FC-SW standard, that is, it is not limited to switches as its name suggests. A Fibre Channel fabric is one or more fabric switches in a single, sometimes extended, configuration. Switched fabrics provide full 200 MB/s bandwidth per port (or 100 MB/s for devices based on the older 1 Gb/s infrastructure), compared to the shared bandwidth per port in Arbitrated loop implementations.

If you add a new device into the arbitrated loop, you further divide the shared bandwidth. However, in a switched fabric, adding a new device or a new connection between existing ones actually increases the bandwidth. For example, an 8-port switch (based on 2 Gb/s technology) with three initiators and three targets can support three concurrent 200 MB/s conversations or a total of 600 MB/s throughput (1,200 MB/s if full-duplex applications were available).

A switched fabric configuration is shown in Figure 3-8 on page 58.

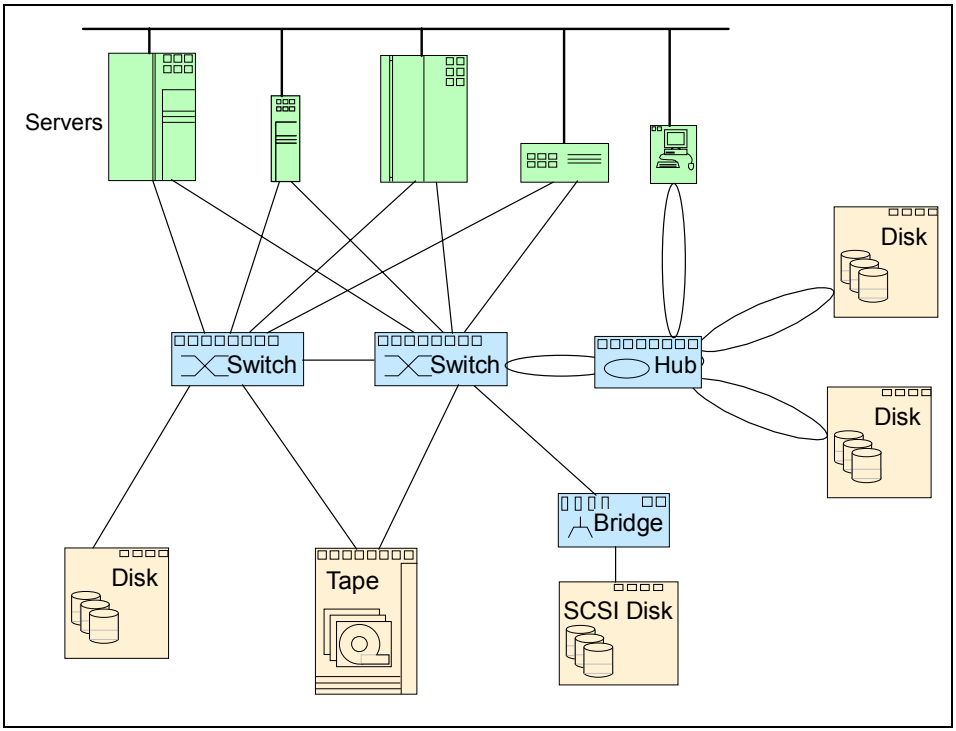


Figure 3-8 Sample switched fabric configuration

Expanding the fabric

As the demand for access to the storage grows, a switched fabric can be expanded to service these needs. Not all storage requirements can be satisfied with fabrics alone. For some applications, the 200 MB/s per port and advanced services are overkill, and they amount to wasted bandwidth and unnecessary cost. When you design a storage network, you need to consider the application's needs and not just rush to implement the latest technology available. SANs are often combinations of switched fabric and arbitrated loops.

Cascading

Expanding the fabric is called switch cascading. Cascading is basically interconnecting Fibre Channel switches and/or directors. The cascading of switches provides the following benefits to a SAN environment:

- The fabric can be seamlessly extended. Additional switches can be added to the fabric, without powering down existing fabric.
- You can easily increase the distance between various SAN participants.

- By adding more switches to the fabric, you increase connectivity by providing more available ports.
- Cascading provides high resilience in the fabric.
- With inter-switch links (ISLs), you can increase the bandwidth. The frames between the switches are delivered over all available data paths. So the more ISLs you create, the faster the frame delivery will be, but careful consideration must be employed to ensure that a bottleneck is not introduced.
- When the fabric grows, the name server is fully distributed across all the switches in fabric.
- With cascading, you also provide greater fault tolerance within the fabric.

Hops

The maximum number of switches allowed in the fabric is 239.

The other limitation is that only seven hops are allowed between any source and destination using IBM 2109 switches with FCP protocol.

There is a maximum of three hops using the Cisco, McDATA, and CNT directors.

There is a maximum of one longwave or extended distance ISL in any fabric path.

However, this is likely to change between vendors and over time and depending on the hosts.

For FICON™ cascading, only a two-director (or single hop) configuration is supported.

We show a sample configuration that illustrates this in Figure 3-9, with “Hoppy”, the hop count kangaroo.

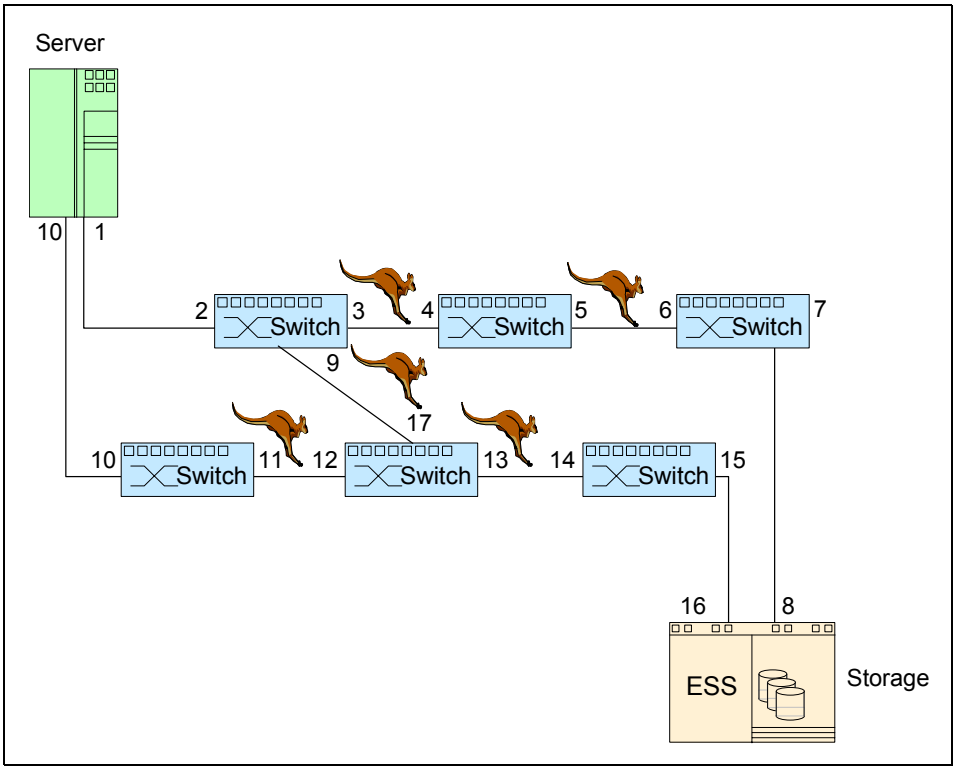


Figure 3-9 Cascading in a switched fabric

The hop count limit is set by the fabric operating system and is used to derive a frame holdtime value for each switch. This holdtime value is the maximum amount of time that a frame can be held in a switch before it is dropped (Class 3) or the fabric is busy (F_BSY, Class 2) is returned. A frame would be held if its destination port is not available. The holdtime is derived from a formula using the error detect time-out value (E_D_TOV) and the resource allocation time-out value (R_A_TOV).

The value of seven hops is not “hard-coded”, and if manipulation of E_D_TOV or R_A_TOV were to take place, the reasonable limit of seven hops could be exceeded. However, be aware that any hop suggestion was not a limit that was arrived at without careful consideration of a number of factors. In the future, the number of hops is likely to increase.

3.1.7 Inter Switch Links

According to the FC-SW Fibre Channel standard, the link joining a pair of E_Ports is called an Inter Switch Link (ISL).

ISLs carry frames originating from the node ports and those generated within the fabric. The frames generated within the fabric serve as control, management, and support for the fabric.

Before an ISL can carry frames originating from the node ports, the joining switches have to go through a synchronization process on which operating parameters are interchanged. If the operating parameters are not compatible, the switches may not join, and the ISL becomes “segmented”. Segmented ISLs cannot carry traffic originating on node ports, but they can still carry management and control frames.

Trunking

Depending on the estimated or measured traffic, you may connect some of your switches by parallel ISLs to share the load. The SAN standard routing protocol FSPF allows you to do so and use the cumulative bandwidth of all parallel ISLs (see Figure 3-10).

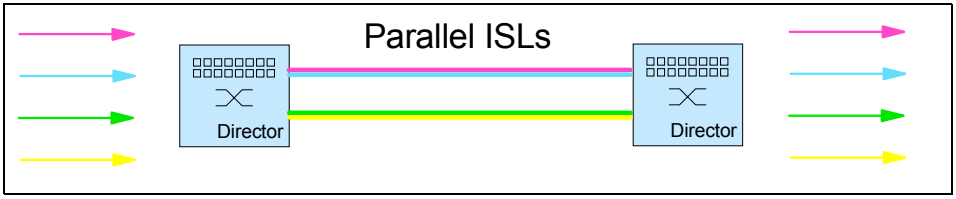


Figure 3-10 Parallel ISLs - low traffic

You need to be aware that load sharing reaches the boundary of its efficiency when servers send high amounts of data at the same time. As the switches dedicate the ISLs to the servers usually in a round-robin fashion, it may easily happen that one server occupies one ISL performing just a low rate of throughput and two other servers have to share the other ISL for their high rate of throughput (see Figure 3-11).

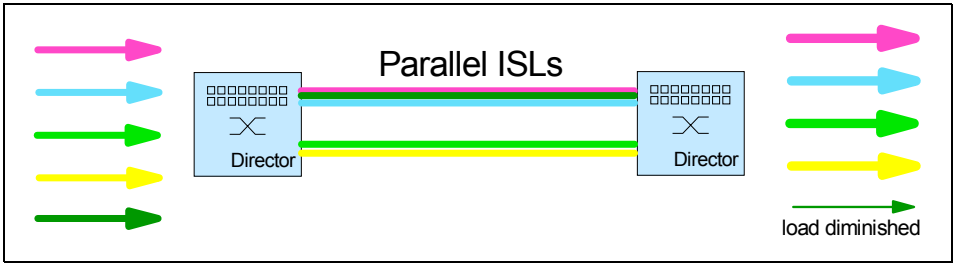


Figure 3-11 Parallel ISLs - high traffic

You may reduce, but not eliminate, this drawback by adding more ISLs in parallel; however, this may be far too expensive and subject to over-provisioning. Instead of this rather inflexible method of load *sharing*, switches may utilize a better way of load *balancing*. The implementation of load balancing is named trunking and is ideal for optimizing SAN performance (see Figure 3-12).

Each vendor of SAN switches will implement trunking in its own way. However, common to all their implementations is that transient workload peaks for one system or application are much less likely to impact the performance of other devices in the SAN fabric.

Note: The Cisco MDS 9000 family uses the term trunking tto refer to an ISL link that carries one or more VSANs. Consult the section on “VSANs” on page 65 for an explanation of Cisco virtual SANs.

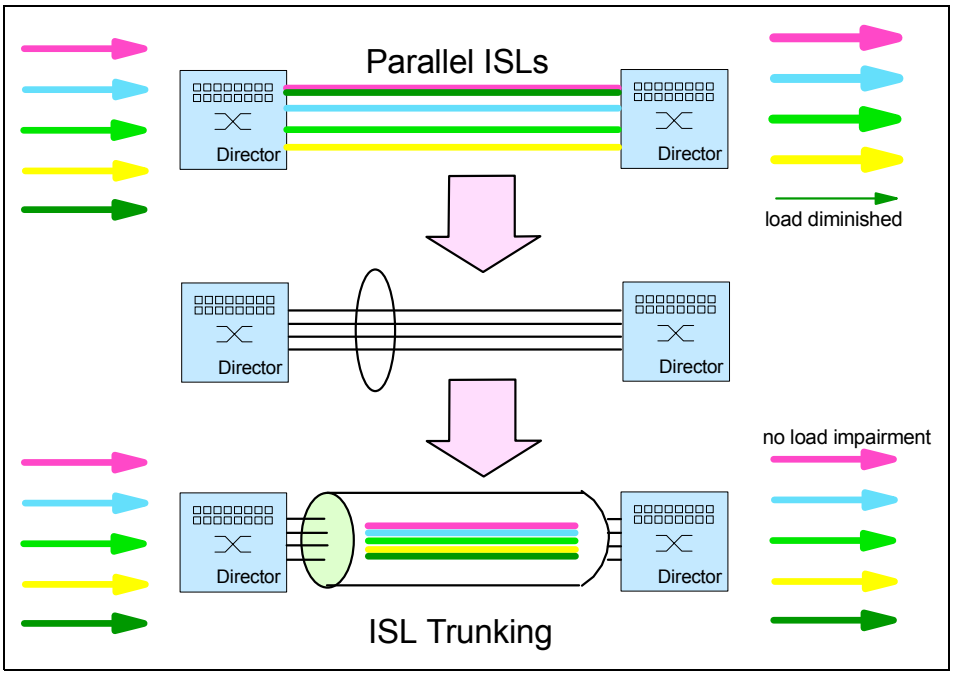


Figure 3-12 ISL Trunking

Load sharing or load balancing: Parallel ISLs always shared load or traffic in a “rough” server-oriented way: next server or next available ISL, regardless of the amount of traffic each server is causing. Load balancing provides the means to find an effective way to use all of the cumulative bandwidth of these parallel ISLs.

Oversubscribing the fabric

We can have several ports in a switch that can communicate with a single port, for example, several servers sharing a path to a storage device. In this case the storage path determines the maximum data rate that all servers can get, and this is usually given by the device and not the SAN itself.

When we start cascading switches, communication between switches are carried by ISLs, as previously stated. It is possible that several ports in one switch need to simultaneously communicate with ports in the other switch through a single ISL. In this case it is possible that the connected devices are able to sustain a data transfer rate higher than 100 MB/s, so the throughput will be limited to what the ISL can handle, and this may impose a throttle or roadblock within the fabric.

We use the term oversubscription to describe a situation when we have several ports trying to communicate with each other, and when the total throughput is higher than what that port can provide. Oversubscription, in itself, is not a bad thing. It is actually good, because it would be too cost prohibitive to dedicate bandwidth and resources for every connection. The problem arises if the oversubscription results in congestion. Congestion occurs when there is not enough bandwidth available for the application or connection. This can happen on storage ports and ISLs.

When designing a SAN, it is important to consider the possible traffic patterns to determine the possibility of oversubscription and which patterns may result in congestion. For example, traffic patterns during backup periods may introduce oversubscription that can affect performance on production systems. In some cases this is not a problem that may even be noticed at first, but as the SAN fabric grows, it is important not to ignore this possibility.

Fabric shortest path first

According to the FC-SW-2 standard, Fabric Shortest Path First (FSPF) is a link state path selection protocol. FSPF keeps track of the links on all switches in the fabric and associates a cost with each link. The protocol computes paths from a switch to all the other switches in the fabric by adding the cost of all links traversed by the path, and choosing the path that minimizes the cost.

For example, as shown in Figure 3-13, if we need to connect a port in switch A to a port in switch D, it will take the ISL from A to D. It will not go from A to B to D, nor from A to C to D.

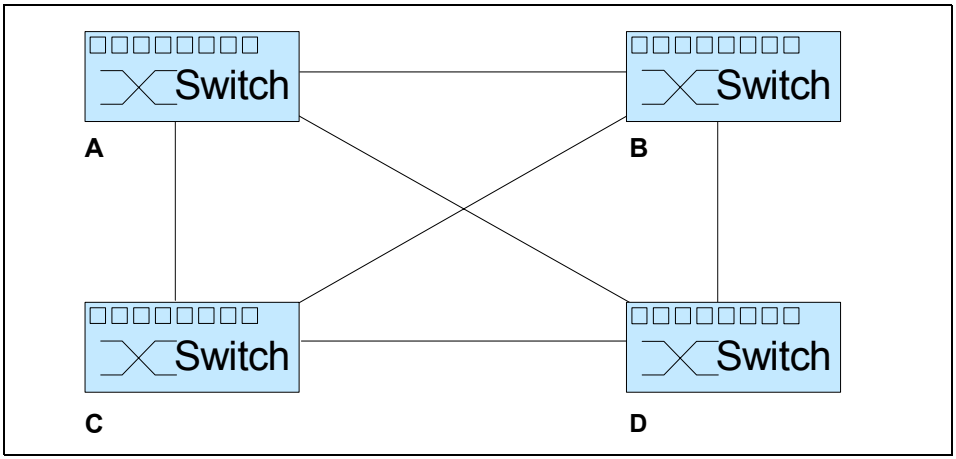


Figure 3-13 Four-switch fabric

This is because FSPF is currently based on the hop count cost.

The collection of link states (including cost) of all switches in a fabric constitutes the topology database (or link state database). The topology database is kept in all switches in the fabric, and they are maintained and synchronized to each other. There is an initial database synchronization, and an update mechanism. The initial database synchronization is used when a switch is initialized, or when an ISL comes up. The update mechanism is used when there is a link state change, for example, an ISL going down or coming up, and on a periodic basis. This ensures consistency among all switches in the fabric.

If we look again at the example in Figure 3-13, and we imagine that the link from A to D goes down, switch A will now have four routes to reach D:

- A-B-D
- A-C-D
- A-B-C-D
- A-C-B-D

A-B-D and A-C-D will be selected because they are the shortest paths based on the hop count cost. The update mechanism ensures that switches B and C will also have their databases updated with the new routing information.

PortChannels

Cisco offer a feature called port channels. Port channels allow users to aggregate up to sixteen physical ISLs into a single logical bundle, designed to optimize bandwidth utilization and resilience between switches. The group of Fibre Channel ISLs designated to act as a Port-Channel may consist of any port on any 16 port switching module within the MDS 9000 chassis, allowing the overall Port-Channel to remain active upon failure of one or more ports, or failure of one or more switching modules. The port channels increase aggregate bandwidth and availability by distributing and load-balancing the fabric traffic across the functional ISLs in the port channel.

Load balancing

The standard does not provide for load balancing when there are multiple paths of the same cost, so it is up to the switch vendor to establish routing algorithms to balance the load across ISLs. The potential routes are stored in routing tables.

Some vendors allow you to adjust the cost of traversing the fabric, and it is wise to check with each vendor as to the adjustments that can be made. Some vendors also allow you to define static routes. Again, it is wise to check with each vendor regarding what you can do to affect the traffic that flows over ISLs.

The balancing is usually done at initialization, assigning the same number of paths to each ISL. However, having the same number of paths does not mean having the same bandwidth requirements. We may end up with different connections that have high performance requirements being assigned to the same ISLs; while other ISLs are not being used due to inactive connections. Current implementations do not include dynamic load balancing, although this is expected to change over time.

Due to the potential performance impact of oversubscribing ISLs, it is recommended to have high volume traffic inside a switch or director. When cascading is not an option, the number of ISLs should be planned, and should take into consideration the expected traffic through them under different conditions, for example, production workload, and backup workload. In the absence of quantitative data, if you plan for the peak workload, that may be as good a rule of thumb as any.

When ISL oversubscription is detected, one solution is to add additional ISLs. It can be done concurrently, and the new path will be automatically included in the routing tables.

VSANs

The MDS 9000 SAN Fabric family offers Cisco's Virtual SAN (VSAN) technology, offering the capability to overlay multiple hardware enforced virtual fabric

environments within a single physical fabric infrastructure. Each VSAN contains separate fabric services designed for enhanced scalability, resilience, and independence among storage resource domains. Each VSAN contains its own complement of hardware-enforced zones, dedicated name server, and management capabilities, just as if the VSAN were configured as a separate physical fabric.

Therefore, VSANs are designed to allow more efficient SAN utilization and flexibility, because SAN resources may be allocated and shared among more users, while supporting secure segregation of traffic and retaining independent control of resource domains on a VSAN-by-VSAN basis.

3.1.8 Adding new devices

Switched fabrics, by their very nature, are dynamic environments. They can handle topology changes as new devices are attached, or previously active devices are removed or taken offline. For these reasons it is important that notification of these types of events can be provided to participants (nodes) in the switched fabric.

Notification is provided by two functions:

- State Change Notification: SCN
- Registered State Change Notification: RSCN

These two functions are not obligatory, so each N_Port or NL_Port must register its interest in being notified of any topology changes, or if another device alters its state.

The original SCN service allowed an N_Port to send a notification change directly to another N_Port. This is not necessarily an optimum solution, as no other participants on the fabric will know about this change. RSCN offers a solution to this and will inform all registered devices about the change.

Perhaps the most important change that you would want to be notified about, is when an existing device goes offline. This information is very meaningful for participants that communicate with that device. For example, a server in the fabric environment would want to know if their resources are powered off or removed, or when new resources became available for its use.

Changed notification provides the same functionality for the switched fabric as loop initialization provides for arbitrated loop.

The Registered State Change Notification (RSCN) is part of the Extended Link Service (ELS) in the Fibre Channel protocol. It was defined within the Fabric Loop Attachment group (FC-FLA) as a replacement for State Change Notification

(SCN). You may consider RSCN similar to SCN plus the opportunity as a Fibre Channel device to register (subscribe) to that service or not. RSCN, like SCN, is used to notify FC devices about the status changes of other ports which may be of interest for them. For example, when a storage port becomes active or inactive, the switch will let the registered servers know by issuing a RSCN notification to them. RSCN notifications flow either from:

- Node ports to switch — by addressing the well-known fabric controller address of 0xFF FF FD (FC-FLA definition modified in FC-DA)
- Switch to switch — by addressing the fabric controller (FC-FLA definition modified in FC-MI-2)
- Switch to node port — from fabric controller to node fabric address (FC-FLA definition modified in FC-MI-2)

After a server has been notified via RSCN that another SCSI storage device has come online, the server may try and attach to that storage by performing a login to it. Or if the server was notified that some storage has gone offline, the server may like to verify the current status of that device. Without RSCN, in the latter case, the server probably wouldn't find out until it was sending SCSI-READs or WRITEs to that storage. These are the types of RSCNs:

- **Fabric Format:** Sent when a zone configuration is activated or deactivated or when and ISL in a fabric goes up or down.
- **Port Format:** Occurs when a device logs in or out of a fabric.
  - Sent to local devices on the same switch.
  - Sent to remaining switches in the fabric.
- **Area Format:** Occurs when an entire arbitrated loop goes up or down.
- **Domain Format:** Occurs when a switch is added or removed from a fabric.

3.2 Classes of service

In Fibre Channel, we have a combination of traditional I/O technologies with networking technologies.

We need to keep the functionality of traditional I/O technologies to preserve data sequencing and data integrity, and we need to add networking technologies that allow for a more efficient exploitation of available bandwidth.

Based on the methodology with which the communication circuit is allocated and retained, and in the level of delivery integrity required by an application, the Fibre Channel standards provide different classes of service:

3.2.1 Class 1

In a Class 1 service, a dedicated connection between source and destination is established through the fabric for the duration of the transmission. Each frame is acknowledged by the destination device back to the source device. This class of service ensures the frames are received by the destination device in the same order they are sent, and reserves full bandwidth for the connection between the two devices. It does not provide for a good utilization of the available bandwidth, since it is blocking another possible contender for the same device. Because of this blocking and the necessary dedicated connections, Class 1 is rarely used.

3.2.2 Class 2

In a Class 2 service there is no dedicated connection; each frame is sent separately using switched connections allowing several devices to communicate at the same time. For this reason Class 2 is also called “connectionless”. Although there is no dedicated connection, each frame is acknowledged from destination to source to confirm receipt. The use of delivery acknowledgments in Class 2 allows for quickly identifying communications problems at both the sending and receiving ports. Class 2 makes a better use of available bandwidth since it allows the fabric to multiplex several messages on a frame by frame basis. As frames travel through the fabric they can take different routes, so Class 2 does not guarantee in-order delivery. Class 2 relies on upper layer protocols to take care of frame sequence. The use of acknowledgments reduced available bandwidth which needs to be considered in large scale busy networks.

3.2.3 Class 3

Like Class 2, there is no dedicated connection in Class 3, the main difference is that received frames are not acknowledged. The flow control is based on BB Credit, but there is no individual acknowledgement of received frames. Class 3 is also called “datagram connectionless” service. It optimizes the use of fabric resources, but it is now up to the upper layer protocol to ensure all frames are received in the proper order, and to request to the source device the retransmission of any missing frame. Class 3 is the commonly used class of service in Fibre Channel networks.

Note: Classes 1, 2, and 3 are well defined and stable. They are defined in the FC-PH standard.

IBM 2109 switches, Cisco switches and directors, CNT and McDATA directors support Class 2 and Class 3 service.

3.2.4 Class 4

Class 4 is a connection oriented service like Class 1, but the main difference is that it allocates only a fraction of the available bandwidth of a path through the fabric that connects two N_Ports. Virtual Circuits (VCs) are established between N_Ports with guaranteed Quality of Service (QoS) including bandwidth and latency. The Class 4 circuit between two N_Ports consists of two unidirectional VCs, not necessarily with the same QoS. An N_Port may have up to 254 Class 4 circuits with the same or different N_Port. Like Class 1, Class 4 guarantees in-order frame delivery and provides acknowledgment of delivered frames, but now the fabric is responsible for multiplexing frames of different VCs. Class 4 service is mainly intended for multimedia applications such as video and for applications that allocate an established bandwidth by department within the enterprise. Class 4 was added in the FC-PH-2 standard.

3.2.5 Class 5

Class 5 is called isochronous service and it is intended for applications that require immediate delivery of the data as it arrives, with no buffering. It is not clearly defined yet. It is not included in the FC-PH documents.

3.2.6 Class 6

Class 6 is a variant of Class 1 known as multicast class of service. It provides dedicated connections for a reliable multicast. An N_Port may request a Class 6 connection for one or more destinations. A multicast server in the fabric will establish the connections and get the acknowledgment from the destination ports, and send it back to the originator. Once a connection is established it should be retained and guaranteed by the fabric until the initiator ends the connection. Only the initiator can send data and the multicast server will transmit that data to all destinations.Class 6 was designed for applications like audio and video requiring multicast functionality. It appears in the FC-PH-3 standard.

3.2.7 Class F

Class F Service is defined in the FC-SW and FC-SW2 standard for use by switches communicating through ISLs. It is a connectionless service with notification of non-delivery between E_Ports, used for control, coordination and configuration of the fabric. Class F is similar to Class 2 since it is a connectionless service, the main difference is that Class 2 deals with N_Ports sending data frames, while Class F is used by E_Ports for control and management of the fabric.

3.2.8 Communication

FC-2 is the protocol level that defines protocol signaling rules and defines the organization or structure of Fibre Channel communications. This structure allows for efficient flow control and allows the network to quickly identify where a network error is occurring. The following describes the levels within this structure, they are listed according to size from the largest to the smallest.

- **Exchanges:** The highest level Fibre Channel mechanism used for communication. An exchange contains one or more non-concurrent sequences being exchanged between a pair of Fibre Channel ports.
- **Sequences:** A sequence is a collection of frames related to one message element or information unit.
- **Frames:** A Fibre Channel frame consists of maximum 2112 bytes of data. It is considered as a basic unit of data transmission. It consists of a start delimiter, destination and source address, protocol metadata, data payload, CRC (error check value) and an end delimiter.
- **Words:** An addressable unit of data in memory. The smallest Fibre Channel data element consisting of 40 serial bits representing either a flag (K28.5) plus 3 encoded data bytes (10 encoded bits each) or four 10-bit encoded data bytes. An *ordered set* is a 4-byte transmission word that has the special character, K28.5 as its first character and 3 bytes used to define the meaning or function of the ordered set. They either identify the start of frame, the end of frame, or occur between Fibre Channel frames.

3.3 Buffers

Ports need memory, or “buffers”, to temporarily store frames as they arrive and until they are assembled in sequence, and delivered to the upper layer protocol.

The number of buffers (the number of frames a port can store) is called its “Buffer Credit”.

BB_Credit

During login, N_Ports and F_Ports at both ends of a link establish its Buffer to Buffer Credit (BB_Credit).

EE_Credit

During login all N_Ports establish End to End Credit (EE_Credit) with each other.

During data transmission, a port should not send more frames than the buffer of the receiving port can handle before getting an indication from the receiving port that it has processed a previously sent frame. Two counters are used for that purpose. BB_Credit_CNT and EE_Credit_CNT, and both are initialized to 0 during login.

Each time a port sends a frame, it increments BB_Credit_CNT and EE_Credit_CNT by 1. When it receives R_RDY from the *adjacent* port it decrements BB_Credit_CNT by 1, when it receives ACK from the *destination* port it decrements EE_Credit_CNT by 1. Should at any time BB_Credit_CNT become equal to the BB_Credit, or EE_Credit_CNT become equal to the EE_Credit of the receiving port, the transmitting port has to stop sending frames until the respective count is decremented.

The previous statements are true for Class 2 service. Class 1 is a dedicated connection, so it does not care about BB_Credit and only EE_Credit is used (EE Flow Control). Class 3 on the other hand is an unacknowledged service, so it only uses BB_Credit (BB Flow Control), but the mechanism is the same on all cases.

Here we can see the importance that the number of buffers has in overall performance. We need enough buffers to make sure the transmitting port can continue sending frames without stopping in order to use the full bandwidth.This is particularly true with distance.

BB_Credit considerations for long distance

BB_Credit needs to be taken into consideration on Fibre Channel devices that are several kilometers apart from each other and you need to know the distance separating the adjacent partners. We will assume that Fibre Channel devices A and B are connected by a 10 km fiber optic cable, as shown in Figure 3-14.

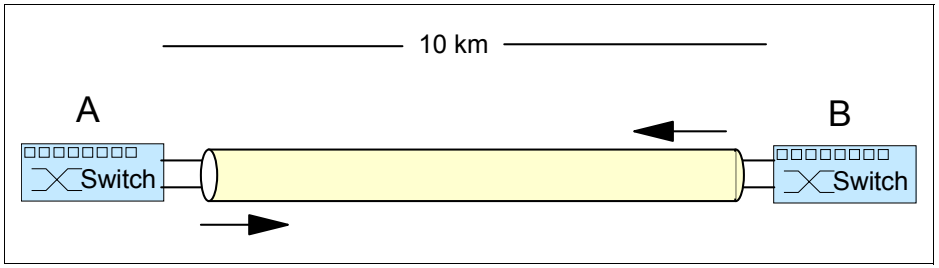


Figure 3-14 Adjacent FC devices

Light travels at approximately 300,000 km/s through a vacuum and at about 200,000 km/s through glass fiber. Latency is the inverse function of speed — so the optical signal of a Fibre Channel frame ends up with a latency of 5 ns/m.

$$latency = \frac{1}{speed} = \frac{1}{0.2\,\mathrm{exp}9\,\frac{m}{s}} = 5\,\mathrm{exp}{-9}\,\frac{s}{m} = 5\frac{ns}{m}$$

Light has a finite speed, and we need to take that into account when we figure out the maximum amount of frames that will be in transit over the fiber optic cable from A to B. A distance of 10 km over Fibre Channel means a round-trip of 20 km in total. The optical signal takes 100 µs propagation time (t_p) to travel that round-trip.

$$t_p = distance \times latency = 20\,\mathrm{exp}3\,m \times 5\,\mathrm{exp}{-9}\,\frac{s}{m} = 100\,\mathrm{exp}{-6}\,s = 100\,\mu s$$

That is the shortest possible time that A can expect to get an R_RDY back from B which, once received, signals that more frames can be sent.

Round-trip: We assume a data frame is sent from A and when it arrives at B a Receiver Ready (R_RDY) travels back to A. So for our equations, it is based on one frame which would make the round-trip.

Fibre Channel frames are usually 2 KB large, but because of 8b/10b encoding, they will become larger, as the encoding causes 1 byte to become 10 bits. Sending 2 KB Fibre Channel frames over fiber optic cable with 1 Gb/s bandwidth computes to a sending time (t_s) of 20 µs per frame. In other words: 20 µs is the time A needs to send 2000 bits.

$$t_s = \frac{Framelength}{Bandwidth} = \frac{20\,\mathrm{exp}3\,b}{1\,\mathrm{exp}9\frac{b}{s}} = 2\,\mathrm{exp}{-6}\,s = 20\mu s$$

To give an idea of how long a Fibre Channel frame spreads out on the fiber optic cable link, in 20 µs the light travels 4000 m (l), so the 2 KB frame occupies 4 km of fiber optic cable from the first bit transmitted to the last bit transmitted.

$$l = t_s \times speed = 20\,\mathrm{exp}{-6}\,s \times 0.2\,\mathrm{exp}9\frac{m}{s} = 4\,\mathrm{exp}3\,m = 4000\,m$$

The ratio between propagation time and sending time gives us the maximum amount of frames which *A* may send out before *B's* response would arrive back at *A*.

$$BB_Credit = \frac{t_p}{t_s} = \frac{100\mu s}{20\mu s} = 5$$

So, *A* may send out five consecutive frames to fill up the whole Fibre Channel during the time it is waiting for a response from *B*. In order to do so, *A* needs to hold at least 5 BB_Credits to use the Fibre Channel effectively. Distances in the range of a few hundred meters and below are not usually effected. It becomes an area for consideration with longer distances in the region of 50-100 km or more when extending Fibre Channel links over DWDM or ATM. To guarantee the same effectiveness for the optical signal's propagation time for a 100 km distance, you would need to make sure that *A* has 50 BB_Credits available. Doubling the bandwidth of the fiber optic link from 1 Gb/s to 2 Gb/s means there may be twice as many Fibre Channel frames on the link at the same time and so we will need twice as many BB_Credits to reach the same efficiency. That is theoretically 100 BB_Credits on a 100 km, 2 Gb/s fiber optic link.

Practically speaking, you may not need that much BB_Credit, as it is unlikely that the FC device will fill up the Fibre Channel 100% over a sustained period.

3.4 Addressing

All participants in a Fibre Channel environment have an identity. The way that the identity is assigned and used depends on the format of the Fibre Channel fabric. For example, there is a difference between the way that addressing is done in an Arbitrated Loop and a switch fabric.

Each participant in the Fibre Channel environment has a unique ID, which is called the World Wide Name (WWN). This WWN is a 64-bit address, and if two WWN addresses are put into the frame header, this leaves 16 bytes of data just for identifying destination and source address. So 64-bit addresses can impact routing performance.

Because of this, there is another addressing scheme used in Fibre Channel networks. This scheme is used to address the ports in the switched fabric. Each port in the switched fabric has its own unique 24-bit address. With this 24-bit addressing scheme, we get a smaller frame header, and this can speed up the routing process. With this frame header and routing logic, the Fibre Channel fabric is optimized for high-speed switching of frames.

With a 24-bit addressing scheme, this allows for up to 16 million addresses, which is an address space larger than any practical SAN design in existence in today's world. This 24-bit address has to somehow be connected to and with the 64-bit address associated with World Wide Names.

3.4.1 World Wide Name

All Fibre Channel devices have a unique identity called the World Wide Name (WWN). This is similar to the way that all ethernet cards have a unique MAC address.

Each N_Port will have its own WWN but it is also possible for a device with more than one Fibre Channel adapter to have its own WWN as well. Thus, for example, an IBM TotalStorage Enterprise Storage Server has its own WWN as well as incorporating the WWNs of the adapters within it. This means that a soft zone can be created using the entire array, or individual zones could be created using particular adapters. In the future, this will be the case for servers as well.

This WWN is a 64-bit address and if two WWN addresses are put into the frame header, this leaves 16 bytes of data just for identifying destination and source address. So 64-bit addresses can impact routing performance.

3.4.2 WWN and WWPN

Each device in the SAN is identified by a unique world wide name (WWN). The WWN contains a vendor identifier field, which is defined and maintained by the IEEE, and a vendor specific information field.

For further information, visit the Web site:

http://standards.ieee.org/

Currently, there are two formats of the WWN as defined by the IEEE. The original format contains either a hex 10 or hex 20 in the first two bytes of the address. This is then followed by the vendor specific information.

The new addressing scheme starts with a hex 5 or 6 in the first half-byte followed by the vendor identifier in the next 3 bytes. The vendor specific information is then contained in the following fields.

Both the old and new WWN formats are shown in Figure 3-15.

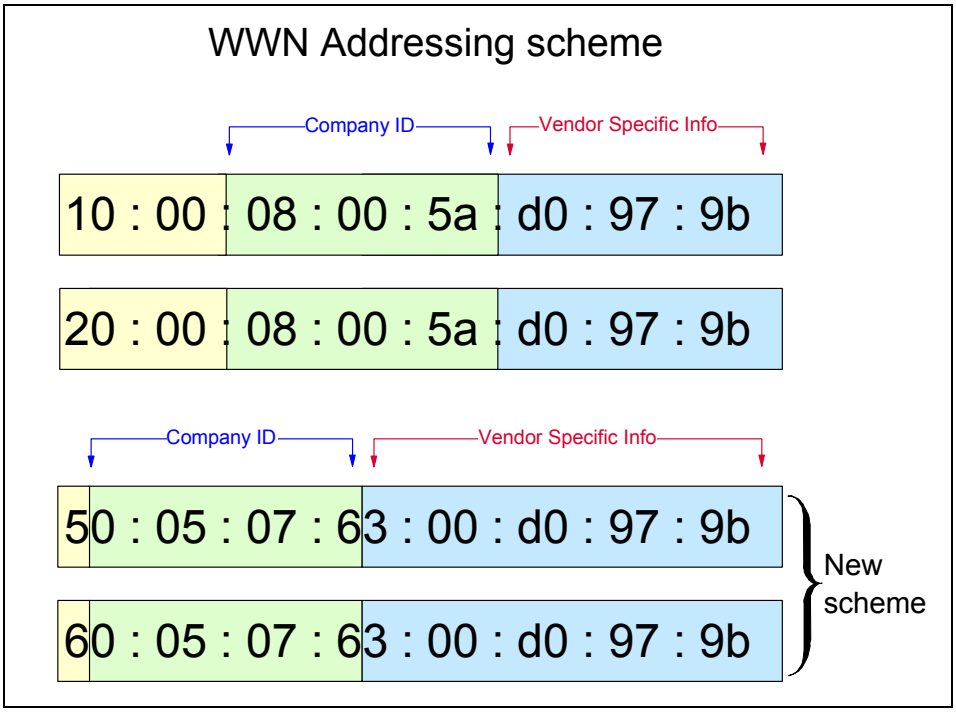


Figure 3-15 World Wide Name addressing scheme

The complete list of vendor identifiers as maintained by the IEEE is available at:

http://standards.ieee.org/regauth/oui/oui.txt

Table 3-1 lists a few of these vendor identifiers.

Table 3-1 WWN company identifiers

WWN (hex)	Company
00-50-76	IBM Corporation
00-60-69	Brocade Communications
08-00-88	McDATA Corporation
00-60-DF	CNT Technologies Corporation

Some devices may have multiple Fibre Channel adapters, like an ESS, for example. In this case the device also has an identifier for each of its Fibre Channel adapters. This identifier is called the world wide port name (WWPN). This way it is possible to uniquely identify all Fibre Channel adapters and paths within a device.

This is illustrated in Figure 3-16.

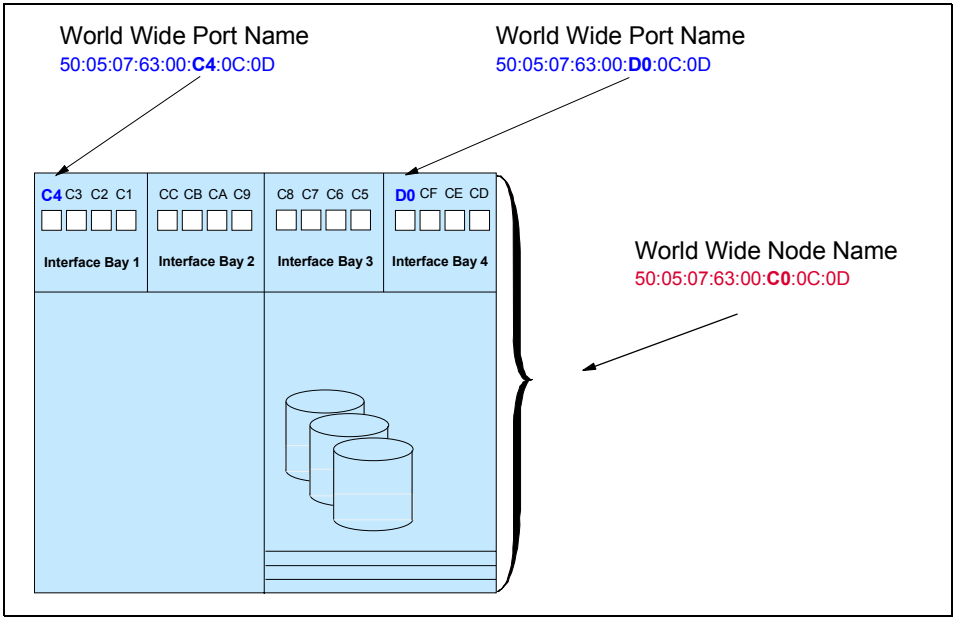


Figure 3-16 WWN and WWPN

This diagram shows how each of the ESS's Fibre Channel adapters has a unique WWPN. In the case of the ESS, the vendor specific information field is

used to identify each Fibre Channel adapter according to which bay and slot position it is installed in within the ESS.

Shown in Figure 3-17 is a screen capture of the name server table for a test SAN in the ITSO lab. This shows that the two devices (DEC HSG80 and IBM 1742) both have multiple HBAs. The name server table shows the WWN for each device as being the same, but the WWPN is different for each HBA within these devices.

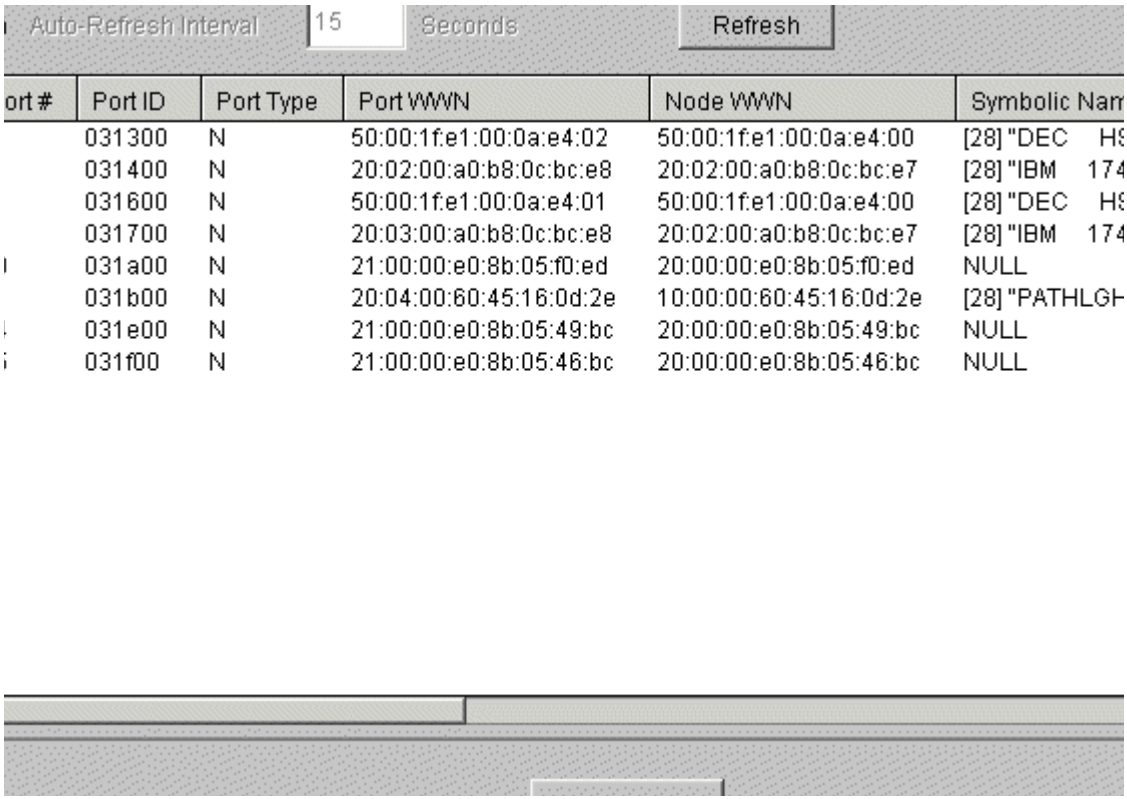


Figure 3-17 WWN and WWPN entries in a name server table

The switch must make a correlation between the WWN and the 24-bit port address.

3.4.3 24-bit Port address

Each port in the switched fabric has its own unique 24-bit address. The relationship between this 24-bit address and the 64-bit address associated with World Wide Names is explained below.

The 24-bit address scheme removes the overhead of manual administration of addresses by allowing the topology itself to assign addresses. This is not like WWN addressing, in which the addresses are assigned to the manufacturers by the IEEE standards committee, and are built in to the device at time of manufacture, similar to naming a child at birth. If the topology itself assigns the 24-bit addresses, then somebody has to be responsible for the addressing scheme from WWN addressing to port addressing.

In the switched fabric environment, the switch itself is responsible for assigning and maintaining the port addresses. When the device with its WWN logs into the switch on a specific port, the switch will assign the port address to that port and the switch will also maintain the correlation between the port address and the WWN address of the device on that port. This function of the switch is implemented by using a Name Server .

The Name Server is a component of the fabric operating system, which runs inside the switch. It is essentially a database of objects in which fabric-attached device registers its values.

Dynamic addressing also removes the potential element of human error in address maintenance, and provides more flexibility in additions, moves, and changes in the SAN.

A 24-bit port address consists of three parts:

- Domain (bits from 23 to 16)
- Area (bits from 15 to 08)
- Port or Arbitrated Loop physical address: AL_PA (bits from 07 to 00)

We show how the address is built up in Figure 3-18 on page 78.

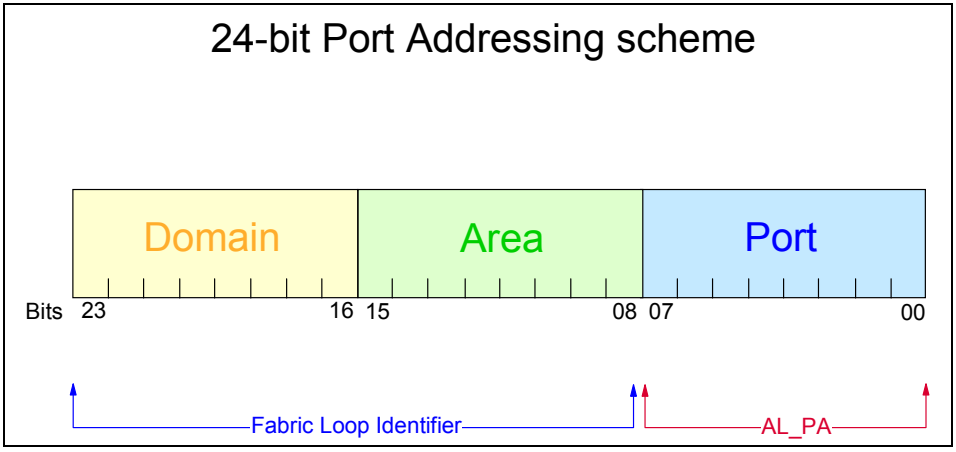


Figure 3-18 Fabric port address

We explain the significance of some of the bits that make up the port address in the following sections.

- Domain: The most significant byte of the port address is the domain. This is the address of the switch itself. One byte allows up to 256 possible addresses. Because some of these are reserved (like the one for broadcast) there are only 239 addresses actually available. This means that you can theoretically have as many as 239 switches in your SAN environment. The domain number allows each switch to have a unique identifier if you have multiple interconnected switches in your environment.
- Area: The area field provides 256 addresses. This part of the address is used to identify the individual FL_Ports supporting loops or it can be used as the identifier for a group of F_Ports; for example, a card with more ports on it. This means that each group of ports has a different area number, even if there is only one port in the group.
- Port; The final part of the address provides 256 addresses for identifying attached N_Ports and NL_Ports.See “3.4.4, “Loop address” on page 79”.

To arrive at the number of available addresses is a simple calculation based on:

```
Domain x Area x Ports
```

This means that there are 239 x 256 x 256 = 15,663,104 addresses available.

3.4.4 Loop address

An NL_Port, like an N_Port, has a 24-bit port address. If no switch connection exists, the two upper bytes of this port address are zeroes (x’00 00’) and referred to as a private loop. The devices on the loop have no connection with the outside world. If the loop is attached to a fabric and an NL_Port supports a fabric login, the upper two bytes are assigned a positive value by the switch. We call this mode a public loop.

As fabric-capable NL_Ports are members of both a local loop and a greater fabric community, a 24-bit address is needed as an identifier in the network. In the case of public loop assignment, the value of the upper two bytes represents the loop identifier, and this will be common to all NL_Ports on the same loop that performed login to the fabric.

In both public and private Arbitrated Loops, the last byte of the 24-bit port address refers to the Arbitrated Loop physical address (AL_PA). The AL_PA is acquired during initialization of the loop and may, in the case of fabric-capable loop devices, be modified by the switch during login.

The total number of the AL_PAs available for Arbitrated Loop addressing is 127. This number is based on the requirements of 8b/10b running disparity between frames.

As a frame terminates with an end-of-frame character (EOF), this will force the current running disparity negative. In the Fibre Channel standard each transmission word between the end of one frame and the beginning of another frame should also leave the running disparity negative. If all 256 possible 8-bit bytes are sent to the 8b/10b encoder, 134 emerge with neutral disparity characters. Of these 134, seven are reserved for use by Fibre Channel. The 127 neutral disparity characters left have been assigned as AL_PAs. In fact one of those is reserved to be an FL_Port. Put another way, there is an absolute limit of 126 L_Ports on any loop. This does not imply that we recommend this amount, or load, for a 100MB/s shared transport, but only that it is possible.

3.4.5 FICON addressing

Ficon generates the 24-bit FC port address field in yet another way. When communication is required from the FICON channel port to the FICON CU port, the FICON channel (using FC-SB-2 and FC-FS protocol information) will provide both the address of its port, the source port address identifier (S_ID), and the address of the CU port, the destination port address identifier (D_ID) (when the communication is from the channel N_Port to the CU N_Port).

The Fibre Channel architecture does not specify how a Server N_Port determines the destination port address of the Storage Device N_Port it requires to communicate to. This is Node and N_Port implementation dependent. Basically, there are two ways that a server can determine the address of the N_Port that it wishes to communicate to:

- The “discovery” method, by knowing the World Wide Name (WWN) of the target Node N_Port and then requesting a WWN for the N_Port port address from a Fibre Channel Fabric Service called the fabric Name Server.
- The “defined” method, by the Server (Processor channel) N_Port having a known predefined port address of the Storage Device (CU) N_Port that it requires to communicate to. This later approach is referred to as the “port address definition approach”, and is the approach that is implemented for the FICON channel in FICON native (FC) mode by the zSeries and the 9672 G5/G6, using either the z/OS HCD function or an IOCP program to define a 1-byte switch port (1-byte FC “area” field of the 3-byte fiber channel N_Port port address).

The Fibre Channel architecture (FC-FS) uses a 24-bit FC port address (3 bytes) for each port in an FC switch. The switch port addresses in a FICON native (FC) mode are always assigned by the switch fabric.

For the FICON channel in FICON native (FC) mode, the Accept (ACC ELS) response to the Fabric Login (FLOGI), in a switched point-to-point topology, provides the channel with the 24-bit N_Port address to which the channel is connected. This N_Port address is in the ACC destination address field (D_ID) of the FC-2 header.

The FICON CU port will also perform a fabric login to obtain its 24-bit FC port address. Figure 3-19 shows the FC-FS 24-bit FC port address identifier is divided into three fields:

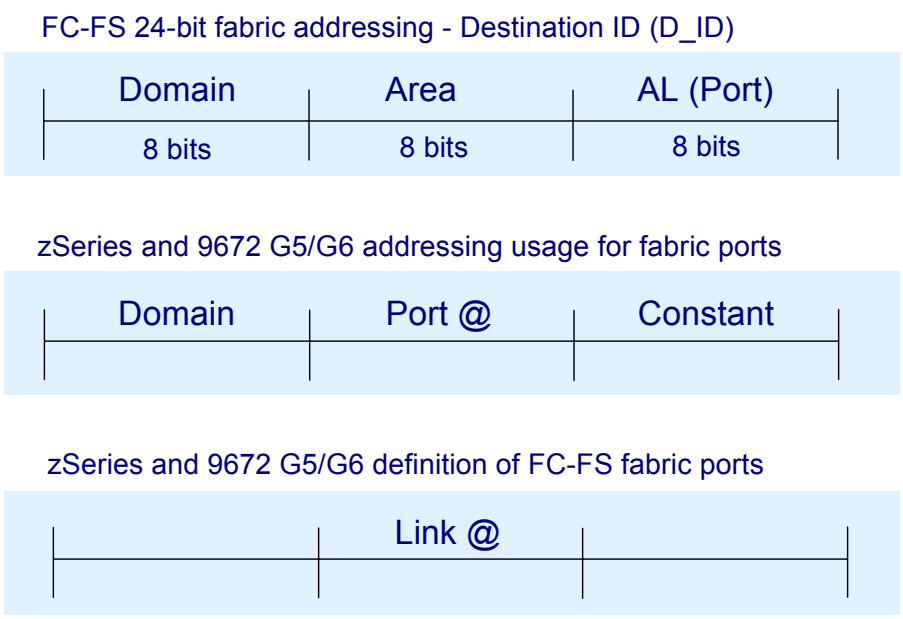


Figure 3-19 Ficon port addressing

It shows the FC-FS 24-bit port address and the definition usage of that 24-bit address in a zSeries and 9672 G5/G6 environment. Only the 8 bits making up the FC port address are defined for the zSeries and 9672 G5/G6 to access a FICON CU. The FICON channel in FICON native (FC) mode working with a switched point-to-point FC topology (single switch) provides the other 2 bytes that make up the 3-byte FC port address of the CU to be accessed.

The zSeries and 9672 G5/G6 processors, when working with a switched point-to-point topology, require that the Domain and the AL_Port (Arbitrated

Loop) field values be the same for all the FC F_Ports in the switch. Only the “area” field value will be different for each switch F_Port.

For the zSeries and 9672 G5/G6 the “area” field is referred to as the F_Port’s “port address” field. It is just a 1-byte value, and when defining access to a CU that is attached to this port (using the zSeries HCD or IOCP), the port address is referred to as the Link address.

As shown in Figure 3-20 on page 82, the 8 bits for the domain address and the 8-bit constant field are provided from the Fabric Login initialization result, while the 8 bits (1 byte) for the port address (1-byte Link address) are provided from the zSeries or 9672 G5/G6 CU link definition (using HCD and/or IOCP).

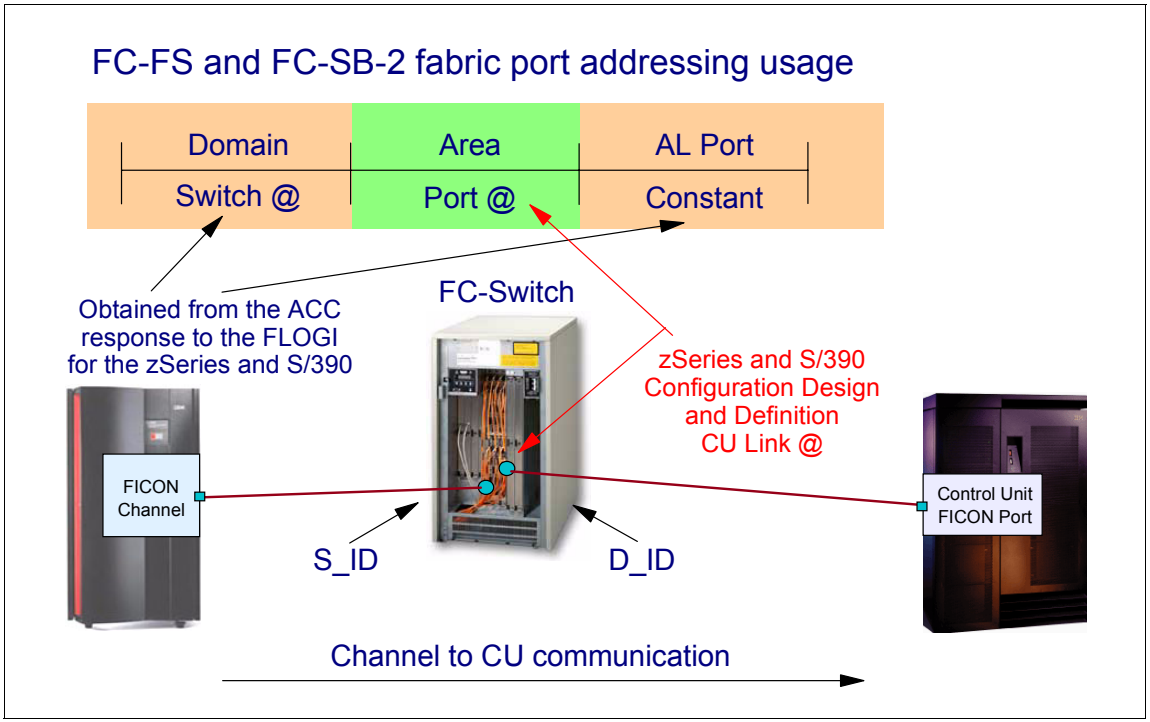


Figure 3-20 FICON single switch - switched point-to-point link address

FICON address support for cascaded switches

The Fibre Channel architecture (FC-FS) uses a 24-bit FC port address (3 bytes) for each port in an FC switch. The switch port addresses in a FICON native (FC) mode are always assigned by the switch fabric.

For the FICON channel in FICON native (FC) mode, the Accept (ACC ELS) response to the Fabric Login (FLOGI) in a 2-switch cascaded topology, provides

the channel with the 24-bit N_Port address to which the channel is connected. This N_Port address is in the ACC destination address field (D_ID) of the FC-2 header.

The FICON CU port will also perform a fabric login to obtain its 24-bit FC port address.

Figure 3-21 on page 83 shows that the FC-FS 24-bit FC port address identifier is divided into three fields:

- Domain
- Area
- AL Port

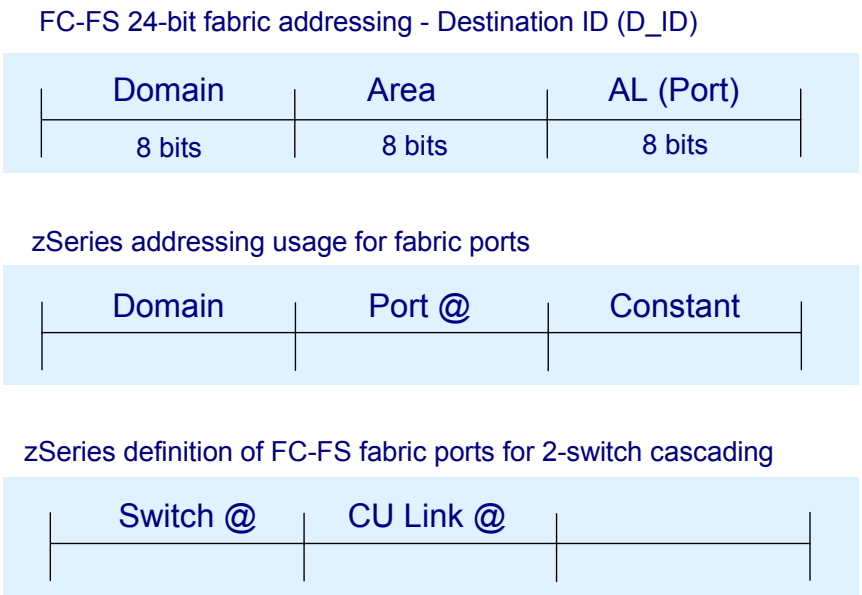


Figure 3-21 ficon addressing for cascaded directors

It shows the FC-FS 24-bit port address and the definition usage of that 24-bit address in a zSeries environment. Now 16 bits making up the FC port address must be defined for the zSeries to access a FICON CU in a cascaded environment. The FICON channel in FICON native (FC) mode working with a

cascaded FC topology (2-switch) provides the remaining byte that make up the full 3-byte FC port address of the CU to be accessed.

It is required that the Domain (switch @) and the AL_Port (Arbitrated Loop) field value be the same for all the FC F_Ports in the switch. Only the “area” field value will be different for each switch F_Port.

The zSeries “domain” and “area” fields are referred to as the F_Port’s “port address” field. It is a 2-byte value, and when defining access to a CU that is attached to this port (using the zSeries HCD or IOCP), the port address is referred to as the Link address.

As shown in Figure 3-22 on page 84, the 8 bits for the constant field are provided from the Fabric Login initialization result, while the 16 bits for the port address (2-byte Link address) are provided from the zSeries CU link definition (using HCD and/or IOCP)..

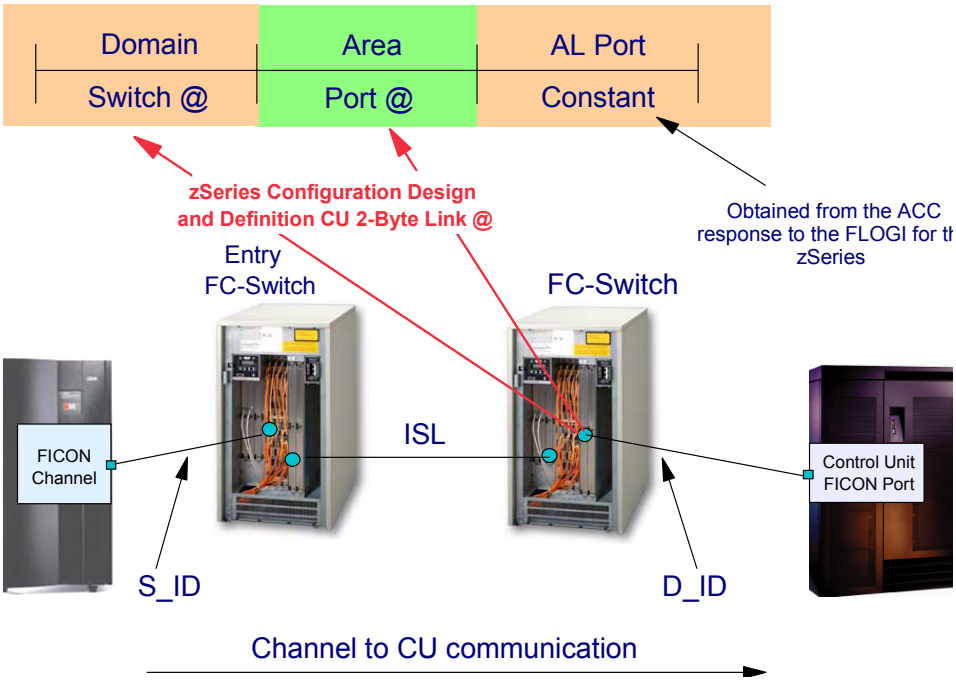


Figure 3-22 Two cascaded director ficon addressing

3.5 Fabric services

There are a set of services available to all device participating in a fabric. They are known as fabric services, and include:

- Management services
- Time services
- Name services
- Login services
- Registered State Change Notification (RSCN)

The services are implemented by switches and directors participating in the SAN. Generally speaking, the services are distributed across all the devices, and a node can make use of whichever switching device it is connected to.

3.5.1 Management services

This is an inband fabric service which allows data to be passed from devices to management platforms. This will include such information as the topology of the SAN. A critical feature of this service is that it allows management software access to the SNS bypassing any potential block caused by zoning. This means that a management suite can have a view of the entire SAN. The well known port used for the Management Server is 0xFFFFFA.

3.5.2 Time services

This is defined but has not yet been implemented (as far as we know). The assigned port is 0xFFFFFB.

3.5.3 Name services

Fabric switches implement a concept known as the Name Server. All switches in the fabric keep the SNS updated, and are therefore all aware of all devices in the SNS. After a node has successfully logged into the fabric, it performs a PLOGI into a well known port, 0xFFFFFC. This allows it to register itself and pass on critical information such as class of service parameters, its WWN/address and the Upper Layer Protocols which it can support.

3.5.4 Login services

In order to do a fabric login, a node communicates with the login server at address 0xFFFFFE. For more details see 3.6.1, “Fabric login” on page 86.

3.5.5 Registered State Change Notification

This service, Registered State Change Notification (RSCN), is critical as it propagates information about a change in state of one node to all other nodes in the fabric. This means that in the event of, for example, a node being shutdown, that the other nodes on the SAN will be informed and can take necessary steps to stop communicating with it. This prevents the other nodes trying to communicate with the node that has been shutdown, timing out and retrying.

3.6 Logins

There are three different types of login for Fibre Channel. These are:

- Fabric login
- Port login
- Process login

3.6.1 Fabric login

After the fabric capable Fibre Channel device is attached to a fabric switch, it will carry out a fabric login (FLOGI).

Similar to port login, FLOGI is an extended link service command that sets up a session between two participants. With FLOGI a session is created between an N_Port or NL_Port and the switch. An N_Port will send a FLOGI frame that contains its Node Name, its N_Port Name, and service parameters to a well-known address of 0xFFFFFE.

A public loop NL_Port first opens the destination AL_PA 0x00 before issuing the FLOGI request. In both cases the switch accepts the login and returns an accept (ACC) frame to the sender. If some of the service parameters requested by the N_Port or NL_Port are not supported, the switch will set the appropriate bits in the ACC frame to indicate this.

When the N_Port logs in it uses a 24-bit port address of 0x000000. Because of this the fabric is allowed to assign the appropriate port address to that device, based on the Domain-Area-Port address format. The newly assigned address is contained in the ACC response frame.

When the NL_Port logs in a similar process starts, except that the least significant byte is used to assign AL_PA and the upper two bytes constitute a fabric loop identifier. Before an NL_Port logs in it will go through the LIP on the loop, which is started by the FL_Port, and from this process it has already

derived an AL_PA. The switch then decides if it will accept this AL_PA for this device or not. If not a new AL_PA is assigned to the NL_Port, which then causes the start of another LIP. This ensures that the switch assigned AL_PA does not conflict with any previously selected AL_PAs on the loop.

After the N_Port or public NL_Port gets its fabric address from FLOGI, it needs to register with the SNS. This is done with port login (PLOGI) at the address 0xFFFFFC. The device may register values for all or just some database objects, but the most useful are its 24-bit port address, 64-bit Port Name (WWPN), 64-bit Node Name (WWN), class of service parameters, FC-4 protocols supported, and port type, such as N_Port or NL_Port.

3.6.2 LPort login

Port login is also known as PLOGI.

Port login is used to establish a session between two N_Ports (devices) and is necessary before any upper level commands or operations can be performed. During the port login, two N_Ports (devices) swap service parameters and make themselves known to each other.

3.6.3 Process login

Process login is also known as PRLI. Process login is used to set up the environment between related processes on an originating N_Port and a responding N_Port. A group of related processes is collectively known as an image pair. The processes involved can be system processes, system images, such as mainframe logical partitions, control unit images, and FC-4 processes. Use of process login is optional from the perspective of Fibre Channel FC-2 layer, but may be required by a specific upper-level protocol as in the case of SCSI-FCP mapping.

3.7 Path routing mechanisms

A complex fabric can be made of interconnected switches and directors, perhaps even spanning a LAN/WAN connection. The challenge is to route the traffic with the minimum of overhead, latency, and reliability, and to prevent out-of-order delivery of frames. Here are some of the mechanisms.

3.7.1 Spanning tree

In case of failure, it is important to consider having an alternative path between source and destination available. This will allow the data still to reach its

destination. However, having different paths available could lead to the delivery of frames being out of the order of transmission, due to a frame taking a different path and arriving earlier than one of its predecessors.

A solution, which can be incorporated into the meshed fabric, is called a spanning tree and is an IEEE 802.1 standard. This means that switches keep to certain paths, as the spanning tree protocol will block certain paths to produce a simply connected active topology. Then the shortest path in terms of hops is used to deliver the frames and, most importantly, only one path is active at a time. This means that all associated frames go over the same path to the destination. The paths that are blocked can be held in reserve and used only if, for example, a primary path fails. The fact that one path is active at a time means that in the case of a meshed fabric, all frames will arrive in the expected order.

Path selection

For path selection, link state protocols are popular and extremely effective in today's networks. Examples of link state protocol are OSPF for IP and PNNI for ATM.

The most commonly used path selection protocol is Fabric Shortest Path First (FSPF). This type of path selection is usually performed at boot time, and no configuration is needed. All paths are established at start time, and only if the inter switch link (ISL) is broken or added will reconfiguration take place.

If multiple paths are available and if the primary path goes down, the traffic will be rerouted to another path. If the route fails, this can lead to congestion of frames, and any new frames delivered over the new path could potentially arrive at the destination first. This will cause an out-of-sequence delivery.

One possible solution for this is to prevent the activation of the new route for a while (this can be configured from milliseconds to a few seconds), so the congested frames are either delivered or rejected. Obviously, this can slow down the routing, so it should only be used when the devices connected to the fabric are not in a position to, or cannot tolerate, occasional out-of-sequence delivery. For instance, video can tolerate an out-of-sequence delivery, but financial and commercial data cannot.

But today, Fibre Channel devices are much more sophisticated, and this is a feature that is not normally required. FSPF allows a fabric to still benefit from load balancing the delivery of frames by using multiple paths.

We discuss FSPF in greater depth in “Fabric shortest path first” on page 63.

Route definition

Routes are usually dynamically defined. Static routes can also be defined. In the event that a static route fails, a dynamic route will take over. Once the static route becomes available again, frames will return to utilizing that route.

If dynamic paths are used, FSPF path selection is used. This guarantees that only the shortest and fastest paths will be used for delivering the frames.

3.7.2 Fabric Shortest Path First

According to the FC-SW-2 standard, Fabric Shortest Path First (FSPF) is a link state path selection protocol.

The concepts used in FSPF were first proposed by Brocade and have since been incorporated into the FC-SW-2 standard. Since then, it has been adopted by most, if not all, manufacturers. Certainly, all of the switches and directors in the IBM portfolio implement and utilize FSPF.

3.7.3 What is FSPF?

FSPF keeps track of the links on all switches in the fabric and associates a cost with each link. At the time of writing, the cost is always calculated as being directly proportional to the number of hops For example, in “Fabric shortest path first” on page 63, the path chosen is shown in RED.

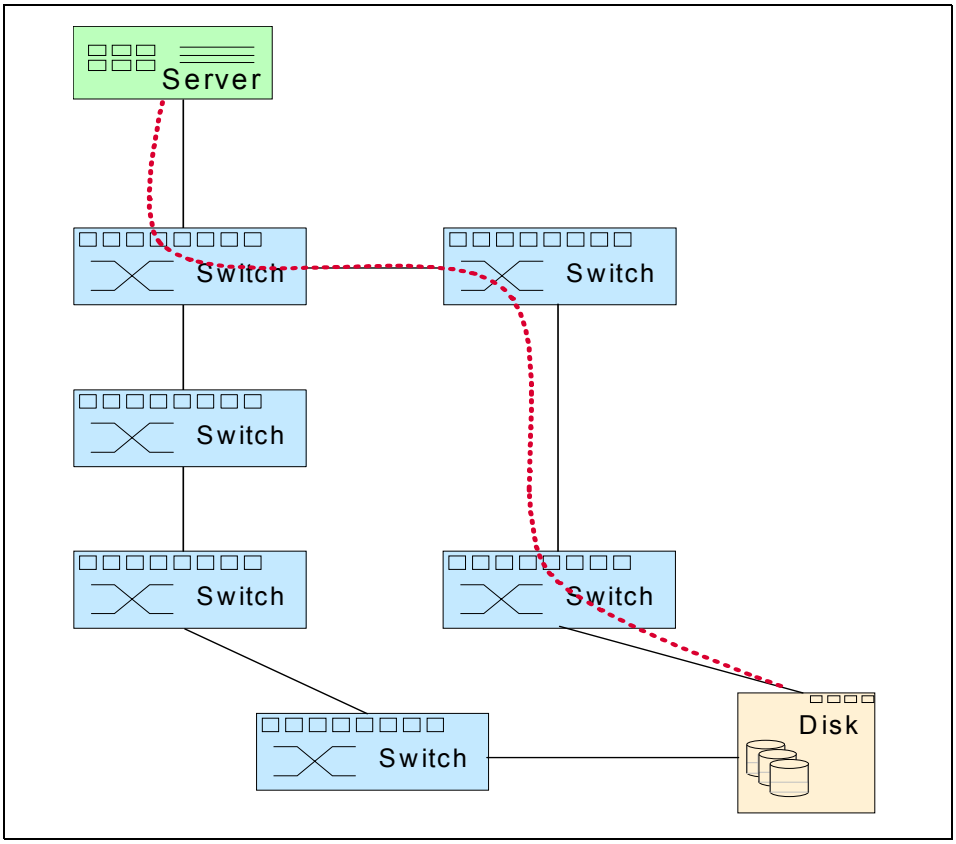


Figure 3-23 Fabric shortest path first

The protocol computes paths from a switch to all the other switches in the fabric by adding the cost of all links traversed by the path, and choosing the path that minimizes the cost.

For example, in Figure 3-24, if we need to connect a port in switch A to a port in switch D, it will take the ISL from A to D.

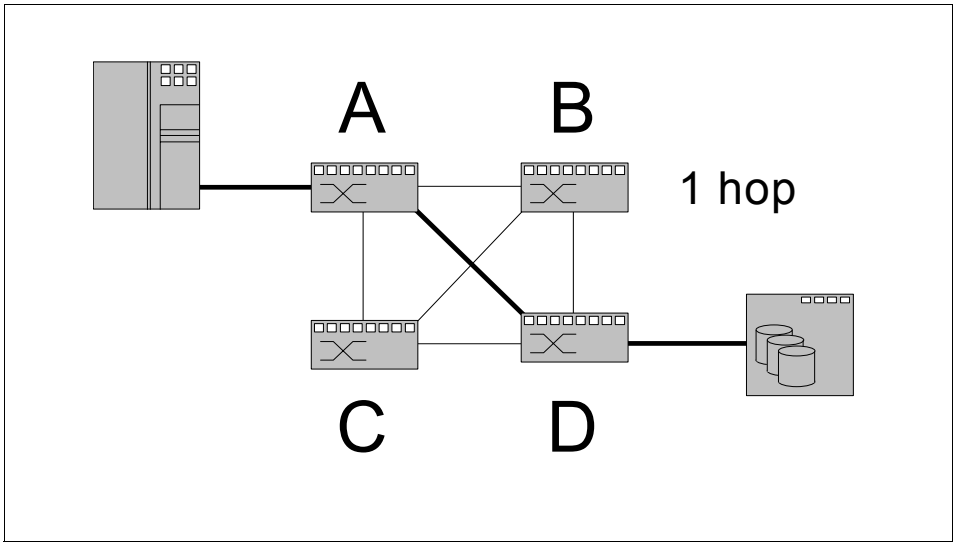


Figure 3-24 FSPF calculates the route taking the least hops

The other possible paths are shown in Figure 3-25.

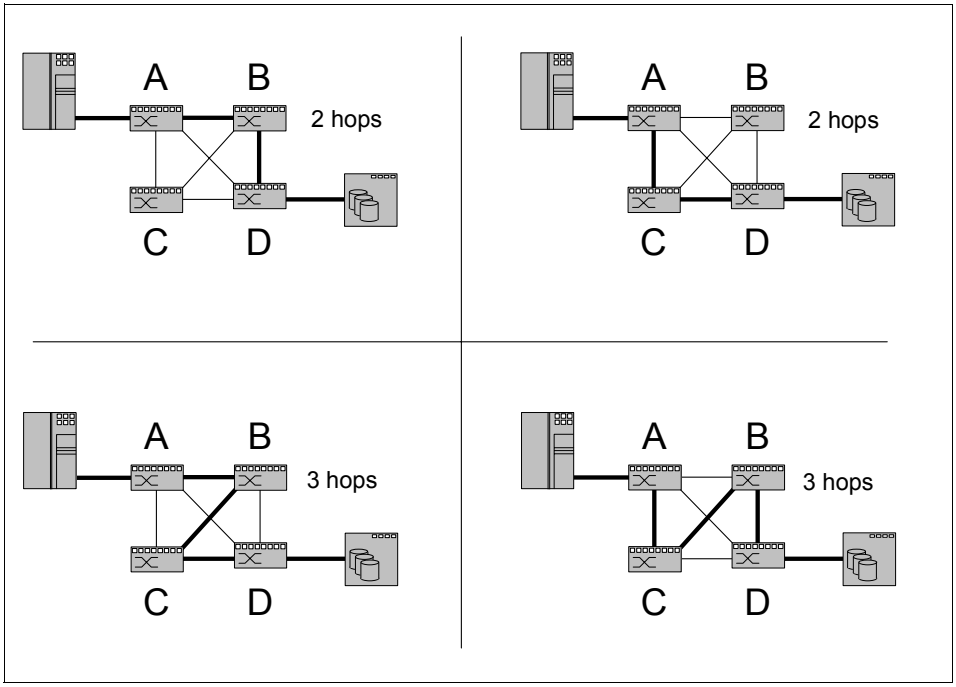


Figure 3-25 Other possible paths

It will not go from A to B to D, neither from A to C to D. This is because FSPF is currently based on the hop count cost.

3.7.4 How does FSPF work?

The collection of link states (including cost) of all switches in a fabric constitutes the topology database (or link state database). The topology database is kept in all switches in the fabric and they are maintained and synchronized to each other. There is an initial database synchronization, and an update mechanism. The initial database synchronization is used when a switch is initialized, or when an ISL comes up. The update mechanism (see 3.5.5, “Registered State Change Notification” on page 86) is used when there is a link state change, for example, an ISL going down or coming up, and on a periodic basis. This ensures consistency among all switches in the fabric.

3.7.5 How does FSPF help?

In the situation where there are multiple routes, FSPF will ensure that the route that is used is the one with the lowest number of hops. If all the hops:

- Have the same latency
- Operate at the same speed
- Have no congestion

then FSPF will ensure that the frames get to their destinations by the fastest route.

3.7.6 What happens when there is more than one shortest path?

If we look again at the example in Figure 3-24, and we imagine that the link from A to D goes down, switch A will now have four routes to reach D:

- A-B-D
- A-C-D
- A-B-C-D
- A-C-B-D

A-B-D and A-C-D will be selected because they are the equal shortest paths based on the hop count cost. The update mechanism ensures that switches B and C will also have their databases updated with the new routing information.

So, which of the two routes will be used? The answer is that the decision of which way to send a frame is up to the manufacturer of each switch. In our case, Switch B and Switch C will send frames directly to Switch D. The firmware in Switch A will make a decision about which way to send frames to Switch D, either via Switch B or Switch C. The way that this decision is made is by a round robin algorithm based on the order of connection. Let us consider the situation illustrated in Figure 3-26.

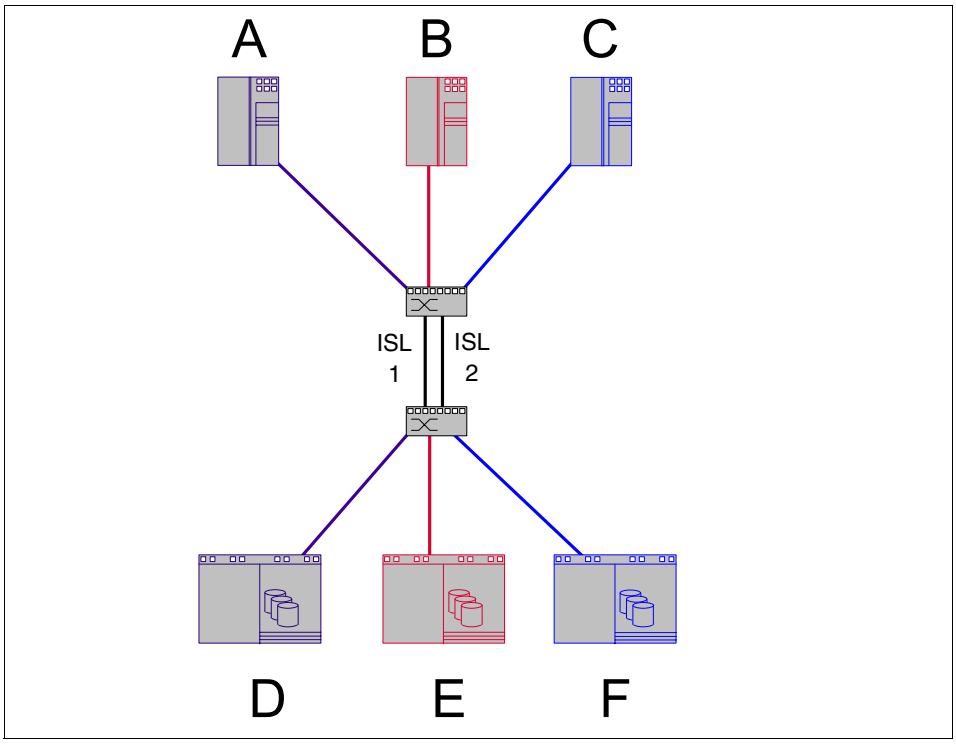


Figure 3-26 FSPF and round robin

There are three servers A, B and C which all need to communicate with the storage devices D, E and F respectively (We are assuming that there is no zoning or trunking enabled, and that all of the links are operating at the same bandwidth.

Let us assume that the three servers connect in the order A then B then C. Server A will be given a route from the upper switch to the lower switch. For the sake of this example, let us assume that it is via ISL1. The second server, Server B in the example will be assigned a route via ISL2. and the Server C will have a route via ISL1. This will have the result of sharing the load between the two switches over the two ISLs.

Important: This implements load sharing but not load balancing.

We can see that some traffic will flow via each of the ISLs, but we must stress that this is not the same as load balancing.

3.7.7 Can FSPF cause any problems?

There are some occasions when FSPF does not produce an ideal situation.

Oversubscription

Let us consider the diagram in Figure 3-27.

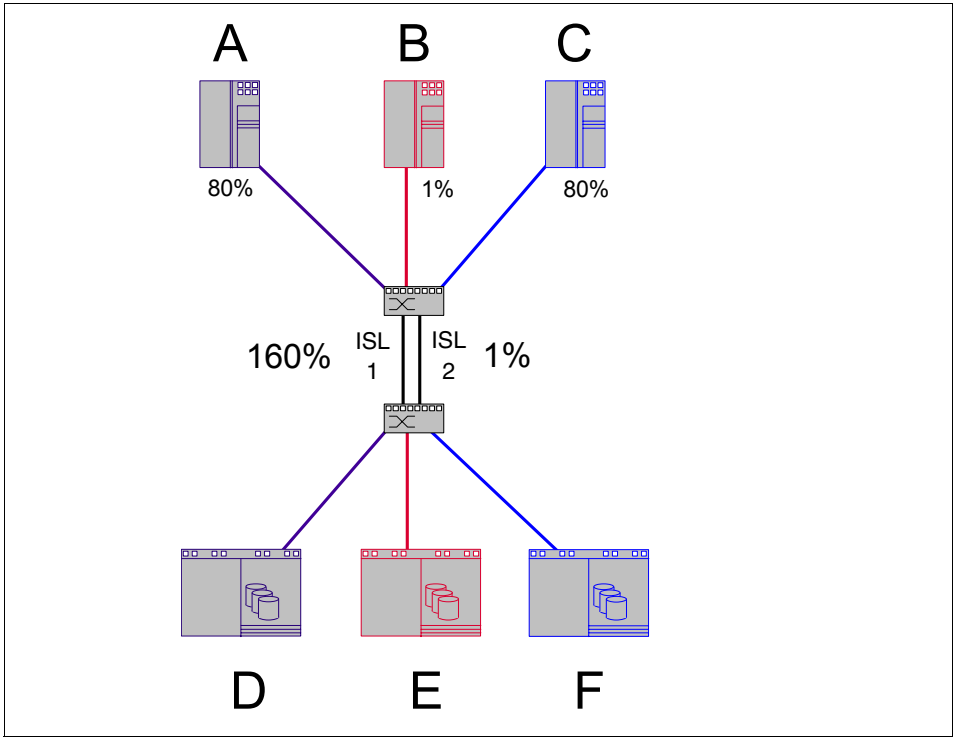


Figure 3-27 FSPF can lead to oversubscription and congestion

In this scenario, Server A and C have routes to their storage via ISL1 and Server B sends its frames via ISL2. These routes were assigned by FSPF, as discussed above. As we can see, Servers A and C are trying to utilize 80% of the maximum bandwidth of the links, but server B has a much lighter I/O requirement. We can see that ISL1 is very much oversubscribed and that ISL2 is hardly utilized at all.

This feature is something that switch and director manufacturers are aware of. There is nothing in the specification of FSPF to work around this problem. Possible solutions are being worked on and may become available in the future. Until then, possible solutions include the use of zoning and trunking. For more information on zoning, see 3.8, “Zoning” on page 104. Trunking is discussed in 3.9, “Trunking” on page 112.

Length and speed of hops

When FSPF is counting the cost of possible routes, all it considers is the number of hops.

It could be that a particular route has, two hops and an alternative has one hop. At first sight, the one hop route would seem to be better, and that is what FSPF will decide. Let us look at Figure 3-28, however.

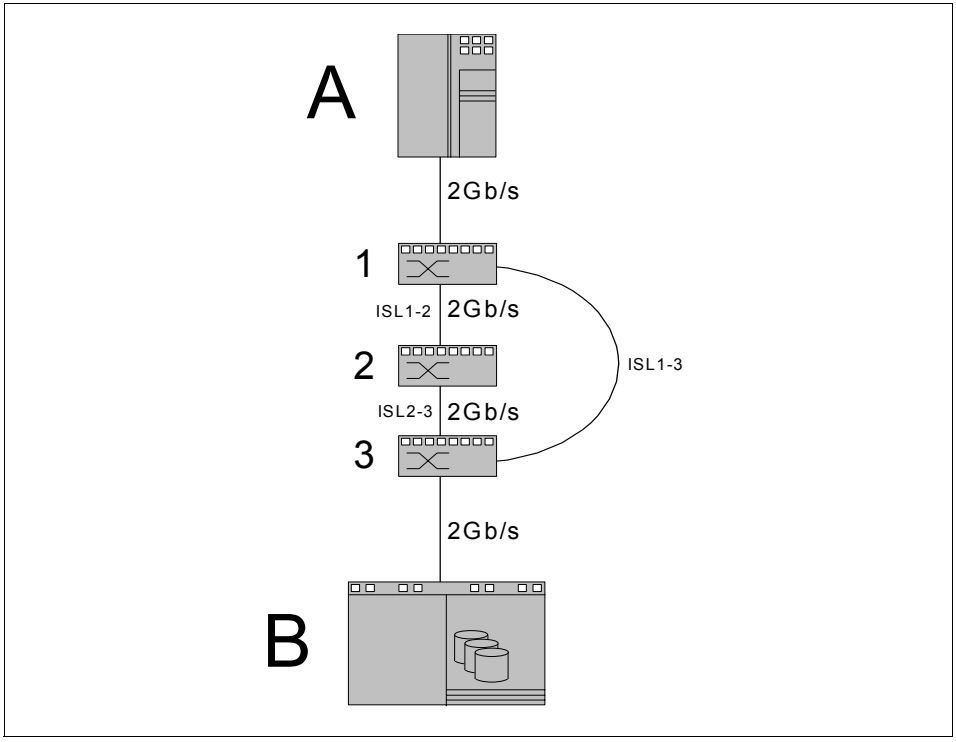


Figure 3-28 Hops and their cost, speed

Clearly, the path from A to B via ISL1-3 uses a single hop, and the path via Switch 2 takes two hops, through ISL1-2 and ISL2-3. FSPF will select the path via ISL1-3.

If ISL1-3 is running at 2 Gb/s (or faster see 3.9, “Trunking” on page 112, and 3.7.9, “1 Gb/s, 2 Gb/s and beyond” on page 98) then the fastest path will be via ISL1-3, so FSPF will be giving us the fastest path as well as the shortest path.

If on the other hand, the ISL is running at 1 Gb/s or slower then it would actually be better to use the other route, via Switch 2. In this case FSPF would not be giving us the fastest route!

Even if ISL1-3 is running at 2 Gb/s it might be better to use the path via Switch 2. For example if ISL1-3 is very long and ISL1-2 and ISL2-3 are very short, then the added latency in ISL1-3 may cause a significant delay.

Note: The rule is that the latency per 1 Km of fibre cable is 5 microseconds.

So, if for example ISL1-3 is about 100 km long it introduces a latency of 500µs, whereas the typical latency through a switch or director is much less than 5µs. We can make a reasonable approximation that in this case the *shortest* path has about 100 times the latency of the *longest* route!

This particular scenario is unlikely to occur in the real world, but it illustrates the point.

Getting around these problems

The switch and director manufacturers are aware of these problems and are trying to produce a mechanism for ensuring that the route chosen is actually the best one. The areas that they are working on include:

- Manually assigning a notional cost to each ISL
- Manually forcing a static route

3.7.8 FC-PH-2 and speed

Present day Fibre Channel devices generally operate at 2 Gb/s, that is, at two gigabits per second. What does this actually mean?

Data flows from the transducer (GBIC, GLM, transceiver) at a rate of 2 Gb/s. so, for example, in the case of a GBIC we actually tend to have a dataflow of 2,125,000,000 bits per second. (various speed GBICs are available but that is the normal speed). Those are *optical* bits. In other words it is bits after 8b/10b encoding has taken place, so that equates to 212,500,000 bytes. These are 10 bit bytes on the fiber but would be 8-bit bytes at the application level.

When data is sent over the fibre, it is carried as the payload of a frame. The maximum payload being 2112 bytes. The frame also occupies 36 bytes for the start of frame, frame headers, CRC, and end of frame.

So, we can see that in fact to send 2112 bytes of data we actually send 2148 bytes altogether. There will also be some IDLEs, R_RDYs or other primitive signals between each frame (FC-PH specifies that an N_Port will transmit a minimum of six primitive signals between consecutive frames). There may also be the need for each frame to be acknowledged by a special frame called an ACK. An ACK frame does not carry any data but its headers define which particular frame that it is acknowledging.

The details are as follows:

- Speed
  - 2,125,000,000 bits per second
  - 212,500,000 bytes per second
- Data payload and frame size
  - Frame length is payload size + 36 bytes
  - Maximum data payload 2112 bytes
  - Typical data payload is 2048 bytes
- Size of a Primitive Signal, for example IDLE or R_RDY
  - 4 bytes
  - There will be a minimum of six between frames
- Size of an ACK frame
  - 36 bytes
- Total overhead in bytes (without acknowledgment)
  - 36 for frame overhead
  - 24 for primitive signals after the data frames
  - Total overhead is 36 + 24 = 60 bytes
- Total overhead in bytes (with acknowledgment)
  - 36 for frame overhead
  - 36 for ACK frame
  - 24 for primitive signals after each of the data and ACK frames
  - Total overhead is 36 + 36 + 24 + 24 = 120 bytes

This allows us to build up the data shown in Table 3-2.

Table 3-2 Application bytes per second on 2 Gb/s link using 2048-byte payload

Acknowledged	Overhead Bytes	Bytes/sec	Megabytes/second
No	60	206,451,613	196.88
Yes	120	200,738,006	191.42

So, when the term 200 MB/s is used, it is a generic term for the goal throughput. If you prefer marketing (1000 x 1000) megabytes, then there is a maximum possible throughput of over 200 MB/s whether frames are acknowledged or not. If you prefer real computer (1024 x 1024) megabytes, then the maximum possible throughput is about 191.42 MB/s for an acknowledged class of service. This is still not a bad throughput, but, arguably, not 200 MB/s.

These are not the actual figures that we would expect to be achieved in the real world, but are theoretical figures. There are external factors which will determine the actual, sustained throughput which can be achieved on a given link. These will include, for example:

- The efficiency of the software or firmware stack in a server or disk array
- The bandwidth of the bus that the HBA is plugged into
- Any other bottleneck in the fabric

A typical throughput might be 160 MB/s but this is by no means guaranteed, it might be higher or it might be lower.

3.7.9 1 Gb/s, 2 Gb/s and beyond

The FC-PH document describes and defines communication at 1 Gb/s, and the FC-PH-2 document extends on this and defines 2 Gb/s. In fact the speeds are 1,062,500,000 and 2,125,000,000. FC-PH-2 goes even further and defines 4,250,000,000 as well.

While 2 Gb/s is standard at the time of writing, 4Gb/s is an emerging speed and manufacturers are beginning to implement it on their new products, or by making upgrades to their existing products. Many manufacturers are actively working towards 10 Gb/s.

Interoperability

As the transition was made from 1 Gb/s to 2 Gb/s a natural consideration was whether or not it was possible to have both 1Gb/s and 2Gb/s components in the same SAN.

The answer is yes, *but*. By design and definition, there should be no problem. The way that devices communicate when first connected allows for each node to declare certain parameters including the maximum and minimum speeds that they can communicate at. The normal procedure being to both agree to use the highest common speed. Thus a 1 Gb/s GBIC in a node, connected to a 2 Gb/s transceiver in a switch should operate happily at 1 Gb/s. Equally, it should be fine to operate the other way around, that is, with a 2 Gb/s transceiver in a node and a 1 Gb/s GBIC in a switch.

So, we now ask, can we have a 2 Gb/s connection into a switch or director with a 1 Gb/s connection into storage, and if so, what do we need to consider?

Firstly, let us see Figure 3-29.

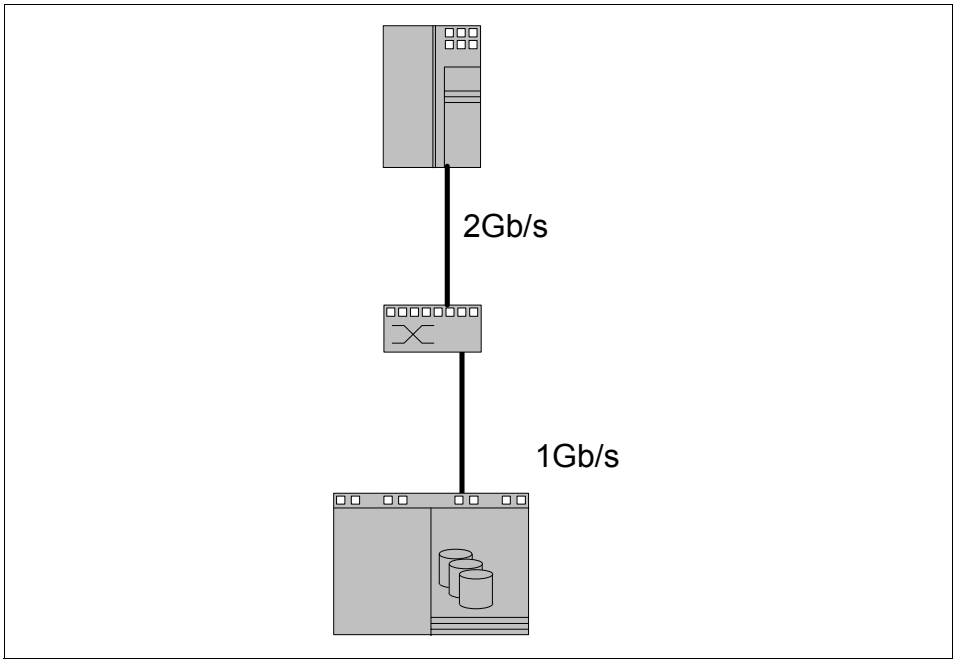


Figure 3-29 Mixing 2 Gb/s and 1 Gb/s

In this example, the host connection has negotiated a 2 Gb/s link and the server has negotiated a 1Gb/s link.

This is perfectly acceptable under the rules of Fibre Channel. The data throughput between this particular server and the storage will be at 100 MB/s maximum due to the 1 Gb/s link, however the communication between the server and the switch can happen faster. This might mean that the server sees a busy port at the switch, but flow control should handle this using BB_Credit (FC-PH, FC-PH-2, and FC-PH-3).

3.7.10 FC-PH, FC-PH-2, and FC-PH-3

The American National Standards Institute T11 technical committee documents and defines standards for the Fibre Channel world.

The FC-PH documents define standards for FC-0, FC-1 and FC-2 (see 3.7.11, “Layers” on page 101). There are several documents covering the development of Fibre Channel Physical Hardware. These include FC-PH, FC-PH-2.

The original FC-PH document defined the original standard and covered speeds up to 100 MB/s of bandwidth in each direction over a full duplex fiber.

The FC-PH-2 document covers additional speeds of 200MB/s and 400MB/s. There are also some other enhanced features defined in the document.

The FC-PH-3 document covers some enhancements to both the FC-PH and FC-PH-2 documents.

FC-PH

This is the document which defines the Physical Hardware of Fibre Channel. It is the basis on which the technology has been based, and is growing.

It defines many features including:

- The way that the ports and the components operate at a very low level.
- The three initial Classes of communication are defined (see also Table 3.2 on page 67).
- The signalling layers (see also “Physical and signaling layers” on page 102)
- Speeds up to and including 100 MB/s
- Frames, Sequences, and Exchanges (see also 3.10, “Ordered set, frames, sequences, and exchanges” on page 119.)

FC-PH-2

In addition to the extra speeds of 200 MB/s and 400 MB/s FC-PH-2 also defines the following extra features, including:

- Hunt Groups
- Multicast
- Dedicated simplex
- Fractional bandwidth

Not all equipment will support the features of FC-PH-2. Many parts will support some of the features, but not all.

Hunt Groups

A Hunt Group is a set of one or more N_Ports, which may be addressed using a Hunt Group Identifier (HG_ID). This allows the fabric to select any one of the N_Ports in the Hunt Group as the destination of a frame. This means that we may achieve either or both increased bandwidth and reduced latency.

Multicast

This allows for a multicast service. It is based on Class 3 Fibre Channel communications and is unacknowledged. A frame which is sent to a Multicast Group which has a Multicast Group Identifier (MG_ID) will be delivered to every N_Port in the group.

Dedicated simplex

The type of connection defined in FC-PH is duplex or bidirectional and end to end. Thus any Fibre Channel communication is effectively point to point, no matter what the intervening Fabric may be. There is a cunning extension to this concept defined in FC-PH-2. It allows a particular N_Port to send data outbound to one N_Port whilst, at the same time, receive inbound data from a different N_Port. This tends to increase the overall throughput of data by increases the chances of bidirectional transfer.

Fractional bandwidth

This is Class 4 and is discussed in “Class 4” on page 69.

FC-PH-3

As well as some enrichments and advancements to the functions defined in the previous specifications, the main new feature in this document is Class 6 communication. For a further description of this, see “Class 6” on page 69.

It has been known to make Unix machines pretend to be printers or card punchers and readers so that they can communicate with mainframes.

3.7.11 Layers

Fibre Channel (FC) is broken up into a series of five layers. The concept of layers, starting with the ISO/OSI seven-layer model, allows the development of one layer to remain independent of the adjacent layers. Although, FC contains five layers, those layers follow the general principles stated in the ISO/OSI model.

The five layers are divided into two:

- Physical and signaling layer
- Upper layer

The five layers are illustrated in Figure 3-30.

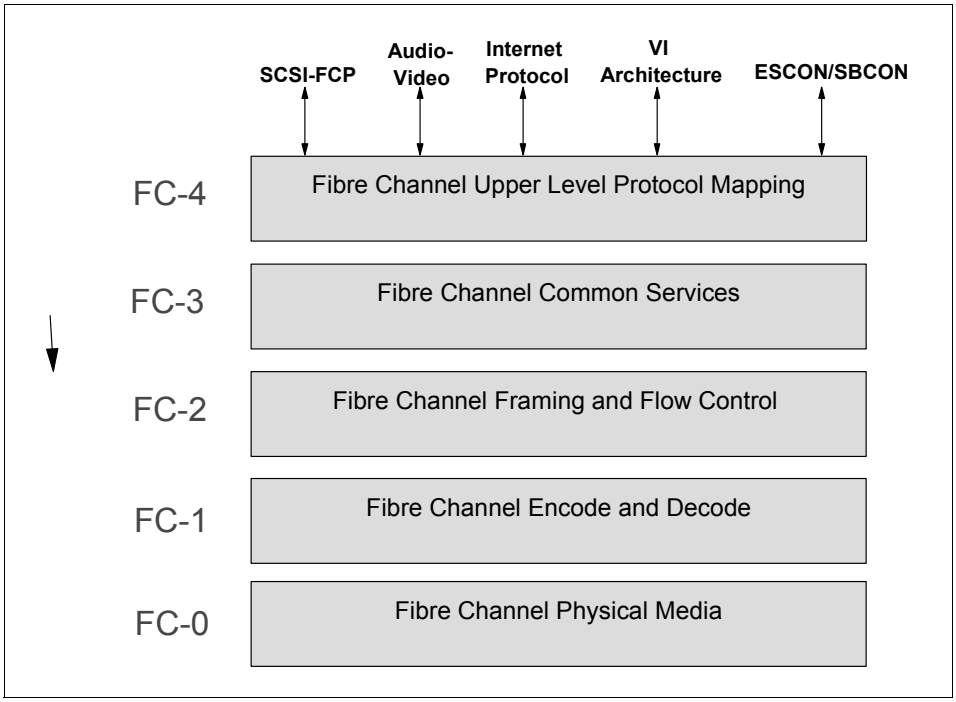


Figure 3-30 Fibre Channel layers

Physical and signaling layers

The physical and signaling layers include the three lowest layers: FC-0, FC-1, and FC-2.

Physical interface and media: FC-0

The lowest layer (FC-0) defines the physical link in the system, including the cabling, connectors, and electrical parameters for the system at a wide range of data rates. This level is designed for maximum flexibility, and allows the use of a large number of technologies to match the needs of the desired configuration.

A communication route between two nodes may be made up of links of different technologies. For example, in reaching its destination, a signal may start out on copper wire and become converted to single-mode fibre for longer distances. This flexibility allows for specialized configurations depending on IT requirements.

Laser safety

Fibre Channel often uses lasers to transmit data, and can, therefore, present an optical health hazard. The FC-0 layer defines an open fibre control (OFC) system, and acts as a safety interlock for point-to-point fibre connections that use semiconductor laser diodes as the optical source. If the fibre connection is broken, the ports send a series of pulses until the physical connection is re-established and the necessary handshake procedures are followed.

Transmission protocol: FC-1

The second layer (FC-1) provides the methods for adaptive 8B/10B encoding to bind the maximum length of the code, maintain DC-balance, and provide word alignment. This layer is used to integrate the data with the clock information required by serial transmission technologies.

Framing and signaling protocol: FC-2

Reliable communications result from Fibre Channel's FC-2 framing and signaling protocol. FC-2 specifies a data transport mechanism that is independent of upper layer protocols. FC-2 is self-configuring and supports point-to-point, Arbitrated Loop, and switched environments.

FC-2, which is the third layer of the FC-PH, provides the transport methods to determine:

- Topologies based on the presence or absence of a fabric
- Communication models
- Classes of service provided by the fabric and the nodes
- General fabric model
- Sequence and exchange identifiers
- Segmentation and reassembly

Data is transmitted in 4-byte ordered sets containing data and control characters. Ordered sets provide the availability to obtain bit and word synchronization, which also establishes word boundary alignment.

Together, FC-0, FC-1, and FC-2 form the Fibre Channel physical and signaling interface (FC-PH).

Upper layers

The Upper layer includes two layers: FC-3 and FC-4.

Common services: FC-3

FC-3 defines functions that span multiple ports on a single-node or fabric. Functions that are currently supported include:

- Hunt Groups:

- A Hunt Group is a set of associated N_Ports attached to a single node. This set is assigned an alias identifier that allows any frames containing the alias to be routed to any available N_Port within the set. This decreases latency in waiting for an N_Port to become available.

- **Striping:**
  - Striping is used to multiply bandwidth, using multiple N_Ports in parallel to transmit a single information unit across multiple links.

- **Multicast:**
  - Multicast delivers a single transmission to multiple destination ports. This includes the ability to broadcast to all nodes or a subset of nodes.

Upper layer protocol mapping (ULP): FC-4

The highest layer (FC-4) provides the application-specific protocols. Fibre Channel is equally adept at transporting both network and channel information and allows both protocol types to be concurrently transported over the same physical interface.

Through mapping rules, a specific FC-4 describes how ULP processes of the same FC-4 type interoperate.

A channel example is Fibre Channel Protocol (FCP). This is used to transfer SCSI data over Fibre Channel, while a networking example is sending IP (Internet Protocol) packets between nodes. FICON is another ULP in use today for mainframe systems. The term is a contraction of FIbre CONnection and refers to running ESCON traffic over Fibre Channel

3.8 Zoning

Zoning allows for finer segmentation of the switched fabric. Zoning can be used to instigate a barrier between different environments. Only the members of the same zone can communicate within that zone and all other attempts from outside are rejected.

For example, it may be desirable to separate a Windows NT environment from a UNIX environment. This is very useful because of the manner in which Windows attempts to claim all available storage for itself. Because not all storage devices are capable of protecting their resources from any host seeking available resources, it makes sound business sense to protect the environment in another manner.We show an example of zoning in Figure 3-31 on page 105 where we have separated AIX® from NT and created Zone 1 and Zone 2. This diagram also shows how a device can be in more than one zone.

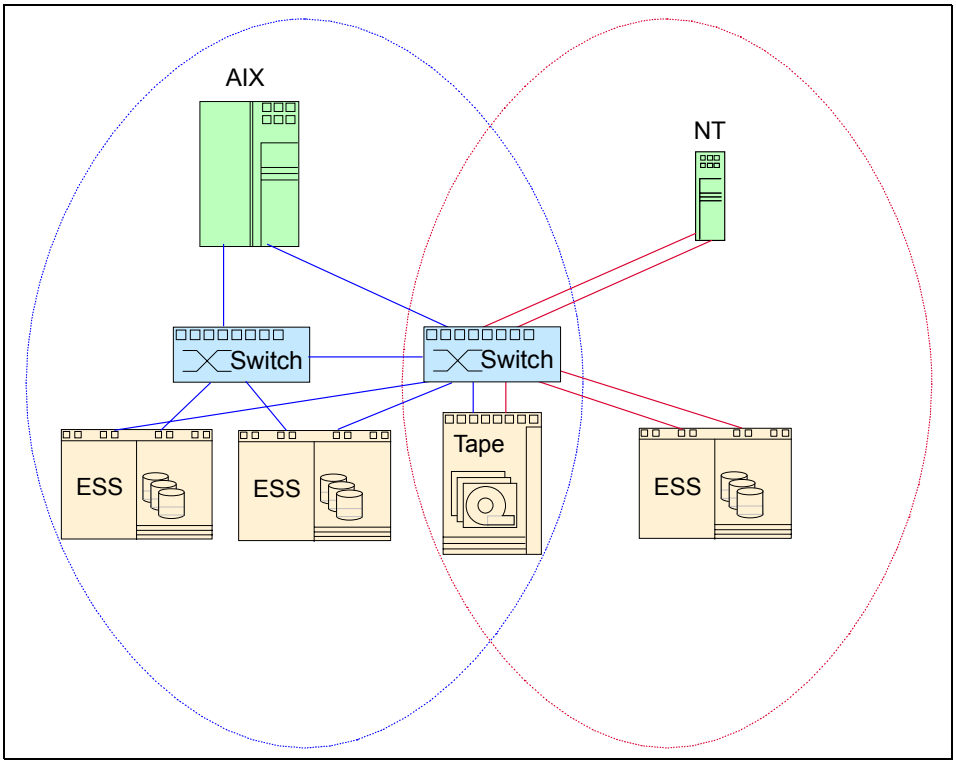


Figure 3-31 Zoning

Looking at zoning in this way, it could also be considered as a security feature and not just for separating environments. Zoning could also be used for test and maintenance purposes. For example, not many enterprises will mix their test and maintenance environments with their production environment. Within a fabric, you could easily separate your test environment from your production bandwidth allocation on the same fabric using zoning.

An example of zoning is shown in Figure 3-32. In this case:

- Server A and Storage A can communicate with each other.
- Server B and Storage B can communicate with each other.
- Server A cannot communicate with Storage B.
- Server B cannot communicate with Storage A.
- Both servers and both storage devices can communicate with the tape.

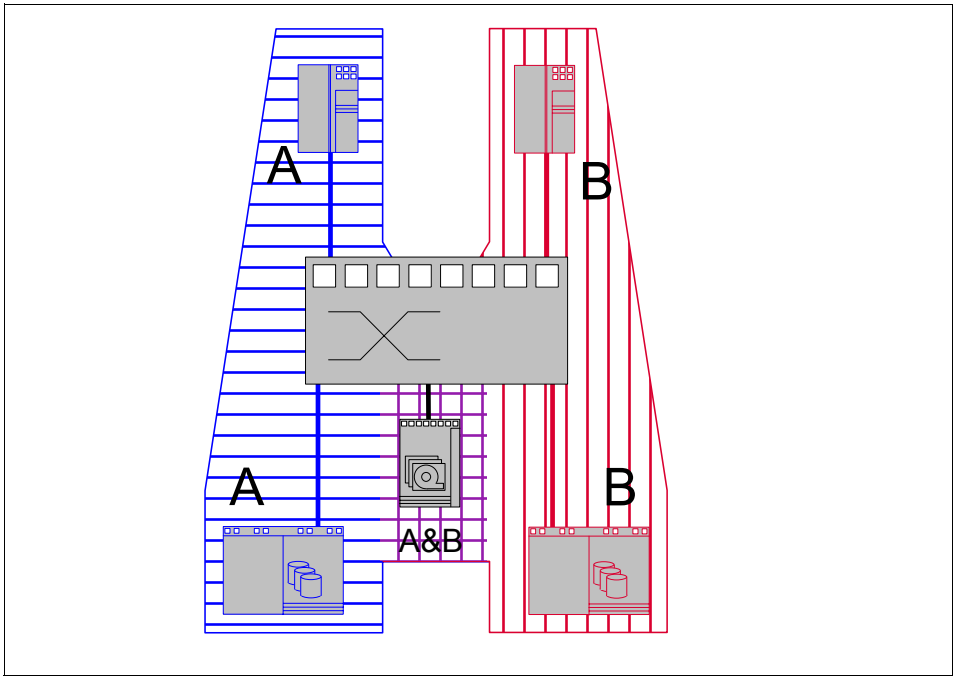


Figure 3-32 An example of zoning

Zoning also introduces the flexibility to manage a switched fabric to meet different user groups objectives.

Zoning can be implemented in two ways:

- Hardware zoning
- Software zoning

These forms of zoning are different, but are not necessarily mutually exclusive. Depending upon the particular manufacturer of the SAN hardware, it is possible for hardware zones and software zones to overlap. This adds to the flexibility, but can make the solution complicated increasing the need for good management software and documentation of the SAN.

3.8.1 Hardware zoning

Hardware zoning is based on the physical fabric port number. The members of a zone are physical ports on the fabric switch. It can be implemented in the following configurations:

- One-to-one
- One-to-many

- Many-to-many

Figure 3-33 on page 107 shows an example of zoning based on the switch port numbers.

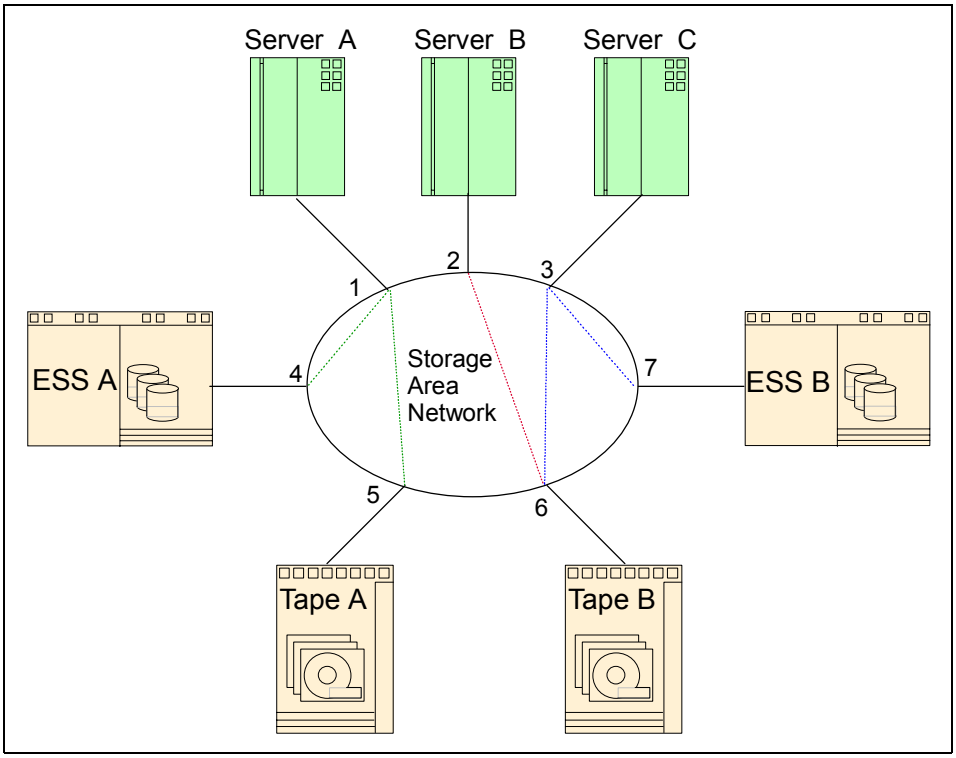


Figure 3-33 Zoning based on the switch port-number

In this example, port-based zoning is used to restrict Server A to only see storage devices that are zoned to port 1, that is, ports 4 and 5.

Server B is also zoned so that it can only see from port 2 through to port 6.

Server C is zoned so that it can see both ports 6 and 7 even though port 6 is also a member of another zone.

A single port can also belong to multiple zones.

We show an example of hardware zoning in Figure 3-34. This example illustrates another way of considering the hardware zoning as an array of connections. This technique is utilized for the concept of port prohibits and port allows in “IBM TotalStorage SAN n-type family and replication appliance” on page 345.

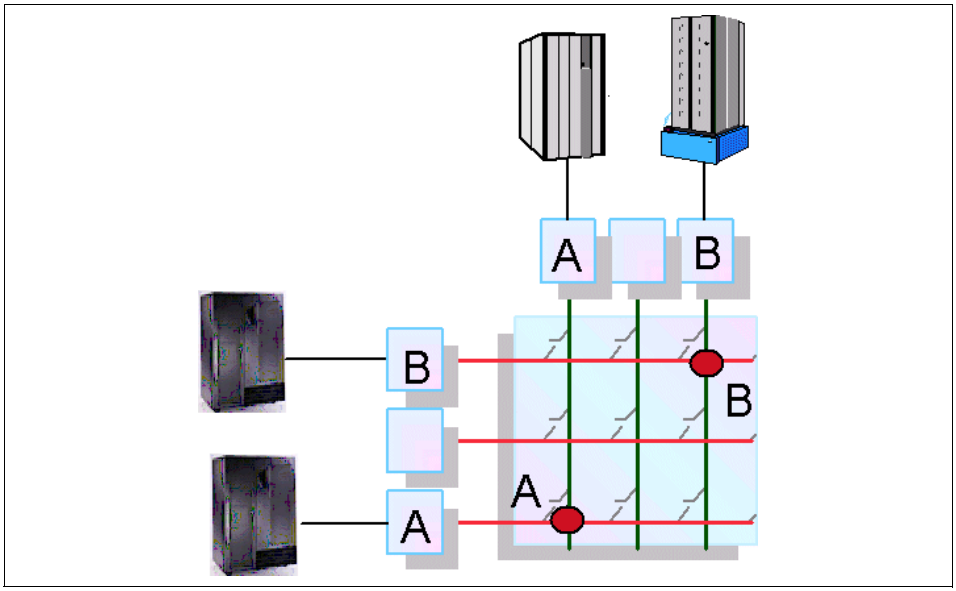


Figure 3-34 Hardware zoning

In this example, device A can only access storage device A through connection A; and device B can only access storage device B through connection B.

In a hardware enforced zone, switch hardware (usually at the ASIC level) ensures that there is no data transferred between unauthorized zone members. However, devices can transfer data between ports within the same zone. Consequently, hard zoning provides the highest level of security. The availability of hardware enforced zoning and the methods to create hardware enforced zones depends on the switch hardware used.

One of the disadvantages of hardware zoning is that devices have to be connected to a specific port, and the whole zoning configuration could become unusable when the device is connected to a different port. In cases where the device connections are not permanent the use of software zoning is recommended.

The advantage of hardware zoning is that it can be implemented into a routing engine by filtering. As a result, this kind of zoning has a very low impact on the performance of the routing process.

If possible the designer may include some unused ports into a hardware zone. So, in the event of a particular port failing (maybe a caused by a GBIC problem), the cable could be moved to a different port in the same zone. This would mean that the zone would not need to be reconfigured.

3.8.2 Software zoning

Software zoning is implemented by the fabric operating systems within the fabric switches. When using software zoning the members of the zone can be defined using their World Wide Names:

- Node WWN
- Port WWN

Usually, zoning software also allows you to create symbolic names for the zone members and for the zones themselves. Dealing with the symbolic name or aliases for a device is often easier than trying to use the WWN address.

The number of members possible in a zone is limited only by the amount of memory in the fabric switch. A member can belong to multiple zones. You can define multiple sets of zones for the fabric, but only one set can be active at any time. You can activate another zone set any time you want, without the need to power down the switch.

With software zoning there is no need to worry about the physical connections to the switch. If you use WWNs for the zone members, even when a device is connected to another physical port, it will still remain in the same zoning definition, because the device's WWN remains the same. The zone follows the WWN.

Shown in Figure 3-35 is an example of WWN based zoning. In this example symbolic names are defined for each WWN in the SAN to implement the same zoning requirements, as shown in the previous Figure 3-33 on page 107for port zoning:

- **Zone_1** contains the aliases **alex**, **ben**, and **sam**, and is restricted to only these devices.
- **Zone_2** contains the aliases **robyn** and **ellen**, and is restricted to only these devices.
- **Zone_3** contains the aliases **matthew**, **max**, and **ellen**, and is restricted to only these devices.

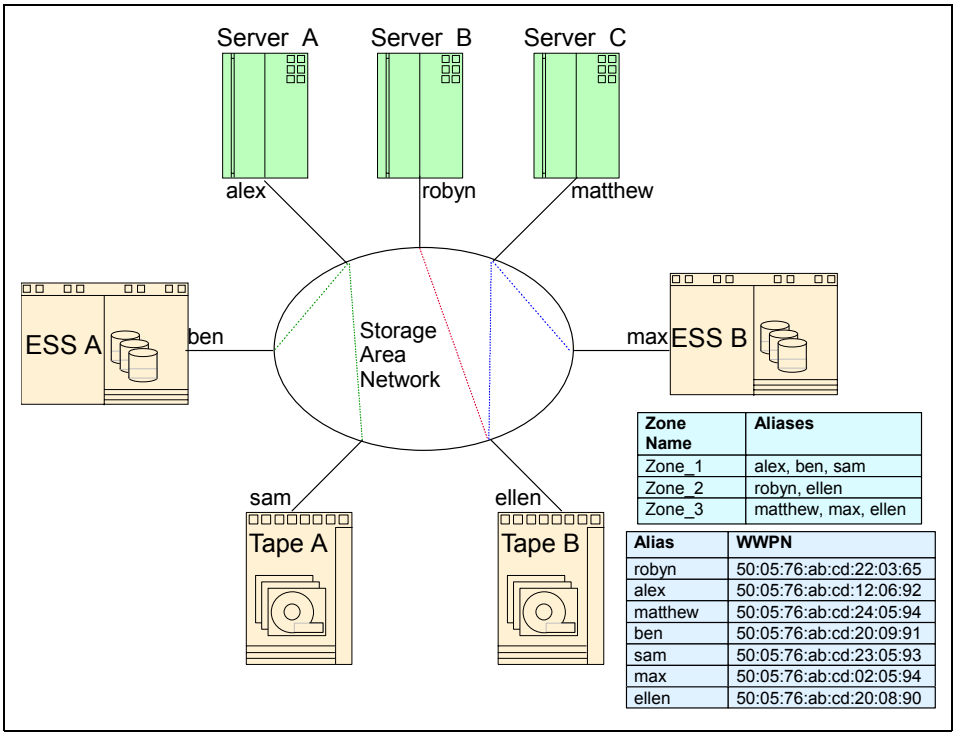


Figure 3-35 Zoning based on the devices WWN

There are some potential security leaks with software zoning:

- When a specific host logs into the fabric and asks for available storage devices, the SNS will look into the software zoning table to see which storage devices are allowable for that host. The host will only see the storage devices defined in the software zoning table. But, the host can also make a direct connection to the storage device, while doing device discovery, without asking SNS for the information it has.
- It is possible for a device to define the WWN that it will be use, rather than using the one designated by the manufacturer of the HBA. This is known as WWN spoofing. So, an unknown server could masquerade as a trusted server and thus gain access to data on a particular storage device. Some fabric operating systems will allow the fabric administrator to prevent this risk by allowing the WWN to be tied to a particular port.
- Any device that does any form of probing for WWNs may be able to discover devices and talk to them. A simple analogy would be that of an unlisted telephone number where, although the telephone number is not publicly available, there is nothing to stop a person dialling that number whether by design or accident. The same holds true for WWN and there

are devices that will randomly probe for WWNs to see if they can start a conversation with them.

A number of switch vendors offer hardware enforced WWN zoning, which can prevent this security exposure. Hardware Enforced zoning utilizes hardware mechanisms to restrict access rather than relying on the servers to follow the fibre channel protocols.

Note: In a software enforced zone, when a device logs in, it queries the name server for devices within the fabric. If zoning is in effect, only the devices in the same zone(s) are returned. Other devices are hidden from the name server query reply. When using software enforced zones, the switch does not control data transfer and there is no guarantee of data being transferred from unauthorized zone members. Use software zoning where flexibility and security are ensured by the cooperating hosts.

VSANs and logical domains

Another layer of partitioning and separating fabric traffic is also available on some switches. The Cisco MDS 9000 family of switches offers virtual SANs (VSAN). Each port in the fabric can be identified to be in a certain virtual SAN. Multiple logical SANS can consist on the fabric. The VSANs are isolated via hardware enforced zoning from each other. VSANs can stretch across switches via ISLs and Cisco considers these trunks.

THE CNT Ultranet Multi-service Director offers logical domains to partition a director into a maximum of four isolated domains enforced via hardware.The CNT logical domains are partitionable on a I/O blade basis, and can be maintained or upgraded separately from each other.

The Cisco VSAN can consist of any ports in the fabric, which is more flexible for configuring, but the switch is a single unit for maintenance and upgrades.

LUN masking

The term logical unit number (LUN) was originally used to represent the entity within a SCSI target which actually executes I/Os. A single SCSI device would usually only have a single LUN but some devices; for example, tape libraries might have more than one LUN.

In the case of a storage array, the array makes virtual disks available to servers. These virtual disks are identified by logical unit numbers.

It is absolutely possible for more than one host to see the same storage device or LUN. This is potentially a problem, both from a practical and security perspective. Another approach to securing storage devices from hosts wishing to take over

already assigned resources is logical unit number (LUN) masking. Every storage device offers its resources to the hosts by means of LUNs.

For example, each partition in the storage server has its own LUN. If the host (server) wants to access the storage, it needs to request access to the LUN in the storage device. The purpose of LUN masking is to control access to the LUNs. The storage device itself accepts or rejects access requests from different hosts.

The user defines which hosts can access which LUN by means of the storage device control program. Whenever the host accesses a particular LUN, the storage device will check its access list for that LUN, and it will allow or disallow access to the LUN.

3.9 Trunking

Trunking is a feature of switches that enables traffic to be distributed across available inter-switch links (ISLs) while still preserving in-order delivery. On some Fibre Channel protocol devices, frame traffic between a source device and destination device must be delivered in order within an exchange.

This restriction forces current devices to fix a routing path within a fabric. Consequently, certain traffic patterns in a fabric can cause all active routes to be allocated to a single available path and leave other paths unused. Trunking creates a trunking group (a set of available paths linking two adjacent switches). Ports in the trunking group are called trunking ports.

We illustrate the concepts of trunking in Figure 3-36.

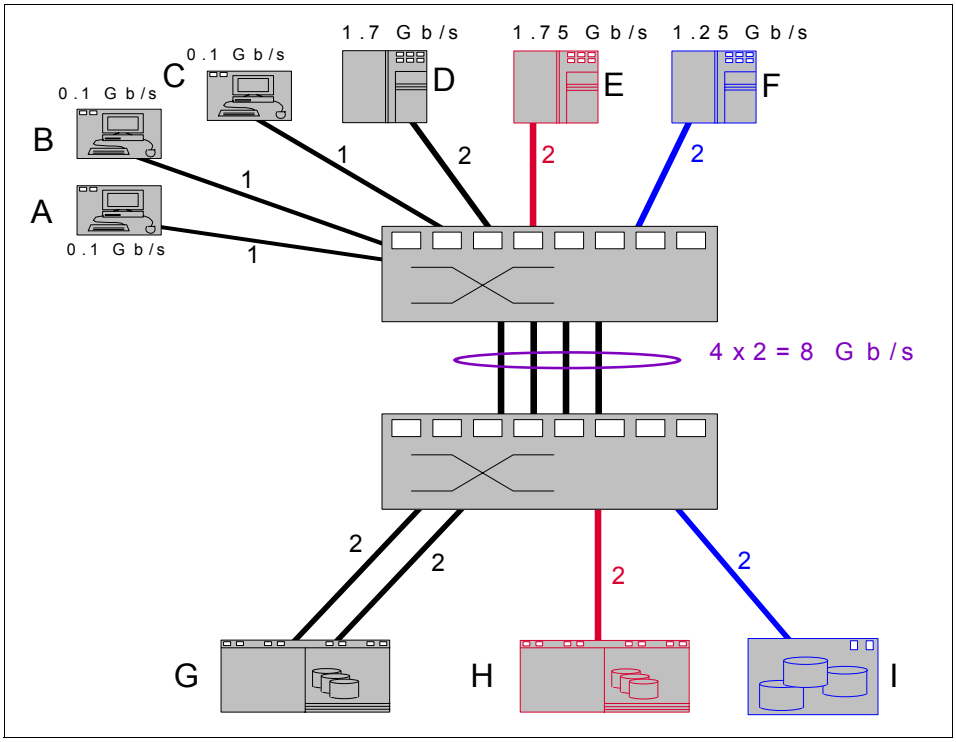


Figure 3-36 Trunking

In this example we have six computers which are accessing three storage devices. Computers A, B, C and D are communicating with Storage G. Server E is communicating with Storage H and Server F uses disks in storage device I.

The speeds of the links are shown in Gb/s and the target throughput for each computer is shown. If we let FSPF alone decide the routing for us, we could have a situation where servers D and E were both utilizing the same ISL. This would lead to oversubscription and hence congestion as 1.7 added to 1.75 is greater than 2.

If all of the ISLs are gathered together into a trunk, then effectively they can be seen as a single, big ISL. In effect, they appear to be an 8 Gb/s ISL. This bandwidth is greater than the total requirement of all of the servers. In fact the nodes require an aggregate bandwidth of 5 Gb/s, so we could even suffer a failure of one of the ISLs and still have enough bandwidth to satisfy their needs.

When the nodes come up, FSPF will simply see one route and they will all be assigned a route over the same trunk. The fabric operating systems in the switches will share the load over the actual ISLs which combine to make up the trunk. This is done by distributing frames over the physical links, and then re-assembling them at the destination switch so that in-order delivery can be assured, if necessary. To FSPF, a trunk will appear as a single, low-cost ISL.

3.9.1 Frame filtering

Zoning is a fabric management service that can be used to create logical subsets of devices within a SAN and enable partitioning of resources for management and access control purposes. Frame filtering is another feature that enables devices to provide zoning functions with finer granularity. Frame filtering can be used to set up port level zoning, world wide name zoning, device level zoning, protocol level zoning, and LUN level zoning. Frame filtering is commonly carried out by an ASIC. This has the result that, after the filter is set up, the complicated function of zoning and filtering can be achieved at wire speed.

3.9.2 Oversubscription

There can be several ports in a switch that can communicate with one particular port, for example, several servers sharing a path to a storage device. In this case the storage path determines the maximum data rate that all servers can get, and this is usually given by the device and not the SAN itself.

When we start cascading switches, communications between switches are carried by ISLs. It is possible that several ports in one switch need to simultaneously communicate with ports in the other switch through a single ISL. In this case it is possible that the connected devices could sustain a combined data transfer rate higher than the ISL can provide, so the throughput will be limited to what the ISL can handle and this may impose a throttle or bottleneck within the fabric.

We use the term oversubscribing to describe the occasion when we have several ports trying to communicate with each other, and when the total throughput is higher than what that port can provide.

This can happen on storage ports and ISLs. When designing a SAN it is important to consider the possible traffic patterns to determine the possibility of oversubscription and which may result in degraded performance. For example, traffic patterns during backup periods may introduce oversubscription that can affect performance on production systems. In some cases this is not a problem that may even be noticed at first, but as the SAN fabric grows, it is important not to ignore this possibility.

Oversubscription of an ISL may be overcome by adding a parallel ISL. Oversubscription to a storage device may be overcome by adding another adapter to the storage array and connecting into the fabric. There are other considerations as well, such as:

- More ports will be used for ISLs and less ports will be available for nodes.
- The cost of retrofitting additional ISLs may be significant if the sites are remote.

3.9.3 Congestion

When oversubscription occurs, it leads to a condition called congestion. When a node is unable to utilize as much bandwidth as it would like to, due to contention with another node, then there is congestion. A port, link, or fabric can be congested.

3.9.4 Information units

A Fibre Channel Information Unit (IU), is defined as:

*"*A related set of data specified by a Fibre Channel upper layer protocol, which is transferred as a single Fibre Channel sequence."

Upper Layer Protocols are discussed in "Upper layer protocol mapping (ULP): FC-4" on page 104. The Fibre Channel sequence is described in 3.10, "Ordered set, frames, sequences, and exchanges" on page 119.

3.9.5 The movement of data

To move data bits with integrity over a physical medium, there must be a mechanism to check that this has happened and integrity has not been compromised. This is provided by a reference clock which ensures that each bit is received as it was transmitted. In parallel topologies this can be accomplished by using a separate clock or strobe line. As data bits are transmitted in parallel from the source, the strobe line alternates between high or low to signal the receiving end that a full byte has been sent. In the case of 16 and 32-bit wide parallel cable, it would indicate that multiple bytes have been sent.

The reflective differences in fiber optic cabling mean that modal dispersion may occur. This may result in frames arriving at different times. This bit error rate (BER) is referred to as the jitter budget. No products are entirely jitter free, and this is an important consideration when selecting the components of a SAN.

As serial data transports only have two leads, transmit and receive, clocking is not possible using a separate line. Serial data must carry the reference timing which means that clocking is embedded in the bit stream.

Embedded clocking, though, can be accomplished by different means. Fibre Channel uses a byte-encoding scheme, which is covered in more detail in 3.9.6, “Data encoding” on page 116, and clock and data recovery (CDR) logic to recover the clock. From this, it determines the data bits that comprise bytes and words.

Gigabit speeds mean that maintaining valid signaling, and ultimately valid data recovery, is essential for data integrity. Fibre Channel standards allow for a single bit error to occur only once in a million, million bits (1 in 10^{12}). In the real IT world, this equates to a maximum of one bit error every 16 minutes, however actual occurrence is a lot less frequent than this.

3.9.6 Data encoding

In order to transfer data over a high-speed serial interface, the data is encoded prior to transmission and decoded upon reception. The encoding process ensures that sufficient clock information is present in the serial data stream to allow the receiver to synchronize to the embedded clock information and successfully recover the data at the required error rate. This 8b/10b encoding will find errors that a parity check cannot. A parity check will not find even numbers of bit errors, only odd numbers. The 8b/10b encoding logic will find almost all errors.

First developed by IBM, the 8b/10b encoding process will convert each 8-bit byte into two possible 10-bit characters.

This scheme is called 8b/10b encoding, because it refers to the number of data bits input to the encoder and the number of bits output from the encoder.

The format of the 8b/10b character is of the format Ann.m, where:

- A represents ‘D’ for data or ‘K’ for a special character
- nn is the decimal value of the lower 5 bits (EDCBA)
- ‘.’ is a period
- m is the decimal value of the upper 3 bits (HGF)

We illustrate an encoding example in Figure 3-37.

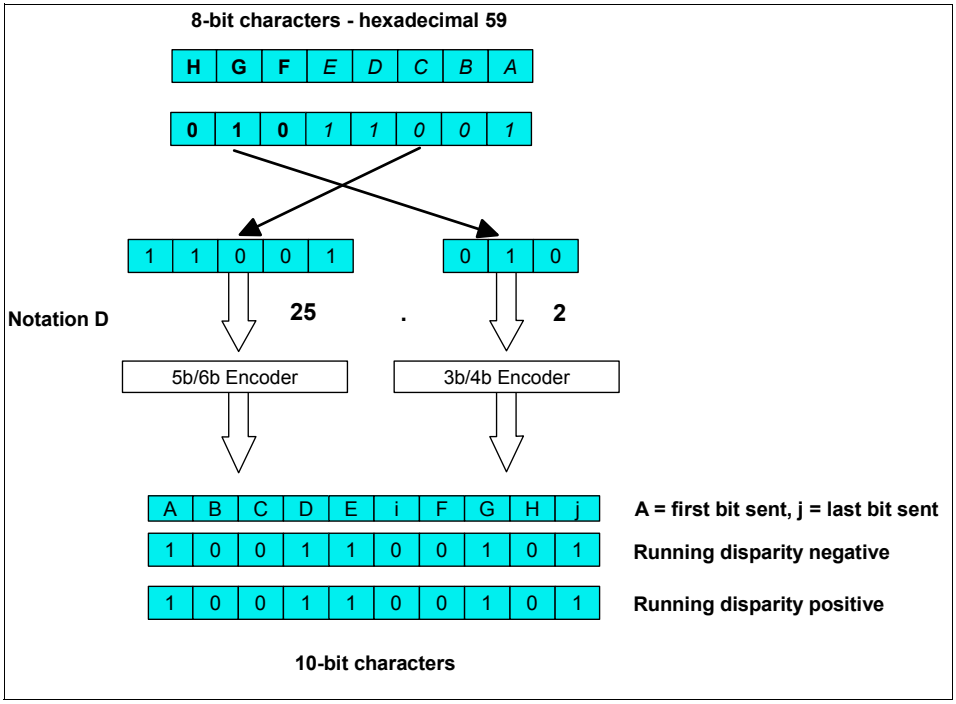


Figure 3-37 8b/10b encoding logic

In the encoding example the following occurs:

1. Hexadecimal representation x'59' is converted to binary: 01011001
2. Upper three bits are separated from the lower 5 bits: 010 11001
3. The order is reversed and each group is converted to decimal: 25 2
4. Letter notation D (for data) is assigned and becomes: D25.2

Running disparity

As we illustrate, the conversion of the 8-bit data bytes has resulted in two 10-bit results. The encoder needs to choose one of these results to use. This is achieved by monitoring the running disparity of the previously processed character. For example, if the previous character had a positive disparity, then the next character issued should have an encoded value that represents negative disparity.

You will notice that in our example the encoded value, when the running disparity is either positive or negative, is the same. This is legitimate. In some cases it (the encoded value) will differ, and in others it will be the same.

It should be noticed that in the above example the encoded 10 bit byte has 5 bits which are set and 5 bits which are unset. The only possible results of the 8b/10b encoding are as follows:

- If 5 bits are set, then:
  - The byte is said to have neutral disparity
- If 4 bits are set and 6 are unset, then:
  - The byte is said to have negative disparity
- If 6 bits are set and four are unset then:
  - The byte is said to have positive disparity

The rules of Fibre Channel define that a byte which is sent cannot take the positive or negative disparity above one unit. Thus, if the current running disparity is negative, then the next byte that is sent must either have

- **Neutral disparity**
  - Keeping the current running disparity negative
  - The subsequent byte would need to have either neutral or positive disparity.
- **Positive disparity**
  - Making the new current running disparity neutral
  - The subsequent byte could have either positive, negative or neutral disparity.

Note: By this means, at any point in time, at the end of any byte, the number of set bits and unset bits that have passed over a Fibre Channel link will only differ by a maximum of two.

K28.5

As well as the fact that many 8 bit numbers encode to *two* 10 bit numbers under the 8b/10b encoding scheme, there are some other key features.

Some 10 bit numbers cannot be generated from any 8 bit number. Thus, it should not be possible to see these particular 10 bit numbers as part of a flow of data. his is really a useful fact, as it means that these particular 10 bit numbers can be used by the protocol for signalling or control purposes.

These characters are referred to as Comma characters and rather than having the prefix D have the prefix K.

The only one that actually gets used in Fibre Channel is the character known as K28.5 and it has a very special property.

The two 10-bit encodings of K28.5 are shown in Table 3-3 below.

Table 3-3 10 bit encodings of K28.5

Name of Character	Encoding for current Running Disparity of	
	Negative	Positive
K28.5	001111 1010	110000 0101

It was stated above that all of the 10 bit bytes which are possible using the 8b/10b encoding scheme have either four, five or six bits set. The K28.5 character is special in that it is the only character used in Fibre Channel that has five consecutive bits set or unset, all other characters have four or less consecutive bits of the same setting.

So, what is the significance? There are two things to note here:

The first is that these ones and zeroes are actually representing light and dark on the fibre (assuming fiber optic medium). A 010 pattern would effectively be a light pulse between two periods of darkness. A 0110 would be the same, except that the pulse of light would last for twice the length of time.

As the two devices have their own clocking circuitry, the number of consecutive set bits, or consecutive unset bits, becomes important. Let us say that Device 1 is sending to Device 2 and that the clock on Device 2 is running 10% faster than that on Device 1. If Device 1 sent 20 clock cycles worth of set bits, then Device 2 would count 22 set bits. (Note that this example is just given to illustrate the point). The worst possible case that we can have in Fibre Channel is five consecutive bits of the same setting within one byte: the K28.5

The other key thing is that because this is the *only* character with five consecutive bits of the same setting, Fibre Channel hardware can look out for it specifically. As K28.5 is used for control purposes, this is very useful and allows the hardware to be designed for maximum efficiency.

3.10 Ordered set, frames, sequences, and exchanges

In order for Fibre Channel devices to be able to communicate with each other, there need to be some strict definitions regarding the way that data is sent and received. To this end, some data structures have been defined.

3.10.1 Ordered set

Fibre Channel uses a command syntax, known as an ordered set, to move the data across the network. The ordered sets are four byte transmission words containing data and special characters which have a special meaning. Ordered sets provide the availability to obtain bit and word synchronization, which also establishes word boundary alignment. An ordered set always begins with the special character K28.5. Three major types of ordered sets are defined by the signaling protocol.

The frame delimiters, the Start Of Frame (SOF) and End Of Frame (EOF) ordered sets, establish the boundaries of a frame. They immediately precede or follow the contents of a frame. There are 11 types of SOF and eight types of EOF delimiters defined for the fabric and N_Port Sequence control.

The two primitive signals: idle and receiver ready (R_RDY) are ordered sets designated by the standard to have a special meaning. An Idle is a primitive signal transmitted on the link to indicate an operational port facility ready for frame transmission and reception. The R_RDY primitive signal indicates that the interface buffer is available for receiving further frames.

A primitive sequence is an ordered set that is transmitted and repeated continuously to indicate specific conditions within a port or conditions encountered by the receiver logic of a port. When a primitive sequence is received and recognized, a corresponding primitive sequence or Idle is transmitted in response. Recognition of a primitive sequence requires consecutive detection of three instances of the same ordered set. The primitive sequences supported by the standard are:

- Offline state (OLS)
- Not operational (NOS)
- Link reset (LR)
- Link reset response (LRR)

Offline (OLS): The offline primitive sequence is transmitted by a port to indicate one of the following conditions: The port is beginning the link initialization protocol, or the port has received and recognized the NOS protocol or the port is entering the offline status.

Not operational (NOS): The not operational primitive sequence is transmitted by a port in a point-to-point or fabric environment to indicate that the transmitting port has detected a link failure or is in an offline condition, waiting for the OLS sequence to be received.

Link reset (LR): The link reset primitive sequence is used to initiate a link reset.

Link reset response (LRR): Link reset response is transmitted by a port to indicate that it has recognized a LR sequence and performed the appropriate link reset.

Data transfer

In order to send data over Fibre Channel, though, we need more than just the control mechanisms. Data is sent in frames. One or more related frames make up a Sequence and one or more related sequences make up an exchange.

3.10.2 Frames

Fibre Channel places a restriction on the length of the data field of a frame at 528 transmission words which is 2112 bytes. (See Table 3-4, “Transmission words in a frame” on page 122.) Larger amounts of data must be transmitted in several frames. This larger unit that consists of multiple frames is called a sequence. An entire transaction between two ports is made up of sequences administered by an even larger unit called an exchange.

Framing rules

The following rules apply to the framing protocol:

- A frame is the smallest unit of information transfer.
- A sequence has at least one frame.
- An exchange has at least one sequence.

3.10.3 Sequences

The information in a sequence moves in one direction, from a source N_Port to a destination N_Port. Various fields in the frame header are used to identify the beginning, middle and end of a sequence, while other fields in the frame header are used to identify the order of frames, in case they arrive out of order at the destination.

3.10.4 Exchanges

Two other fields of the frame header identifies the exchange ID. An exchange is responsible for managing a single operation that may span several sequences, possibly in opposite directions. The source and destination can have multiple exchanges active at a time

Using SCSI as an example, a SCSI task is an exchange. The SCSI task is made up of one or more information units. The information units (IU) would be:

- Command IU
- Transfer ready IU

- Data IU
- Response IU

Each IU is one sequence of the exchange. Only one participant sends a sequence at a time.

3.10.5 Frames

A frame consists of the following elements:

- SOF delimiter
- Frame header
- Optional headers and payload (data field)
- CRC field
- EOF delimiter

Transmission word

A transmission word is the smallest transmission unit defined in Fibre Channel. This unit consists of four transmission characters, 4 x 10 or 40 bits. When information transferred is not an even multiple of four bytes the framing protocol adds fill bytes. The fill bytes are stripped at the destination.

Frame

Frames are the building block of Fibre Channel. A frame is a string of transmission words, prefixed by a Start Of Frame (SOF) delimiter and followed by an End Of Frame (EOF) delimiter. The way that Transmission Words make up a frame is shown in Table 3-4.

Table 3-4 Transmission words in a frame

SOF	Frame Header	Data Payload Transmission Words	CRC	EOF
1 TW	6 TW	0-528 TW	1 TW	1 TW

Frame header

Each frame includes a header that identifies the source and destination of the frame as well as control information that manages the frame as well as sequences and exchanges associated with that frame. The structure of the Frame header is shown below in Table 3-5. The abbreviations are explained below the table.

Table 3-5 The frame header

	Byte 0	Byte 1	Byte 2	Byte 3
Word 0	R_CTL	Destination_ID (D_ID)		

	Byte 0	Byte 1	Byte 2	Byte 3
Word 1	Reserved	Source_ID (S_ID)		
Word 2	Type	Frame Control (F_CTL)		
Word 3	SEQ_ID	DF_CTL	SequenceCount (SEQ_CNT)	
Word 4	Originator X_ID (OX_ID)		Responder X_ID (RX_ID)	
Word 5	Parameter			

Routing control (R_CTL)
This field identifies the type of information contained in the payload and where in the destination node it should be routed.

Destination ID
This field contains the address of the frame destination and is referred to as the D_ID.

Source ID
This field contains the address of where the frame is coming from and is referred to as the S_ID.

Type
Type identifies the protocol of the frame content for data frames, (i.e. SCSI) or a reason code for control frames.

F_CTL
This field contains control information that relates to the frame content.

SEQ_ID
The sequence ID is assigned by the sequence initiator and is unique for a specific D_ID and S_ID pair while the sequence is open.

DF_CTL
Data Field control specifies whether there are optional headers present at the beginning of the data field.

SEQ_CNT
This count identifies the position of a frame within a sequence and is incremented by one for each subsequent frame transferred in the sequence.

OX_ID
This field identifies the exchange ID assigned by the originator.

RX_ID
This field identifies the exchange ID to the responder.

Parameter
Parameter specifies relative offset for data frames or information specific to link control frames.

3.10.6 “In order” and “out of order”

When data is transmitted over Fibre Channel, it is sent in frames. These frames only carry a maximum of 2112 bytes of data and that is often not enough to hold the entire set of information to be communicated. In this case, more than one frame is needed. Some classes of Fibre Channel communication guarantee that the frames will arrive at the destination in the same order that they were transmitted. Other classes do not. If the frames do arrive in the same order that they were sent, then we are said to have in order delivery of frames.

In some cases, it is critical that the frames arrive in the correct order, and in others, it is not so important. In the latter case, the receiving port can reassemble the frames into the correct order before passing the data out to the application. It is, however, quite common for switches and directors to guarantee in-order delivery, even if the particular class of communication allows for the frames to be delivered out of sequence.

3.10.7 Latency

The term latency relates to the delay between an action being requested and it actually happening. So, for example, if I am sitting at my chair and want to turn the light on, the latency might be:

- Low: If there is a light on the desk
- Medium: If there is a light switch by the door in the room
- High: If I need to leave the room and find the switch at the end of the corridor

In the realms of science, we like to quantify things, though, so we usually measure latency in terms of time.

Latency occurs almost everywhere. It is simply a fact that in the normal world it takes time and energy to do stuff. The areas where we particularly need to be aware of latency in a SAN are

- Ports
- Hubs, switches, directors
- Long distance links
- Inter Switch links
- ASICs

- Interblade links in a core switch or director

3.10.8 Heterogeneousness

This term refers to whether or not more than one different type of thing is involved in something. So, in the realms of a SAN:

- If a SAN only deals with one type of operating system platform, then it is non-heterogeneous with regard to the servers. If, on the other hand, it deals with more than one type of server then it can be called heterogeneous.
- Similarly, a SAN can be described as heterogeneous or non heterogeneous wit regards to storage, or even SAN components such as switching devices.
- If management software will only manage a single type of equipment, then it is termed non-heterogeneous. The term is often used as well to describe software which will only manage equipment from one manufacturer. Software which will manage equipment from many manufacturers is called heterogeneous.

Generally speaking, the trend is towards having non-heterogeneous SANs. If a heterogeneous SAN is being implemented, perhaps because the situation only calls for a simple SAN architecture, there is a strong incentive to design it in such a way, and using such equipment, that it is not restricted to being that way, in the future.

3.10.9 Open Fiber Control: OFC or Non-OFC

When dealing with lasers there is potential danger to the eyes. Generally, the lasers in use in Fiber Channel are low powered devices designed for quality of light and signalling rather than for maximum power. They can still be dangerous though.

Attention: Never look into a laser light source.

Never look into the end of an optic cable unless you know exactly where the other end is and you also know that nobody could connect a light source to it.

In order to add a degree of safety, the concept of Open Fiber Control (OFC) was developed.

The idea is as follows:

1. A device is powered on and it sends out low powered light.
2. If it does not receive light back, then it assumes that there is no fiber connected. This is a fail-safe option.

3. When it receives light, it assumes that there is a fiber connected and switches the laser to full power.
4. If one of the devices stops receiving light, then it will revert to the low power mode.

When a device is transmitting at low power, it is not able to send data, it is just waiting for a completed optical loop.

The OFC ensures that the laser does not emit light which would exceed the Class1 laser limit when no fiber is connected. Non-OFC devices are guaranteed to be below Class 1 limits at all times.

The key factor is that the devices at each end of a fiber link must either both be OFC or both be Non-OFC.

All modern equipment uses Non-OFC optics, but it is possible that some legacy equipment may be using OFC optics.

3.11 Fibre Channel Arbitrated Loop (FC-AL)

Fibre Channel Arbitrated Loop is sufficiently different from Fibre Channel in a crosspoint switch environment that we are covering some of the specific differences in this section.

3.11.1 Loop protocols

To support the shared behavior of the Arbitrated Loop, a number of loop-specific protocols are used. These protocols are used to:

- Initialize the loop and assign addresses.
- Arbitrate for access to the loop.
- Open a loop circuit with another port in the loop.
- Close a loop circuit when two ports have completed their current use of the loop.

Implement the access fairness mechanism to ensure that each port has an opportunity to access the loop.

Loop initialization and LIP

Loop initialization is a necessary process for the introduction of new participants on to the loop. Whenever a loop port is powered on or initialized, it executes the loop initialization primitive (LIP) to perform loop initialization. Optionally, loop initialization may build a positional map of all the ports on the loop. The positional map provides a count of the number of ports on the loop, their addresses, and their position relative to the loop initialization master.

Following loop initialization, the loop enters a stable monitoring mode and begins with normal activity. An entire loop initialization sequence may take only a few milliseconds, depending on the number of NL_Ports attached to the loop. Loop initialization may be started by a number of causes. One of the most likely reasons for loop initialization is the introduction of a new device. For instance, an active device may be moved from one hub port to another hub port, or a device that has been powered on could re-enter the loop.

A variety of ordered sets have been defined to take into account the conditions that an NL_Port may sense as it starts the initialization process. These ordered sets are sent continuously while a particular condition or state exists. As part of the initialization process, loop initialization primitive sequences (referred to collectively as LIPs) are issued. As an example, an NL_Port must issue at least three identical ordered sets to start initialization. An ordered set transmission word always begins with the special character K28.5.

Once these identical ordered sets have been sent, and as each downstream device receives the LIP stream, devices enter a state known as open-init. This causes the suspension of any current operation and enables the device for the loop initialization procedure. LIPs are forwarded around the loop until all NL_Ports are in an open-init condition.

At this point, the NL_Ports need to be managed. In contrast to a token ring, the Arbitrated Loop has no permanent master to manage the topology.

Therefore, loop initialization provides a selection process to determine which device will be the temporary loop master. After the loop master is chosen, it assumes the responsibility for directing or managing the rest of the initialization procedure. The loop master also has the responsibility for closing the loop and returning it to normal operation.

Selecting the loop master is carried out by a subroutine known as the Loop Initialization Select Master (LISM) procedure. A loop device can be considered for temporary master by continuously issuing LISM frames that contain a port type identifier and a 64-bit World-Wide Name. For FL_Ports the identifier is x'00' and for NL_Ports it is x'EF'.

When a downstream port receives a LISM frame from a upstream partner, the device will check the port type identifier. If the identifier indicates an NL_Port, the downstream device will compare the WWN in the LISM frame to its own. The WWN with the lowest numeric value has priority. If the received frame's WWN indicates a higher priority, that is to say it has a lower numeric value, the device stops its LISM broadcast and starts transmitting the received LISM. Had the received frame been of a lower priority, the receiver would have thrown it away and continued broadcasting its own.

At some stage in proceedings, a node will receive its own LISM frame, which indicates that it has the highest priority, and succession to the throne of *temporary loop master* has taken place. This node will then issue a special ordered set to indicate to the others that a temporary master has been selected.

Hub cascading

Since an Arbitrated Loop hub supplies a limited number of ports, building larger loops may require linking another hub. This is called hub cascading. A server with an FC-AL, shortwave, host bus adapter can connect to an FC-AL hub 500 meters away. Each port on the hub can connect to an FC-AL device up to 500 meters away. Cascaded hubs use one port on each hub for the hub-to-hub connection and this increases the potential distance between nodes in the loop by an additional 500 meters. In this topology, the overall distance is 1500 meters. Both hubs can support other FC-AL devices at their physical locations. Stated distances assume a 50 micron multimode cable.

Loops

There are two different kinds of loops, the private and the public loop.

Private loop

The private loop does not connect with a fabric, only to other private nodes using attachment points called NL_Ports. A private loop is enclosed and known only to itself. In Figure 3-38 we show a private loop.

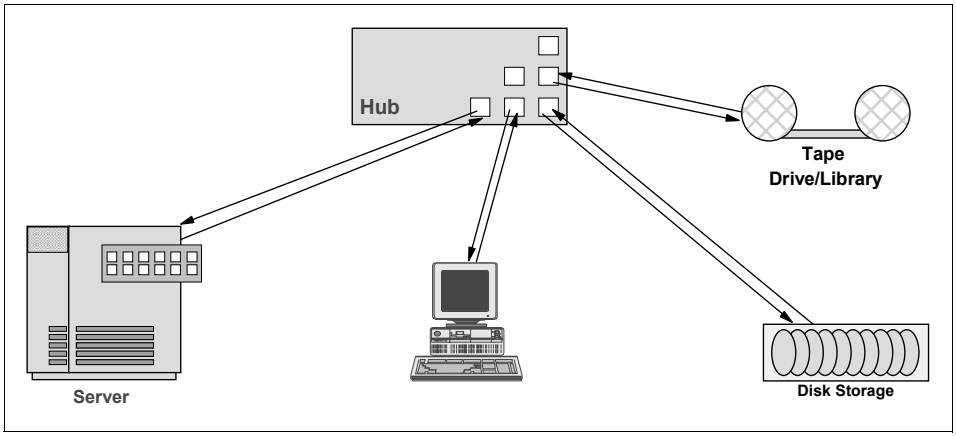


Figure 3-38 Private loop implementation

Public loop

A public loop requires a fabric and has at least one FL_Port connection to a fabric. A public loop extends the reach of the loop topology by attaching the loop to a fabric. Figure 3-39 shows a public loop.

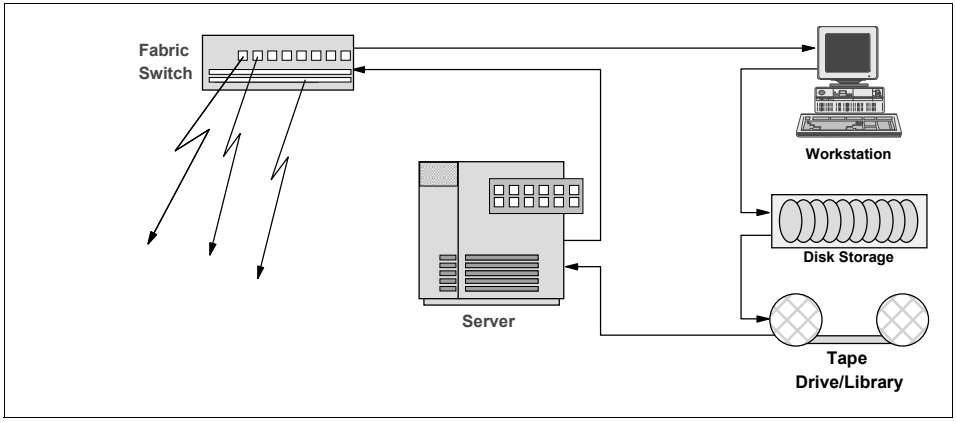


Figure 3-39 Public loop implementation

Arbitration

When a loop port wants to gain access to the loop, it has to arbitrate. When the port wins arbitration, it can open a loop circuit with another port on the loop; a function similar to selecting a device on a bus interface. Once the loop circuit has been opened, the two ports can send and receive frames between each other. This is known as *loop tenancy*.

If more than one node on the loop are arbitrating at the same time, the node with the lower Arbitrated Loop Physical Address (AL_PA) gains control of the loop. Upon gaining control of the loop, the node then establishes a point-to-point transmission with another node using the full bandwidth of the media. When a node has finished transmitting its data, it is not required to give up control of the loop. This is a channel characteristic of Fibre Channel. However, there is a *fairness algorithm*, which states that a device cannot regain control of the loop until the other nodes have had a chance to control the loop.

3.11.2 Fairness algorithm

The way that the fairness algorithm works is based around the IDLE ordered set and the way that arbitration is carried out. In order to determine that the loop is not in use, an NL_Port waits until it sees an IDLE go by and it can arbitrate for the loop by sending an ARB Primitive Signal ordered set. If a higher priority device arbitrates before the first NL_Port sees its own ARB come by, then it loses the arbitration, but if it sees that its own ARB has gone all the way round the loop, then it has won arbitration. It can then open a communication to another NL_Port. When it has finished, it can close the connection and either rearbitrate for the loop or send one or more IDLEs. If it complies with the fairness algorithm (sometimes this is a configurable parameter) then it will take the option of sending IDLEs. That will allow lower priority NL_Ports to successfully arbitrate for the loop. There is no rule that forces any device to operate the fairness algorithm.

3.11.3 Loop addressing

An NL_Port, like a N_Port, has a 24-bit port address. If no switch connection exists, the two upper bytes of this port address are zeroes (x'00 00') and referred to as a private loop. The devices on the loop have no connection with the outside world. If the loop is attached to a fabric and an NL_Port supports a fabric login, the upper two bytes are assigned a positive value by the switch. We call this mode a public loop.

As fabric-capable NL_Ports are members of both a local loop and a greater fabric community, a 24-bit address is needed as an identifier in the network. In the case of public loop assignment, the value of the upper two bytes represents the loop identifier, and this will be common to all NL_Ports on the same loop that performed login to the fabric.

In both public and private Arbitrated Loops, the last byte of the 24-bit port address refers to the Arbitrated Loop physical address (AL_PA). The AL_PA is acquired during initialization of the loop and may, in the case of fabric-capable loop devices, be modified by the switch during login.

The total number of the AL_PAs available for Arbitrated Loop addressing is 127, which is based on the requirements of 8b/10b running disparity between frames.

As a frame terminates with an end-of-frame character (EOF) this will force the current running disparity negative. In the Fibre Channel standard each transmission word between the end of one frame and the beginning of another frame should also leave the running disparity negative. If all 256 possible 8-bit bytes are sent to the 8b/10b encoder, 134 emerge with neutral disparity characters. Of these 134, seven are reserved for use by Fibre Channel. The 127 neutral disparity characters left have been assigned as AL_PAs. Put another way, the 127 AL_PA limit is simply the maximum number, minus reserved values, of neutral disparity addresses that can be assigned for use by the loop. This does not imply that we recommend this amount, or load, for a 100 MB/s shared transport, but only that it is possible.

Arbitrated Loop will assign priority to AL_PAs, based on numeric value. The lower the numeric value, the higher the priority is. For example, an AL_PA of x'01' has a much better position to gain arbitration over devices that have a lower priority or higher numeric value. At the top of the hierarchy it is not unusual to find servers, but at the lower end you would expect to find disk arrays.

It is the Arbitrated Loop initialization that ensures each attached device is assigned a unique AL_PA. The possibility for address conflicts only arises when two separated loops are joined together without initialization.

3.11.4 Private devices on NL_Ports

It is easy to explain how the port to World Wide Name address resolution works when a single device from an N_Port is connected to an F_Port, or when a public NL_Port device is connected to FL_Port in the switch. The SNS will add an entry for the device World Wide Name and connects that with the port address which is selected from the selection of free port addresses for that switch. Problems may arise when a private Fibre Channel device is attached to the switch. Private Fibre Channel devices were designed to only to work in private loops.

When the Arbitrated Loop is connected to the FL_Port, this port obtains the highest priority address in the loop to which it is attached (0x00). Then the FL_Port performs a LIP. After this process is completed, the FL_Port registers all devices on the loop with the SNS. Devices on the Arbitrated Loop use only 8-bit addressing, but in the switched fabric, 24-bit addressing is used. When the FL_Port registers the devices on the loop to the SNS, it adds two most significant bytes to the existing 8-bit address.

The format of the address in the SNS table is 0xPPPPLL, where the PPPP is the two most significant bytes of the FL_Port address and the LL is the device ID on the Arbitrated Loop which is connected to this FL_Port. Modifying the private loop address in this fashion, all private devices can now talk to all public devices, and all public devices can talk to all private devices.

Because we have stated that private devices can only talk to devices with private addresses, some form of translation must take place. We show an example of this in Figure 3-40.

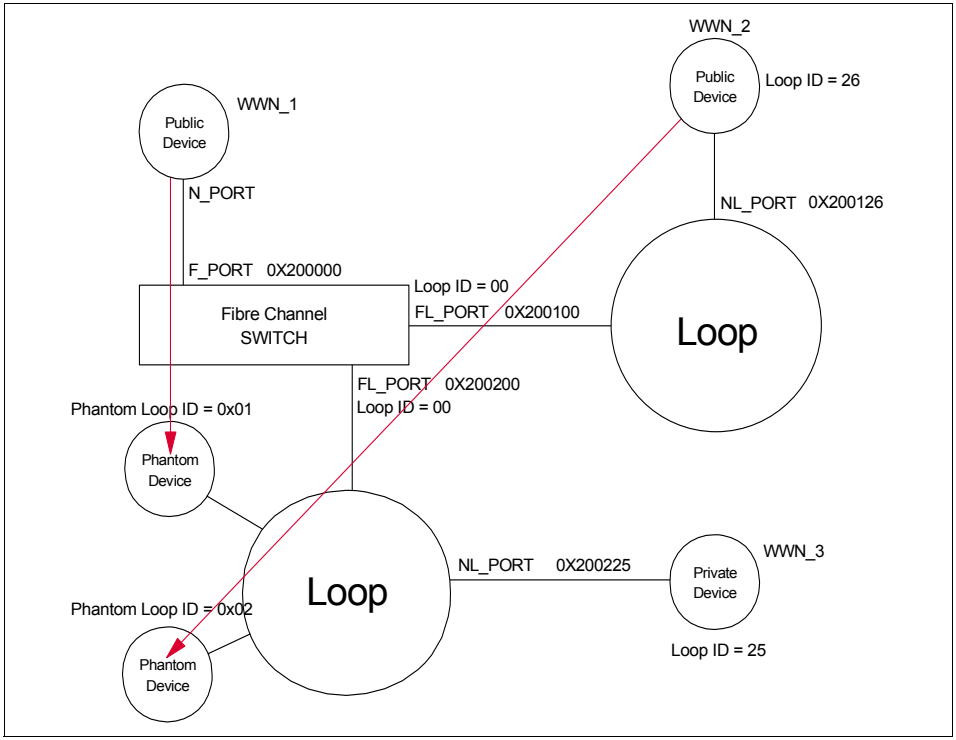


Figure 3-40 Arbitrated loop address translation

As you can see, we have three devices connected to the switch:

- Public device N_Port with WWN address WWN_1 on F_Port with the port address 0x200000
- Public device NL_Port with WWN address WWN_2 on FL_Port with the port address 0x200100. The device has AL_PA 0x26 on the loop, which is attached on the FL_Port

- Private device NL_Port with WWN address WWN_3 on FL_Port with the port address 0x200200. The device has AL_PA 0x25 on the loop, which is attached to the FL_Port

After all FLOGI and PLOGI functions are performed the SNS will have the entries shown in Table 3-6.

Table 3-6 Name server entries

24 bit port address	WWN	FL_Port address
0x200000	WWN_1	n/a
0x200126	WWN_2	0x200100
0x200225	WWN_3	0x200200

We now explain some possible scenarios.

Public N_Port device accesses private NL_Port device

The communication from device to device starts with PLOGI to establish a session. When a public N_Port device wants to perform a PLOGI to a private NL_Port device, the FL_Port on which this private device exists will assign a “phantom” private address to the public device. This phantom address is known only inside this loop, and the switch keeps track of the assignments.

In our example, when the WWN_1 device wants to talk to the WWN_3 device, the following, shown in Table 3-7, is created in the switch.

Table 3-7 Phantom addresses

Switch port address	Phantom Loop Port ID
0x200000	0x01
0x200126	0x02

When the WWN_1 device enters into the loop it represents itself with AL_PA ID 0x01 (its phantom address). All private devices on that loop use this ID to talk to that public device. The switch itself acts as a proxy, and translates addresses in both directions.

Usually the number of phantom addresses is limited, and this number of phantom addresses decreases the number of devices allowed in the Arbitrated Loop. For example, if the number of phantom addresses is 32, this limits the number of physical devices in the loop to 126 - 32 = 94.

Public N_Port device accesses public NL_Port device

If an N_Port public device wants to access an NL_Port public device, it simply performs a PLOGI with the whole 24-bit address.

Private NL_Port device accesses public N_Port or NL_Port

When a private device needs to access a remote public device, it uses the public device's phantom address. When the FL_Port detects the use of a phantom AL_PA ID, it translates that to a switch port ID using its translation table similar to that shown in Table 3-7.

QuickLoop

As we have already explained above, private devices can cooperate in the fabric using translative mode. However, if you have a private host (server), this is not possible. To solve this, switch vendors (including IBM) support a QuickLoop feature. The QuickLoop feature allows the whole switch or just a set of ports to operate as an Arbitrated Loop. In this mode, devices connected to the switch do not perform a fabric login, and the switch itself will emulate the loop for those devices. All public devices can still see all private devices on the QuickLoop in the translative mode. This is described in 3.11.4, "Private devices on NL_Ports" on page 131.

3.12 Factors and considerations

There are other factors that need to be taken into account when contemplating building a SAN from the components we have described.

3.12.1 Limits

There are various limits encountered in Fibre Channel.

Distances

In Table 3-8 we show the copper and Fibre Channel limits.

Table 3-8 Copper and Fibre Channel limits

Type	Fiber type	Distance	Speed**
Extended LW Laser	Single Mode 9 µm	80-100Km*	100MB/s 200MB/s 400MB/s 1000MB

Type	Fiber type	Distance	Speed**
LW Laser	Single Mode 9 µm	10Km	100MB/s 200MB/s 400MB/s 1000MB/s
SW Laser	Multi Mode 50 µm	000500m 000300m 000150m	100MB/s 200MB/s 400MB/s
SW Laser	Multi Mode 62.5 µm	000300m 000150m 000055m	100MB/s 200MB/s 400MB/s
Electrical	75Ω Video Coax	25 meters	100 MB/s
Electrical	75Ω Mini Coax	10 meters	100 MB/s
Electrical	150Ω Shielded Twisted Pair	50 meters	25 MB/s

*The certified and supported length will depend upon the vendor.

**Where multiple speeds are specified, the actual speed will depend upon the SFP, GBIC or other transducer selected. The speeds are the nominal throughput of application data and the considerations discussed in 3.7.8, “FC-PH-2 and speed” on page 96.

3.12.2 Security

A major consideration when setting up a SAN is the security of the data and servers. There is security in terms of:

- Data integrity
- Ensuring that stored data are only accessed by servers authorized to do so

The data integrity is covered by the usual methods of mirroring, and other RAID levels, the use of copying software such as PPRC, and of course backups.

In order to ensure that only the correct servers access the correct data, certain steps can be taken.

Hardware zoning

Hardware zoning can be implemented to ensure that only devices connected to particular ports are logically connected (see 3.8.1, “Hardware zoning” on page 106).

LUN masking

One approach to securing storage devices from hosts wishing to take over already assigned resources is logical unit number (LUN) masking. The user defines which hosts can access which LUN by means of the storage device control program. Whenever the host accesses a particular LUN, the storage device will check its access list for that LUN, and it will allow or disallow access to the LUN.

3.12.3 Interoperability

This is the matter of whether or not different devices will operate with each other. There are several factors to consider which may not be immediately obvious. It is important to consider all aspects of interoperability to ensure that the SAN will provide the required functionality and will be supported by all necessary parties.

Areas to consider will include:

- Whether a particular HBA will operate correctly with a particular storage device
- Whether all SAN devices, directors, switches, hubs, bridges and so on, will all operate with each other
  - Whether or not they are made by the same manufacturer.
  - Will the same pieces of equipment interoperate if they come from different sources? For example they may be the same hardware but there may be specific firmware versions.
  - If a server has a requirement for a particular version of firmware on its HBA and a server has a requirement for a particular version of firmware on its adapter and they are connected to the same switch, will the switch run a version of firmware which is compatible with both?
  - E_Port interoperability

3.13 Standards

Given the strong drive towards SANs from users and vendors alike, one of the most critical success factors is the ability of systems and software from different vendors to operate together in a seamless way. We have mentioned the various fibre channel standards previously. This chapter will document where each can be found. Standards are the basis for the interoperability of devices and software from different vendors.

A good benchmark is the level of standardization in today's LAN and WAN networks. Standard interfaces for interoperability and management have been developed, and many vendors compete with products based on the implementation of these standards. Customers are free to mix and match components from multiple vendors to form a LAN or WAN solution. They are also free to choose from several different network management software vendors to manage their heterogeneous network.

The major vendors in the SAN industry recognize the need for standards, especially in the areas of interoperability interfaces and application programming interfaces (APIs), as these are the basis for wide acceptance of SANs. Standards will allow customers a greater breadth of choice, and will lead to the deployment of cross-platform, multi-vendor, enterprise-wide SAN solutions.

3.13.1 SAN industry associations and organizations

A number of industry associations, standards bodies and company groupings are involved in developing, and publishing SAN standards. The major groups linked with SAN standards are shown in Figure 3-41.

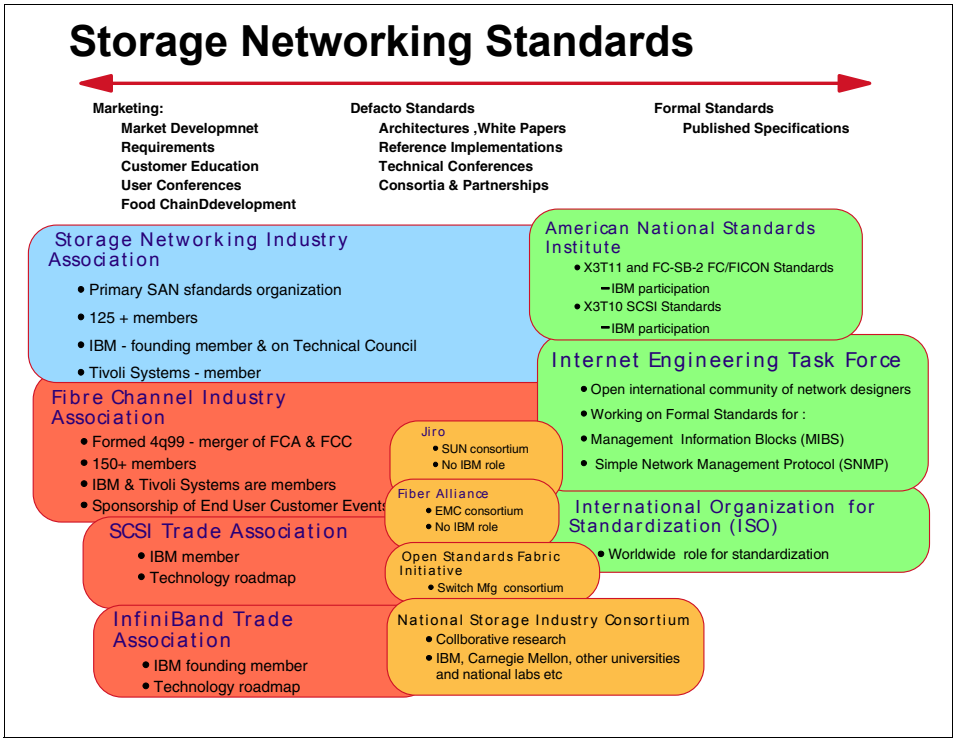


Figure 3-41 Groups involved in setting Storage Networking Standards

The roles of these associations and bodies fall into three categories:

Market development

These associations are involved in market development, establishing requirements, conducting customer education, user conferences, and so on. The main organizations are the Storage Network Industry Association (SNIA); Fibre Channel Industry Association (merging the former Fibre Channel Association and the Fibre Channel Loop Community); and the SCSI Trade Association (SCSITA). Some of these organizations are also involved in the definition of defacto standards.

Defacto standards

These organizations and bodies tend to be formed from two sources. They include working groups within the market development organizations, such as SNIA and FCIA. Others are partnerships between groups of companies in the industry, such as Jiro, Fibre Alliance, and the Open Standards Fabric Initiative (OSFI), which work as pressure groups towards defacto industry standards. They offer architectural definitions, write white papers, arrange technical conferences, and may reference implementations based on developments by their own partner companies. They may submit these specifications for formal standards acceptance and approval.

The OSFI is a good example, comprising the five manufacturers of Fibre Channel switching products. In July 1999, they announced an initiative to accelerate the definition, finalization, and adoption of specific Fibre Channel standards that address switch interoperability.

Formal standards

These are the formal standards organizations like IETF, ANSI, and ISO, which are in place to review, obtain consensus, approve, and publish standards defined and submitted by the preceding two categories of organizations.

IBM and Tivoli Systems are heavily involved in most of these organizations, holding positions on boards of directors and technical councils and chairing projects in many key areas. We do this because it makes us aware of new work and emerging standards. The hardware and software management solutions we develop, therefore, can provide early and robust support for those standards that do emerge from the industry organizations into pervasive use. Secondly, IBM, as the innovation and technology leader in the storage industry, wants to drive reliability, availability, serviceability, and other functional features into standards. The standards organizations in which we participate are included in the following sections.

3.13.2 Storage Networking Industry Association

Storage Networking Industry Association (SNIA) is an international computer industry forum of developers, integrators, and IT professionals who evolve and promote storage networking technology and solutions. SNIA was formed to ensure that storage networks become efficient, complete, and trusted solutions across the IT community. SNIA is accepted as the primary organization for the development of SAN standards, with over 125 companies as its members, including all the major server, storage, and fabric component vendors. SNIA also has a working group dedicated to the development of NAS standards. SNIA is committed to delivering architectures, education, and services that will propel storage networking solutions into a broader market. IBM is one of the founding members of SNIA, and has senior representatives participating on the board and in technical groups. For additional information on the various activities of SNIA, see its Web site at:

www.snia.org

3.13.3 Fibre Channel Industry Association

The Fibre Channel Industry Association (FCIA) was formed in the autumn of 1999 as a result of a merger between the Fibre Channel Association (FCA) and the Fibre Channel Community (FCC). The FCIA currently has more than 150 members in the United States and through its affiliate organizations in Europe and Japan. The FCIA mission is to nurture and help develop the broadest market for fibre channel products. This is done through market development, education, standards monitoring and fostering interoperability among members' products. IBM is a principal member of the FCIA.

The FCIA offers a fibre channel products qualifying program called SANmark. The purpose of the program is to provide the industry with an objective indication of how Fibre Channel products perform against reasonable standards and to permit the use of the trademarked term *SANmark*, and any associated logo(s), in the identification and promotion of products meeting the published test indices.

The SANmark Qualified Program goals are to:

- Make Fibre Channel solutions easy to use, easy to install, manage, configure, diagnose, and troubleshoot
- Ensure that Fibre Channel continues to attain the highest performance and installed base maturity available in the market
- Proliferate heterogeneous shared SAN resources and heterogeneous management framework over WAN connections

For additional information on the various activities of FCIA, see its Web site at:

www.fibrechannel.org

3.13.4 The SCSI Trade Association

The SCSI Trade Association (SCSITA) was formed to promote the use and understanding of small computer system interface (SCSI) parallel interface technology. The SCSITA provides a focal point for communicating SCSI benefits to the market, and influences the evolution of SCSI into the future. IBM is a founding member of the SCSITA. For more information, see its Web site at:

www.scsita.org

3.13.5 InfiniBand (SM) Trade Association

The demands of the Internet and distributed computing are challenging the scalability, reliability, availability, and performance of servers. To meet this demand a balanced system architecture with equally good performance in the memory, processor, and input/output (I/O) subsystems is required. A number of leading companies have joined together to develop a new common I/O specification beyond the current PCI bus architecture, to deliver a channel based, switched fabric technology that the entire industry can adopt. InfiniBand™ Architecture represents a new approach to I/O technology and is based on the collective research, knowledge, and experience of the industry's leaders. IBM is a founding member of InfiniBand (SM) Trade Association. For additional information, see its Web site at:

www.infinibandta.org

3.13.6 National Storage Industry Consortium

The National Storage Industry Consortium membership consists of over fifty US corporations, universities, and national laboratories with common interests in the field of digital information storage. A number of projects are sponsored by NSIC, including network attached storage devices (NASD), and network attached secure disks. The objective of the NASD project is to develop, explore, validate, and document the technologies required to enable the deployment and adoption of network attached devices, subsystems, and systems. IBM is a founding member of the NSIC. For more information, see its Web site at:

www.nsic.org

3.13.7 Internet Engineering Task Force

The Internet Engineering Task Force (IETF) is a large, open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture, and the smooth operation of the Internet. It is responsible for the formal standards for the Management Information Blocks (MIB) and for Simple Network Management Protocol (SNMP) for SAN management. For additional information on IETF, see its Web site at:

www.ietf.org

3.13.8 American National Standards Institute

American National Standards Institute (ANSI) does not itself develop American national standards. It facilitates development by establishing consensus among qualified groups. IBM participates in numerous committees, including those for Fibre Channel and storage area networks. For more information on ANSI, see its Web site at:

www.ansi.orgAddressing

3.13.9 INCITS

The International Committee for Information Technology Standards (INCITS) is the primary U.S. focus of standardization in the field of Information and Communications Technologies (ICT), encompassing storage, processing, transfer, display, management, organization, and retrieval of information. As such, INCITS also serves as ANSI's Technical Advisory Group for ISO/IEC Joint Technical Committee (JTC) 1. JTC 1 is responsible for International standardization in the field of Information Technology. From 1997 until 2001, INCITS operated under the name, Accredited Standards Committee, National Committee for Information Technology (NCITS). You'll find all their projects listed in 3.13.10, "List of evolved Fibre Channel standards" on page 141. For more information, see INCITS Web site at:

http://www.incits.org/

3.13.10 List of evolved Fibre Channel standards

Table 3-9 lists all current T11 Fibre Channel projects that are either approved standards or in proposal status. For the most recent status, visit the T11 Web site at:

http://www.t11.org

Table 3-9 T11 projects

Acronym	Title	Status
10 Bit Interface TR	10-bit Interface Technical Report	X3.TR-18:1997
10GFC	Fibre Channel - 10 Gigabit	Project 1413-D
FC-10KCR	Fibre Channel - 10 km Cost-Reduced Physical variant	INCITS 326: 1999
FC-AE	Fibre Channel Avionics Environment	INCITS TR-31-2002
FC-AE-2	Fibre Channel - Avionics Environment – 2	Project 1605-DT
FC-AL	FC Arbitrated Loop	ANSI X3.272:1996
FC-AL-2	Fibre Channel 2nd Generation Arbitrated Loop	INCITS 332: 1999
FC-AV	Fibre Channel - Audio-Visual	ANSI/INCITS 356:2001
FC-BB	Fibre Channel – Backbone	ANSI NCITS 342
FC-BB-2	Fibre Channel - Backbone – 2	Project 1466-D
FC-CU	Fibre Channel Copper Interface Implementation Practice Guide	Project 1135-DT
FC-DA	Fibre Channel - Device Attach	Project 1513-DT
FC-FG	FC Fabric Generic Requirements	ANSI X3.289:1996
FC-FLA	Fibre Channel - Fabric Loop Attachment	INCITS TR-20:1998
FC-FP	FC - Mapping to HIPPI-FP	ANSI X3.254:1994
FC-FS	Fibre Channel Framing and Signaling Interface	Project 1331-D
FC-FS-2	Fibre Channel - Framing and Signaling – 2	Project

Acronym	Title	Status
FC-GS	FC Generic Services	ANSI X3.288:1996
FC-GS-2	Fibre Channel 2nd Generation Generic Services	ANSI INCITS 288
FC-GS-3	Fibre Channel - Generic Services 3	NCITS 348-2000
FC-GS-4	Fibre Channel Generic Services 4	Project 1505-D
FC-HBA	Fibre Channel - HBA API	Project 1568-D
FC-HSPI	Fibre Channel High Speed Parallel Interface (FC-HSPI)	INCITS TR-26: 2000
FC-LE	FC Link Encapsulation	ANSI X3.287:1996
FC-LS	Fibre Channel - Link Services	Project
FC-MI	Fibre Channel - Methodologies for Interconnects Technical Report	INCITS TR-30-2002
FC-MI-2	Fibre Channel - Methodologies for Interconnects – 2	Project 1599-DT
FC-MJS	Methodology of Jitter Specification	INCITS TR-25:1999
FC-MJSQ	Fibre Channel - Methodologies for Jitter and Signal Quality Specification	Project 1316-DT
FC-PH	Fibre Channel Physical and Signaling Interface	ANSI X3.230:1994
FC-PH-2	Fibre Channel 2nd Generation Physical Interface	ANSI X3.297:1997
FC-PH-3	Fibre Channel 3rd Generation Physical Interface	ANSI X3.303:1998

Acronym	Title	Status
FC-PH:AM 1	FC-PH Amendment #1	ANSI X3.230:1994/AM1:1996
FC-PH:DAM 2	FC-PH Amendment #2	ANSI X3.230/AM2-1999
FC-PI	Fibre Channel - Physical Interface	INCITS 352
FC-PI-2	Fibre Channel - Physical Interfaces – 2	Project
FC-PLDA	Fibre Channel Private Loop Direct Attach	INCITS TR-19:1998
FC-SB	FC Mapping of Single Byte Command Code Sets	ANSI X3.271:1996
FC-SB-2	Fibre Channel - SB 2	INCITS 349-2000
FC-SB-3	Fibre Channel - Single Byte Command Set – 3	Project 1569-D
FC-SP	Fibre Channel - Security Protocols	Project 1570-D
FC-SW	FC Switch Fabric and Switch Control Requirements	INCITS 321:1998
FC-SW-2	Fibre Channel - Switch Fabric – 2	ANSI/INCITS 355-2001
FC-SW-3	Fibre Channel - Switch Fabric – 3	Project 1508-D
FC-SWAPI	Fibre Channel Switch Application Programming Interface	Project 1600-D
FC-Tape	Fibre Channel - Tape Technical Report	INCITS TR-24:1999
FC-VI	Fibre Channel - Virtual Interface Architecture Mapping	ANSI/INCITS 357-2001
FCSM	Fibre Channel Signal Modeling	Project 1507-DT

Acronym	Title	Status
MIB-FA	Fibre Channel Management Information Base	Project 1571-DT
SM-LL-V	FC - Very Long Length Optical Interface	ANSI/INCITS 339-2000
SM-AMD	SAN Management - Attribute & Method Dictionary	Project 1606-DT
SM-MM	SAN Management - Management Model	Project 1606-DT

10 Gb/s

10GFC is a working draft for the extensions to the FC-PH and FC-PI standard to support a data rate of 10.2 Gb/s. The proposal includes five different physical interface types — three shortwave and two longwave solutions:

- SW Parallel interface: the data is spread over four parallel fiber links
- SW Serial interface: 10.2 Gb/s over a single fiber link
- SW Coarse Wavelength Division Multiplexed (CWDM): data is multiplexed over four wavelengths on a single fiber
- LW Serial interface: 10.2 Gb/s over a single fiber link
- LW CWDM: data is multiplexed over four wavelengths on a single fiber

The Fibre Channel Industry Association (FCIA) completed the core content of its proposed 10 Gb/s Fibre Channel standard. The forthcoming 10GFC specification leverages the work done by the IEEE P802.3ae Task Force and shares a common link architecture and common components with Ethernet and InfiniBand. The proposed 10GFC standard will span link distances from 15 m up to 10 km and offer direct support for native dark fiber (DWDM) and SONET/SDH, while preserving the Fibre Channel frame format and size for full backward compatibility.

3.14 SAN software management standards

Traditionally, storage management has been the responsibility of the host server to which the storage resources are attached. With storage networks the focus has shifted away from individual server platforms, making storage management independent of the operating system, and offering the potential for greater flexibility by managing shared resources across the enterprise SAN

infrastructure. Software is needed to configure, control, and monitor the SAN and all of its components in a consistent manner. Without good software tools, SANs cannot be implemented effectively.

The management challenges faced by SANs are very similar to those previously encountered by LANs and WANs. Single vendor proprietary management solutions will not satisfy customer requirements in a multi-vendor heterogeneous environment. The pressure is on the vendors to establish common methods and techniques. For instance, the need for platform independence for management applications, to enable them to port between a variety of server platforms, has encouraged the use of Java™.

The Storage Network Management Working Group (SNMWG) of SNIA is working to define and support open standards needed to address the increased management requirements imposed by SAN topologies. Reliable transport of the data, as well as management of the data and resources (such as file access, backup, and volume management) are key to stable operation. SAN management requires a hierarchy of functions, from management of individual devices and components, to the network fabric, storage resources, data, and applications. This is shown in Figure 3-42.

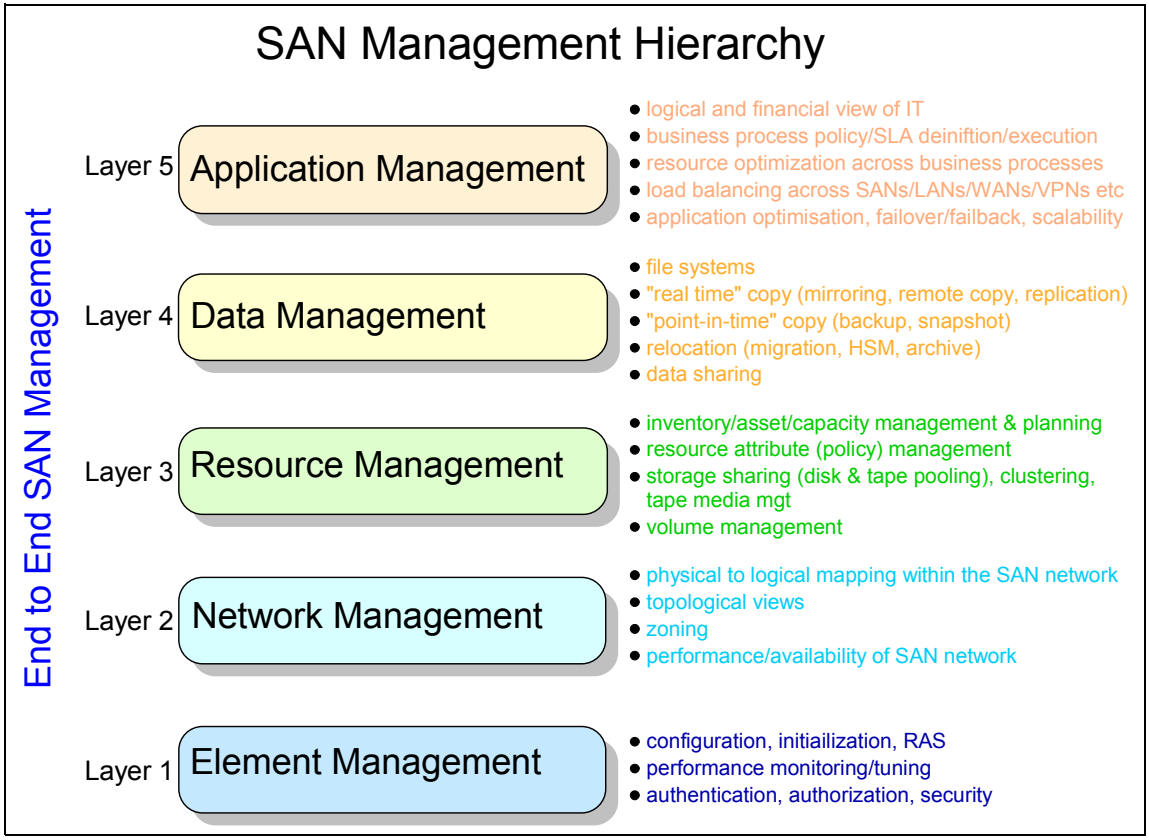


Figure 3-42 SAN management hierarchy

These can be implemented separately, or potentially as a fully integrated solution to present a single interface to manage all SAN resources.

3.14.1 Application management

Application management is concerned with the availability, performance, and recoverability of the applications that run your business. Failures in individual components are of little consequence if the application is unaffected. By the same measure, a fully functional infrastructure is of little use if it is configured incorrectly or if the data placement makes the application unusable. Enterprise application and systems management is at the top of the hierarchy and provides a comprehensive, organization-wide view of all network resources (fabric, storage, servers, applications).

A flow of information regarding configuration, status, statistics, capacity utilization, performance, and so on, must be directed up the hierarchy from lower levels. A number of industry initiatives are directed at standardizing the storage specific information flow using a Common Information Model (CIM) or application programming interfaces (API), such as those proposed by the Jiro initiative, sponsored by Sun Microsystems, and others by SNIA and SNMWG.

Figure 3-43 illustrates a common interface model for heterogeneous, multi-vendor SAN management.

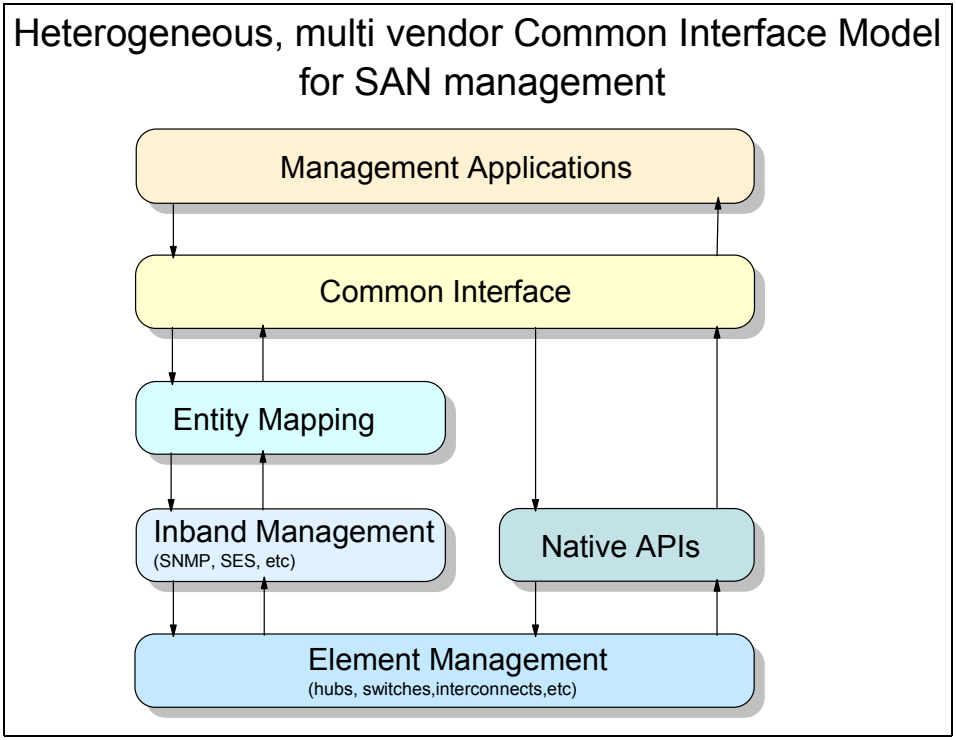


Figure 3-43 Common Interface Model for SAN management

3.14.2 Data management

More than at any other time in history, digital data is fueling business. Data management is concerned with Quality-of-Service (QoS) issues surrounding this data, such as:

- Ensuring data availability and accessibility for applications
- Ensuring proper performance of data for applications
- Ensuring recoverability of data

Data management is carried out on mobile and remote storage, centralized host attached storage, network attached storage (NAS), and SAN attached storage (SAS). It incorporates backup and recovery, archive and recall, and disaster protection.

3.14.3 Resource management

Resource management is concerned with the efficient utilization and consolidated, automated management of existing storage and fabric resources, as well as automating corrective actions where necessary. This requires the ability to manage all distributed storage resources, ideally through a single management console, to provide a single view of enterprise resources.

Without such a tool, storage administration is limited to individual servers. Typical enterprises today may have hundreds, or even thousands, of servers and storage subsystems. This makes impractical the manual consolidation of resource administration information, such as enterprise-wide disk utilization, or regarding the location of storage subsystems. SAN resource management addresses tasks such as:

- Pooling of disk resources
- Space management
- Pooling and sharing of removable media resources
- Implementation of “just-in-time” storage

3.14.4 Network management

Every e-business depends on existing LAN and WAN connections in order to function. Because of their importance, sophisticated network management software has evolved. Now SANs are allowing us to bring the same physical connectivity concepts to storage. And like LANs and WANs, SANs are vital to the operation of an e-business. Failures in the SAN can stop the operation of an enterprise.

SANs can be viewed as both physical and logical entities.

SAN physical view

The physical view identifies the installed SAN components, and allows the physical SAN topology to be understood. A SAN environment typically consists of four major classes of components:

- End-user computers and clients
- Servers
- Storage devices and subsystems
- Interconnect components

End-user platforms and server systems are usually connected to traditional LAN and WAN networks. In addition, some end-user systems may be attached to the Fibre Channel network, and may access SAN storage devices directly. Storage subsystems are connected using the Fibre Channel network to servers, end-user platforms, and to each other. The Fibre Channel network is made up of various interconnect components, such as switches, hubs, and bridges (Figure 3-44).

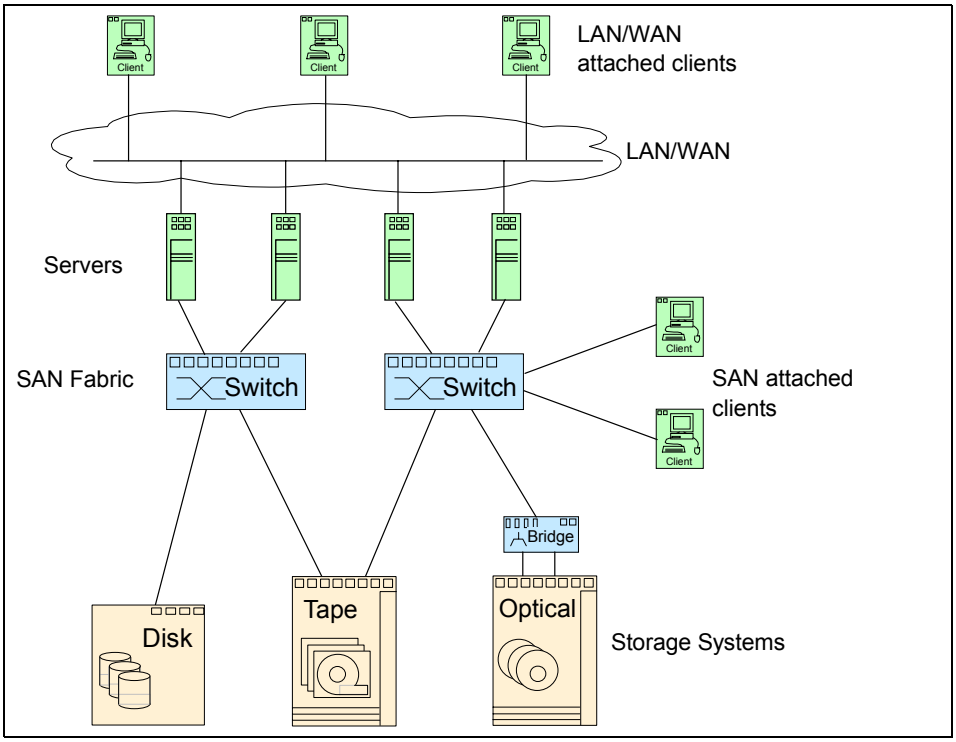


Figure 3-44 Typical SAN environment

SAN logical view

The logical view identifies and understands the relationships between SAN entities. These relationships are not necessarily constrained by physical connectivity, and they play a fundamental role in the management of SANs. For instance, a server and some storage devices may be classified as a logical entity. A logical entity group forms a private virtual network, or zone, within the SAN environment with a specific set of connected members. Communication within each zone is restricted to its members.

Network management is concerned with the efficient management of the Fibre Channel SAN — especially in physical connectivity mapping, fabric zoning, performance monitoring, error monitoring, and predictive capacity planning.

3.14.5 Device Management

The elements that make up the SAN infrastructure include intelligent disk subsystems, intelligent removable media subsystems, Fibre Channel switches, hubs and bridges, meta-data controllers, and out-board storage management controllers. The vendors of these components provide proprietary software tools to manage *their own* individual elements.This non heterogeneous management usually comprises software, firmware and hardware elements such as those shown in Figure 3-45.

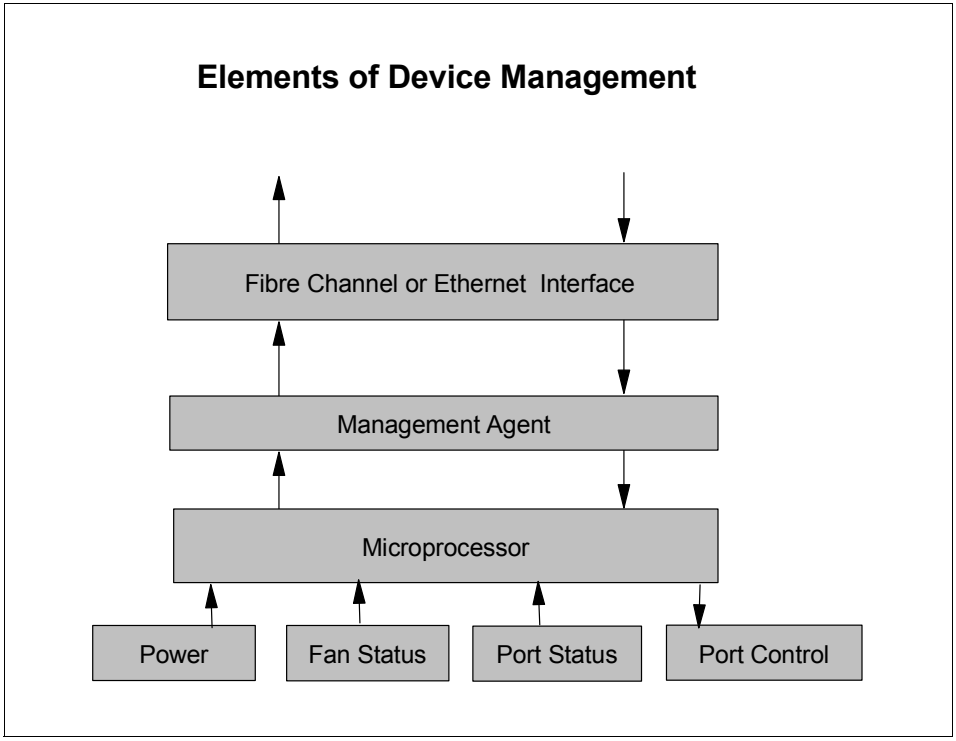


Figure 3-45 Device management elements

For instance, a management tool for a hub will provide information regarding its own configuration, status, and ports, but will not support other fabric components such as other hubs, switches, HBAs, and so on. Vendors that sell more than one element commonly provide a software package that consolidates the management and configuration of all of their elements. Modern enterprises, however, often purchase storage hardware from a number of different vendors.

Other companies, provide heterogeneous management facilities that are able to control, configure and monitor the equipment manufactured by several different companies. An example of this kind of software is Tivoli Storage Network Manager.

Fabric monitoring and management is an area where a great deal of standards work is being focused. Two management techniques are in use: inband and outband management.

Inband management

Device communications to the network management facility is most commonly done directly across the Fibre Channel transport, using a protocol called SCSI Enclosure Services (SES). This is known as inband management. It is simple to implement, requires no LAN connections, and has inherent advantages, such as the ability for a switch to initiate a SAN topology map by means of SES queries to other fabric components. However, in the event of a failure of the Fibre Channel transport itself, the management information cannot be transmitted. Therefore, access to devices is lost, as is the ability to detect, isolate, and recover from network problems. This problem can be minimized by provision of redundant paths between devices in the fabric.

- Inband developments: Inband management is evolving rapidly. Proposals exist for low level interfaces such as Return Node Identification (RNID) and Return Topology Identification (RTIN) to gather individual device and connection information, and for a Management Server that derives topology information. Inband management also allows attribute inquiries on storage devices and configuration changes for all elements of the SAN. Since inband management is performed over the SAN itself, administrators are not required to make additional TCP/IP connections.

Outband management

Outband management means that device management data are gathered over a TCP/IP connection such as Ethernet. Commands and queries can be sent using Simple Network Management Protocol (SNMP), Telnet (a text-only command line interface), or a Web browser Hyper Text Transfer Protocol (HTTP). Telnet and HTTP implementations are more suited to small networks.

Outband management does not rely on the Fibre Channel network. Its main advantage is that management commands and messages can be sent even if a loop or fabric link fails. Integrated SAN management facilities are more easily implemented, especially by using SNMP. However, unlike inband management, it cannot automatically provide SAN topology mapping.

- **Management Information Base (MIB):** A management information base (MIB) organizes the statistics provided. The MIB runs on the SNMP management workstation, and also on the managed device. A number of industry standard MIBs have been defined for the LAN/WAN environment. Special MIBs for SANs are being built by the Storage Networking Industry Association - SNIA. When these are defined and adopted, multi-vendor SANs can be managed by common commands and queries.
- **Outband developments:** Two primary SNMP MIBs are being implemented for SAN fabric elements that allow outband monitoring. The ANSI Fibre Channel Fabric Element MIB provides significant operational and configuration information on individual devices. The emerging Fibre Channel Management MIB provides additional link table and switch zoning information that can be used to derive information about the physical and logical connections between individual devices. Even with these two MIBs, outband monitoring is incomplete. Most storage devices and some fabric devices don't support outband monitoring. In addition, many administrators simply do not attach their SAN elements to the TCP/IP network.
- **Simple Network Management Protocol (SNMP):** This protocol is widely supported by LAN/WAN routers, gateways, hubs, and switches, and is the predominant protocol used for multi-vendor networks. Device status information (vendor, machine serial number, port type and status, traffic, errors, and so on) can be provided to an enterprise SNMP manager. This usually runs on a UNIX or NT workstation attached to the network. A device can generate an alert by SNMP, in the event of an error condition. The device symbol, or icon, displayed on the SNMP manager console, can be made to turn red or yellow, and messages can be sent to the network operator.

Element management is concerned with providing a framework to centralize and automate the management of heterogeneous elements and to align this management with application or business policy.

3.14.6 Fabric management methods

The SAN fabric can be managed using several remote and local access methods. Each vendor will decide on the most appropriate methods to employ on their particular product. Not all vendors are the same and from a management point of view it makes sense to investigate the possibilities before any investment is made.

3.14.7 Common methods

There are several access methods for managing a switch or director. The different methods have advantages and disadvantages. Some utilize inband elements. A comparison of the access methods is shown in Table 3-10.

Switches can be accessed simultaneously from different connections. If this happens, changes from one connection may not be updated to the other, and some may be lost. Make sure when connecting with simultaneous multiple connections, that you do not overwrite the work of another connection.

Table 3-10 Comparison of management access methods

Management method	Description	Local	In-band (Fibre Channel)	Out-of-band (Ethernet)
Serial Port	CLI locally from serial port on the switch	Yes	No	No
Telnet	CLI remotely via Telnet	No	Yes	Yes
SNMP	Manage remotely using the simple network management protocol (SNMP)	No	Yes	Yes
Management Server	Manage with the management server	No	Yes	No
SES	Manage through SCSI-3 enclosure services	No	Yes	No
Web Tools	Manage remotely through graphical user interface	No	Yes	Yes

Common methods

If your switch or director has a front panel display, it may be possible that it can be managed locally using the front panel buttons. See your switch reference manual for more information on this option. In order to manage a switch, you must have access to one of the available management methods. Telnet, SNMP, and IBM StorWatch Specialist require that the switch be accessible using a network connection. The network connection can be from the switch Ethernet port (outband) or from Fibre Channel (inband). We discuss outband in “Outband management” on page 152, and inband in “Inband management” on page 152.

Note: Some switches can be accessed simultaneously from different connections. If this happens changes from one connection may not be updated to the other, and some may be lost. Make sure when connecting with simultaneous multiple connections, that you do not overwrite the work of another connection.

There are several access methods for managing a switch or director. Table 3-11 summarizes the management access methods available.

Table 3-11 Comparison of Management Access Method

Management method	Description	Local	Inband (Fibre Channel)	Outband Ethernet
Switch / Director	**Manage locally from the front panel buttons on the switch / director**	Yes	No	No
Telnet commands	**Manage remotely using Telnet commands**	No	Yes	Yes
SNMP	**Manage remotely using the simple network management protocol (SNMP)**	No	Yes	Yes
Management Server	**Manage with the Management Server.**	No	Yes	No
SES	**Manage through SCSI-3 enclosure services**	No	Yes	No

Hardware setup for switch management

To enable remote connection to the switch, the switch must have a valid IP address. Two IP addresses can be set; one for the external out-of-band Ethernet port and one for inband Fibre Channel network access.

Managing with Telnet

To make a successful Telnet connection to a switch, the user needs:

- Switch name or IP address
- Username
- Password

Any host system that supports Telnet can be used to connect to the switch over the Ethernet. If the host supports a name server, the switch name can be used to effect the Telnet connection. If name service is not used to register network devices, then the IP address is used to connect to the switch. For example:

```
telnet [switch_name]
telnet 192.168.64.9
```

When the Telnet connection is made, the user is prompted for a username and password. The following section defines the default user names and passwords supplied with the switch. Both of these can be changed by the switch administrator.

Managing with SNMP

The resident SNMP agent allows remote switch management using IP over the Ethernet and Fibre Channel interfaces. This section provides an overview of key concepts about switch management that is based on simple network management protocol (SNMP).

Within the SNMP model, a manageable network consists of one or more manager systems (or network management stations), and a collection of agent systems (or network elements):

- A manager system runs a management application that monitors and controls the network elements.
- An agent system is a network device such as a Fibre Channel switch, a hub, or a bridge, that has an agent responsible for carrying out operations requested by the manager. Therefore, an agent is the interface to a managed device.

The manager uses SNMP to communicate with an agent. The switch agent supports both SNMP Version 1 (SNMPv1) and community-based SNMP Version 2 (SNMPv2C). SNMP allows the following management activities:

- A manager can retrieve management information, such as its identification, from an agent. There are three operations for this activity:
  - SNMP-GET
  - SNMP-NEXT
  - SNMP-BULKGET (SNMPv2C)
- A manager can change management information on the agent. This operation is called SNMP-SET.
- An agent can send information to the manager without being explicitly polled for. This operation is called a trap in SNMPv1 or a notification in SNMPv2C. Traps and notifications alert the manager to events that occur on the agent system, such as a restart. For the rest of the document, the term trap is used.

Management information base

The information on an agent is known as the management information base (MIB). The MIB is an abstraction of configuration and status information. A specific type or class of management information is known as an MIB object or variable. For example, the MIB variable, `sysDescr`, defines the description of an agent system. The existence of a particular value for an MIB object in the agent system is known as an MIB object instance, or simply instance. Some MIB objects have only a single instance for a given agent system. For example, the system description and the instance are denoted as `sysDescr.0`. Other MIB objects have multiple instances, for example, the operational status of each Fibre Channel port on a switch, where a particular instance can be denoted as `swFCPortOperStatus.5`.

Figure 3-46 shows that MIB objects are conceptually organized in a hierarchical tree structure. Each branch in the tree has a unique name and numeric identifier. Intermediate branches of the tree serve as a way to group related MIB objects together. The leaves of the tree represent the actual MIB objects. Figure 3-46 illustrates the tree structure, with special attention to the internet MIB tree and the Fibre Channel MIB tree.

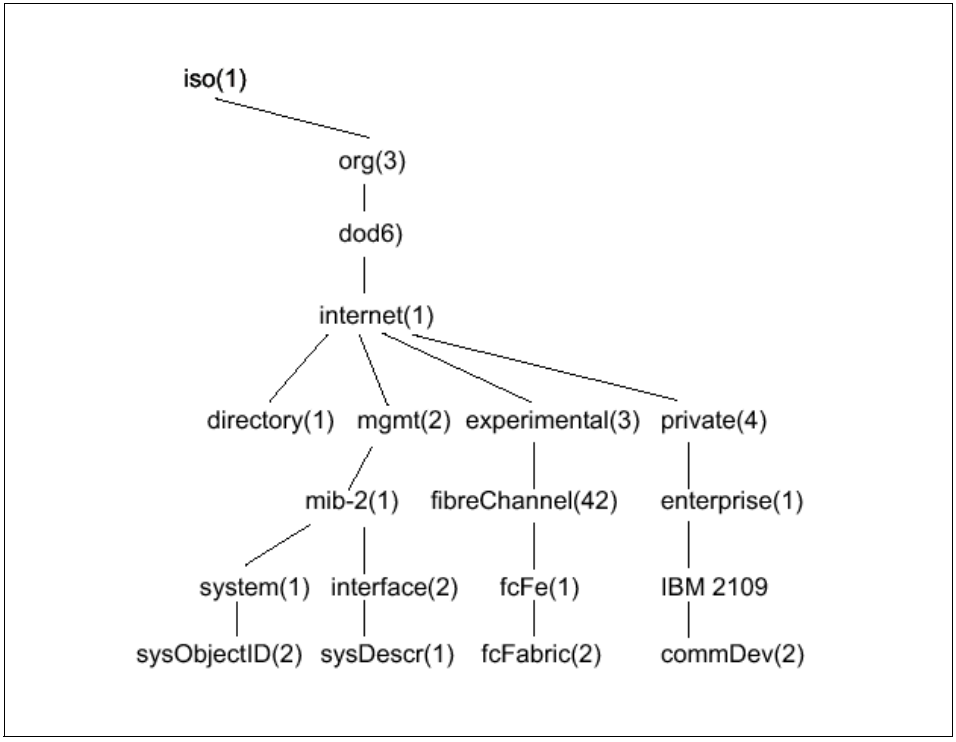


Figure 3-46 MIB tree

A MIB object is uniquely identified or named by its position in the tree. A full object identifier consists of each branch along the path through the tree. For example, the object `sysObjectID` has the full identifier of `1.3.6.1.2.1.1.2`. For readability, notation can be used, for example `{system 1}`.

The switch agent supports the following:

- SNMPv1 and SNMPv2c
- Command line utilities to provide access to configure the agent
- MIB-II system group, interface group, and SNMP group
- Fabric element MIB
- Vendor-specific MIBs
- Standard generic traps
- Enterprise specific traps

SNMP transports

The SNMP agent residing on the embedded processor supports UDP/IP over the Ethernet interface or any FC-IP interface. This transport provides an immediate *plug-and-play* support for the switch, once the IP address has been assigned.

MIB-II support

There are eleven groups of objects specified in MIB-II. The switch SNMP agent supports three of these groups. The eight additional groups do not apply. The three groups that are supported include:

- System group (object ID is {iso, org, dod, internet, mgmt, mib-2, 1})
- Interfaces group (object ID is {iso, org, dod, internet, mgmt, mib-2, 2})
- SNMP group (object ID is {iso, org, dod, internet, mgmt, mib-2, 11})

The following variables are modifiable using the SNMP **`set`** command, given an appropriate community with read-write access.

- sysDescr: System description, the default value is set as `Fiber Channel Switch`.
- sysContact: The identification and contact information for this switch. By default, this is set as `Field Support`.
- sysLocation: The physical location of the switch. The default setting is `End User Premise.`

Fabric element MIB support

There are five object groups defined:

- Configuration group
- Operation group
- Error group

- Accounting group
- Capability group

The agent supports all groups except the accounting group, which is better supported in the Fibre Channel port group of the vendor unique MIB.

Vendor unique MIB

Seven groups of MIBs are defined and supported.

- Switch system group
- Fabric group
- SNMP agent configuration group
- Fibre Channel port group
- Name server group
- Event group
- Fabric watch subsystem group (available with fabric watch license)

For more information, see “Available MIB and trap files” on page 161.

Generic traps

Setting up the switch SNMP connection to an existing managed network allows the network system administrator to receive the following generic traps.

- **coldStart**: Indicates that the agent has re initialized itself such that the agent configuration might be altered. This also indicates that the switch has restarted.
- **linkDown**: Indicates that an IP interface (Ethernet, loop back, or embedded N_Port) has gone down and is not available.
- **linkUp**: Indicates that an IP interface (Ethernet, loop back, or embedded N_Port) has become available.

  Note: *linkUp* and *linkDown* traps are not associated with removing or adding an Ethernet cable. This is strictly a driver indication that the interface is configured, operational, and available, and does not necessarily mean that the physical network cable is affected.

- **authenticationFailure**: Indicates that the agent has received a protocol message that is not properly authenticated. This trap, by default, is disabled but can be enabled using the command **`agtcfgSet`**, or by setting the MIB-II variable snapEnableAuthenTraps to enabled (1).

Enterprise specific traps

Four enterprise specific traps are supported:

- **swFault:** Indicates that the diagnostics detect a fault with the switch.
- **swSensorScn**: Indicates that an environment sensor changes its operational state. For example, a fan stops working. The VarBind in the trap data unit contains the corresponding instance of the sensor status.
- **swFCPortScn**: A notification that a Fibre Channel port changes its operational state. For example, the Fibre Channel port goes from online to offline. The VarBind in the trap data unit contains the corresponding instance of the port operational status.
- **swEventTrap**: A notification that an event has occurred and the event severity level is at or below the value set in the variable, *swEventTrapLevel*. See “Agent configuration” on page 160. The VarBind in the trap data unit contains the corresponding instance of the event index, time information, event severity level, the repeat count, and description.
- **swFabricWatchTrap**: This is sent by fabric watch about an event to be monitored.
- **swTrackChangesTrap:** Sent for tracking login, logout, and configuration changes.

Note: SNMP *swFCPortScn* traps are generated on GBIC insertion and removal even though the state remains offline.

Agent configuration

The list below shows the parameters that can be configured:

- SNMPv1 communities (up to 6)
- trap recipients (1 per community)
- sysName
- sysContact
- sysLocation
- authenticationFailure: indicates the agent has received a protocol message that is not properly authenticated. This trap, by default, is disabled.
- swEventTrap Level: indicates the swEventTrap severity level in conjunction with an event severity level. If the event severity level of an event is at or below the set value, the SNMP trap, swEventTrap, is sent to configured recipients. By default, this value is set at 0, implying that no swEventTrap is sent.

  There are several possible values:

  – 0: none
  – 1: critical

- 2: error
- 3: warning
- 4: informational
- 5: debug

Use the Telnet **agtcfgSet** command or SNMP to change these parameters.

Available MIB and trap files

You can download the MIB definitions and Enterprise trap definitions from:

www.ibm.com/storage/fcswitch

Note: Use the term port number to number the Fibre Channel ports on a switch. The value is from 0 through 15. In the various MIB definition files, there is the notion of port index, which by convention forbids the use of 0 as its value. For the switch, the port index for Fibre Channel ports ranges from 1 through 16 respectively.

Managing using the Management Server

The Management Server allows for the discovery of the physical and logical topology that comprise a Fibre Channel SAN. It provides several advantages for managing a Fibre Channel fabric:

- It is accessed by an external Fibre Channel node at address 0x'FFFFFA'.
- It is distributed on every 2109 Model S16 Switch within a fabric.
- It provides a flat view of the overall fabric configuration (without zones).

Because the Management Server is accessed using its well-known address, an application can access management information with a minimal knowledge of the existing configuration. An application accesses one well-known place to obtain management information about the entire fabric.

The fabric topology view exposes the internal configuration of a fabric for management purposes. It contains interconnect information about switches and devices connected to the fabric.

Under normal optional circumstances, a device (typically an FCP initiator) queries the name server for storage devices within its member zones. Because this limited view is not always sufficient, the Management Server provides the management application with a management view of the name server database.

Using the Management Server

The Management Server provides two management services:

- **Fabric configuration service**: Provides basic configuration management for topology information

- **Unzoned name server access**: Management view of the name server information

It also supports the following fabric configuration service requests:

- Get Interconnect Element List (GIEL)
- Get Interconnect Element Type (GIET)
- Get Domain Identifier (GDID)
- Get Management Identifier (GMID)
- Get Fabric Name (GFN)
- Get Interconnect Element Logical Name (GIELN)
- Get Interconnect Element Management Address List (GMAL)
- Get Interconnect Element Information List (GIEIL)
- Get Port List (GPL)
- Get Port Type (GPT)
- Get Physical Port Number (GPPN)
- Get Attached Port Name List (GAPNL)
- Get Port State (GPS)
- Register Interconnect Element Logical Name (RIELN)

For detailed information, see Fibre Channel Standard FC-GS-3, Revision 6.1, dated January 13, 2000.

syslogd daemon

A UNIX style syslogd daemon (syslogd) process is supported. The syslogd reads system events and forwards system messages to users, and writes the events to log files according to your system configuration.

Introduction

The syslogd daemon reads system events and forwards system messages to users and stores them in log files according to your system configuration. Events are categorized by facility and severity. The log process is used to log errors and system events on the local machine and is sent to a user or system administrator. The daemon is constantly running and ready to receive messages from system processes. The events are logged according to the statements in the configuration file, and syslogd is enabled to receive messages from a remote machine. The syslogd listens to UDP port 514 for system events. A remote machine does not have to be running UNIX to forward messages to syslogd, but it must follow the basic syslogd message format standard.

An example entry in a syslogd log file is:

```
Jul 18 12:48:00 sendmail [9558]: NOQUEUE: SYSERR (uucp):
/etc/mail/sendmail.cf: line 0: cannot open: No such file or directory
```

The first two items are the event date and time (as known by the machine where syslogd is running) and the machine name that issued the error. This is the local machine, if the message is generated by a task running on the same machine as syslogd, or a remote machine, if the message is received on UDP port 514. The first two items are always present. All other entries are message specific.

Note: The log file can be located on a different machine and can be locally mounted. A local error can be an error that occurs where syslogd is running, not on the machine where the error log physically resides.

The syslogd applications for NT and Win95 are available at no charge on several FTP servers on the Internet.

syslogd support

Switch firmware maintains an internal log of all error messages. The log is implemented as a circular buffer, with a storage capability of 64 errors. After 64 errors are logged, the next error message overwrites the messages at the beginning of the buffer.

If configured, the switch sends internal error messages to syslogd by sending the UDP packet to port 514 on the syslogd machine. This allows the storage of switch errors on a syslogd capable machine and avoids the limitations of the circular buffer.

The syslogd provides system error support using a single log file and can notify a system administrator in real time of error events.

Error message format

Each error message logged sends the following information:

- Error number (1 for the first error after startup, increments by one with each new error).
- The error message, exactly as it is stored in the error log and printed using the **`errShow`** command.

The error message includes the switch that reported the error with the event information:

- ID of the task that generated the error.
- Name of the task that generated the error.
- Date and time when the error occurred, as seen by the switch. This can be different from the first item in the log file, which is the time as seen by the syslogd machine. These two time values are different if the clocks in the switch and in the syslogd machine are not in sync.

- The error identifier consisting of a module name, a dash and an error name.
- The error severity
- Optional informational part
- Optional stack trace

Message classification

The syslogd messages are classified according to facility and priority (severity code). This allows a system administrator to take different actions depending on the error. The action taken, based on the message facility and priority, is defined in the syslogd configuration file. The switch uses the facility local7 for all error messages sent to the syslogd.

3.14.8 The SNIA Shared Storage Model

IBM is an active member of SNIA and fully supports SNIA's goals to produce the open architectures, protocols, and APIs required to make storage networking successful. IBM has adopted the SNIA Storage Model and we are basing our storage software strategy and roadmap on this industry-adopted architectural storage model.

IBM is committed to deliver the best-of-breed products in all aspects of the SNIA storage model:

- Block aggregation
- File/record subsystems
- Storage devices/block subsystems
- Services subsystems

In the area of block aggregation, IBM provides the SAN Volume Controller, the SAN Volume Controller for Cisco MDS 9000, and the SAN Integration Server, implemented in an in-band model.

In the area of file/record subsystems, IBM provides the SAN File System, a SAN wide file system implemented in an out-of-band model.

Both of these solutions adhere to open industry standards.You can learn more about the SNIA Shared Storage Model by visiting the SNIA Web site (available at http://www.snia.org).

3.14.9 Long distance links

The first thing to consider regarding a long distance link is whether the technology selected will cope with the distance of the link, see 3.12.1, “Limits” on page 134.

Secondly, longer distances introduce other factors to consider in the SAN design, one of which is latency. Latency increases since the time for the signal to travel the longer links, and has to be added to the normal latency introduced by switches and/or directors. Another point is that the time out values should allow for increased travel times. For this reason, parameters such as the E_D_TOV and R_A_TOV have to be evaluated.

3.14.10 Backup windows

As the amount of data being stored increases, so does the amount of time it takes to back up that data using a particular backup solution. There will come a point when there is not enough time to complete the backup and then it will be necessary to look for a different backup solution. The amount of time available to carry out the backup is known as the backup window.

Typically, the backup window will be a finite period of time during the night.

The SAN design will need to accommodate the need to back up all required data, perhaps from several servers. The data may be written to several backup devices or perhaps just to one. Whatever the strategy is, not only must the backup device or devices be able to offer the bandwidth to allow the data to be written, but the SAN must also have the required bandwidth.

The type of backup carried out can alter the amount of time taken to backup. For example, there may be two servers in a SAN with a three hour backup window. In order to do a full backup of either of the servers, may take two hours. Hence, there is not the time to do full backups of both machines every night. It may, however, be acceptable to fully backup one server on Monday and use incremental backups on other days, while the other server has a full backup on Tuesday with incrementals on other days. This may fit the requirements, but see Restore/disaster recovery time below.

3.14.11 Restore/disaster recovery time

There is also the need to consider the restore of data:

- It may be necessary to design the SAN in such a way that an accidentally deleted file can be restored on an ad hoc basis.
- There may be the requirement to restore a particular set of data in a given time. This may be to recover from the catastrophic failure of a particular application, or perhaps as part of a regular process, such as setting up a standard system for customer demonstrations (perhaps using IBM's Network Installation Manager.)
- The worst case to be considered would be the restore of an entire system or even the whole set of servers, perhaps following a serious environmental problem such as a flood.

The time to fully restore one or more systems may be significantly longer than the regular nightly backup window. For example, in the example above, we never do two full backups at the same time, but in a disaster recovery situation, we have to assume that we will be restoring all systems at the same time.

3.15 zSeries and S/390

The zSeries and S/390 platforms have a dedicated I/O subsystem that offloads workload from the processors allowing for high I/O data rates. Installations with high numbers of I/O devices and channels are very common. To solve the cabling and distance limitations of the Bus and Tag I/O connections, ESCON was introduced more than 10 years ago. ESCON, a serial interface using fiber optics as connecting media, has been delivering customers increased distance, reduced cable bulk, disk and tape pooling, clustering and data sharing, while providing management capabilities. Other vendors also supported ESCON, and it was adopted as a standard by NCITS.

FICON comes as an ESCON evolution. FICON is based on the Fibre Channel standard, so OS/390 and z/OS™ is positioned to participate in heterogeneous Fibre Channel based SANs.

FICON support started bridging from FICON channels to existing ESCON directors and ESCON control units, delivering value using channel consolidation, cable reduction, increased distance and increased device addressability.

We are now in the next phase that includes native FICON control units, attached either point-to-point, or switched point-to-point, using FICON capable directors. Since FICON is an upper layer protocol using standard Fibre Channel transport,

FICON directors are highly available Fibre Channel switches with capabilities that allow in-band management.

From an availability point of view, zSeries and S/390 offer the possibility of a Parallel Sysplex® configuration, the highest available configuration in the market. In a Parallel Sysplex, several processors are normally sharing I/O devices. ESCON directors have traditionally been used to allow sharing while reducing the number of control unit ports and the cabling requirements. With the introduction of FICON, processors and I/O can share a SAN with other platforms.

More information about Parallel Sysplex can be found at this Web site:

http://www-1.ibm.com/servers/eserver/zseries/pso/

Additional information about SAN on the zSeries and S/390 platforms, and future directions can be found at this Web site:

http://www-1.ibm.com/servers/eserver/zseries/san/

3.15.1 pSeries

There are different vendors that offer their own version of UNIX in the market. Each vendor (IBM, SUN, HP) offers its own hardware and different flavors of the UNIX operating system (AIX, Solaris, HP UX), each having some unique enhancements and often supporting different file systems (JFS, AFS®).

There are also several versions of management software available for the UNIX environment, such as Tivoli Storage Manager (TSM), and Veritas.

The IBM version of UNIX is the AIX Operating System and pSeries hardware. IBM currently offers a wide range of SAN ready pSeries servers from the entry servers up to Large Scale SP systems. Additional information regarding connecting pSeries servers to a SAN can be found in the IBM Redbook, *Practical Guide for SAN with pSeries,* SG24-6050.

More details can be found at the Web site:

http://www-1.ibm.com/servers/solutions/pseries/

3.15.2 xSeries

The platform of Intel based servers running Windows is a fast growing sector of the market. More and more of these servers will host mission critical applications that will benefit from SAN solutions such as disk and tape pooling, tape sharing, and remote copy.

IBM offerings in this platform include Netfinity®, NUMA-Q®, and xSeries. Additional information regarding connecting pSeries servers to a SAN can be found in the IBM Redpaper extract, *Implementing IBM server xSeries SANs*, REDP0416.

More details regarding xSeries can be found at the Web site:

http://www-1.ibm.com/servers/solutions/xseries/

3.15.3 iSeries

The iSeries platform uses the concept of single-level storage. The iSeries storage architecture (inherited from its predecessor systems System/38™ and AS/400®) is defined by a high-level machine interface. This interface is referred to as Technology Independent Machine Interface (TIMI). It isolates applications and much of the operating system from the actual underlying systems hardware. They are also unaware of the characteristics of any storage devices on the system because of single-level storage.

The iSeries is a multi-user system. As the number of users increase, you do not need to increase the storage. Users share applications and databases on the iSeries. As far as applications on the iSeries are concerned, there is no such thing as a disk unit. The idea of applications not being aware of the underlying disk structure is similar to the SAN concept.

Additional information regarding connecting pSeries servers to a SAN can be found in the IBM Redbook, *iSeries in Storage Area Networks: Implementing Fibre Channel Disk and Tape with iSeries,* SG24-6220-00.

More details regarding iSeries can be found at the Web site:

http://www-1.ibm.com/servers/solutions/iseries/

3.16 Security

Today, Fibre Channel fabrics are deployed in a variety of applications. Being developed to obtain SCSI connectivity for storage applications, Fibre Channel removes the traditional boundaries associated between SCSI devices. Storage interfaces, once propagated within the server or data center, may be extended outside the typically secure physical cabinet or location.

The Fibre Channel fabric and its components are also considered shared resources. This includes the switched fabric, the storage attached to the fabric and the services the fabric provides. This accessibility, plus the increase in both the quantity of data and users of that data, has heightened the awareness of the

IT community to security exposures. As such, Fibre Channel fabrics and their components are prone to the same types of security breaches once associated mostly with IP networks.

3.16.1 Control types

Many of the security layers can be configured to ensure that only the appropriate users or systems have access to data. Determining the specific values for these different variables and then correctly implementing steps to enforce the policies can be quite challenging. The following three control types are important to creating a complimentary matrix of checks and balances to gain security in a SAN:

- **Operational controls:** Processes that are typically implemented by security personnel. These controls have to work well in conjunction with management controls and technical controls. and are defined in security documents. Refer to , “Operational controls” on page 231 for more information on this type.
- **Management controls:** Procedures and policies that ensure that there is proper management oversight with regards to security.
- **Technical controls:** Security measures implemented by hardware and software. In conjunction with strong operational controls and management controls, technical controls can detect unauthorized access, track changes, implement security policies.

More information about policies can be found at the Web sites:

http://www.sans.org/resources/policies/
http://www.aits.uillinois.edu/security/securestandards.html#introduction

3.16.2 Vulnerabilities

It is a matter for debate if there is such a thing as 100% security for any kind of data. You could disconnect the information database from the rest of the world and “seal” it in concrete to preserve it — but then it would be rather inaccessible and probably not of much value in such an environment. Accessibility for your business applications on SAN will be achieved by interconnection and the network. The network, being a LAN or a SAN with hundreds or more servers, storage devices, and connections, can have plenty of weaknesses in security to consider. You may reduce the possible vulnerabilities into three groups categorized by traffic type, as shown in Figure 3-47.

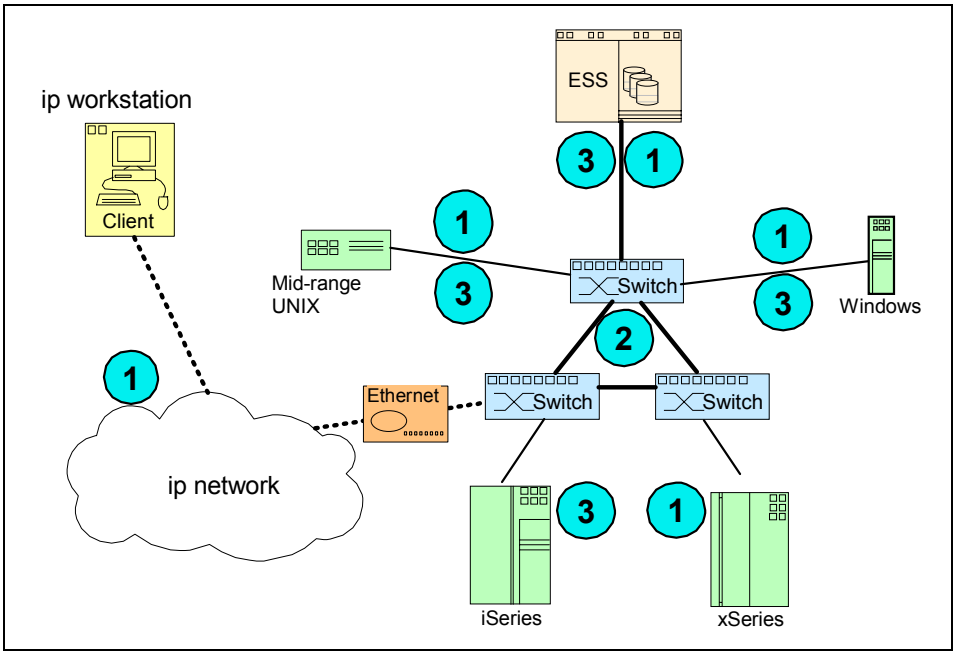


Figure 3-47 Vulnerabilities

Device-fabric traffic

This category includes any traffic that originates on servers and terminates on a fabric switch using Fibre Channel, IP, or other means. The devices can be systems that plug into the Fibre Channel ports of the SAN fabric via an HBA, as well as workstations connected through LANs and used by SAN administrators to manage the SAN. This type of risk can include traffic from unauthorized servers connected to the wrong port, invalid management device connections, and attacks such as denial of service (DOS).

Inter-switch traffic

Inter-switch traffic risks are limited to issues posed by the interconnection of SAN switches in a fabric. Inter-switch traffic includes the traffic generated when switches attempt to create E_Ports, as well as any traffic originating *and* terminating on a fabric switch. This category does not include traffic that only traverses the switches, such as data from a server to a storage device. The concern here is that if an invalid switch is plugged into an existing fabric it can cause significant disruptions, including modifications to zoning, and unauthorized access to fabric devices or resources.

Device-device traffic

This category includes risks coming from devices that do not talk directly to the switches, but instead use the fabric as the medium over which to talk to other connected devices, such as storage elements or hosts. The concerns here are unauthorized access and the possibility of a denial of service (DOS) attack.

The standard bodies are working on security enhancements to the Fibre Channel protocol. The vendors of Fibre Channel products will implement these enhancements using various products and features and adherence to FC security standards over time.

3.16.3 Fibre Channel security

Since April 2002, the ANSI T11 group has been working on FC-SP, a proposal for the development of a set of methods that allow security techniques to be implemented in a SAN.

Up until now, fabric access of Fibre Channel components was attended to by identification (**who are you?**). This information could be used later to decide if this device was allowed to attach to storage (by zoning), or it was just for the propagation of information (for example, attaching a switch to a switch) — but it was not a criteria to refuse an inter-switch connection.

As the fabric complexity increases, more stringent controls are required for guarding against malicious attacks and accidental configuration changes. Additionally, increasingly more in-fabric functionality is being proposed and implemented that requires a closer focus on security.

The customer demand for protecting the access to data within a fabric necessitates the standardization of interoperable security protocols. The security required within a Fibre Channel fabric to cope with attempted breaches of security can be grouped into four areas:

Authorization	**I tell you what you're allowed to do!**
Authentication	**Tell me about yourself; I will decide if you may log in**. A digital verification of who you are, it ensures that received data is from a known and trusted source.
Data confidentiality	Cryptographic protocols ensure that your data was unable to be read or otherwise utilized by any party while in transit.
Data integrity	Verification that the data you sent has not been altered or tampered with in any way.

3.16.4 Security mechanisms

In the topics that follow, we overview some of the common approaches to securing data.

Encryption

in 1976, W.Diffie and M.Hellman (their initials are found in **DH**-CHAP) introduced a new method of encryption and key management. A public-key cryptosystem is a cryptographic system that uses a pair of unique keys (a public key and a private key). Each individual is assigned a pair of these keys to encrypt and decrypt information. A message encrypted by one of these keys can only be decrypted by the other key in the pair:

- The public key is available to others for use when encrypting information that will be sent to an individual. For example, people can use a person's public key to encrypt information they want to send to that person. Similarly, people can use the user's public key to decrypt information sent by that person.
- The private key is accessible only to the individual. The individual can use the private key to decrypt any messages encrypted with the public key. Similarly, the individual can use the private key to encrypt messages, so that the messages can only be decrypted with the corresponding public key.

That means: exchanging keys is no longer a security concern. *A* has a public key and a private key. *A* can send the public key to anyone else. With that public key, *B* can encrypt data to be sent to *A*. Since the data was encrypted with *A's* public key, *only A* can decrypt that data with his private key. If *A* wants to encrypt data to be sent to *B*, *A* needs *B's* public key.

If *A* wants to testify that it was the person that actually sent a document, *A* will encrypt and protect the document with his private key, while others can decrypt it using *A's* public key; they will know that in this case only *A* could have encrypted this document. Each individual involved needs their own public/private key combination.

The remaining question is: when you initially receive someone's public key for the first time, how do you know it is them? If “spoofing” someone's identity is so easy, how do you knowingly exchange public keys and how do you trust the user is who they say they are? The answer is to use a digital certificate. A digital certificate is a digital document that vouches for the identity and key ownership of an individual — it guarantees authentication and integrity.

The ability to perform switch to switch authentication in FC-SP enables a new concept in Fibre Channel: the secure fabric. Only switches that are *authorized* and properly *authenticated* are allowed to join the fabric.

Whereas, authentication in the secure fabric is twofold: the fabric wants to verify the identity of each new switch before joining the fabric, and the switch that is wanting to join the fabric wants to verify that it is connected to the right fabric. Each switch needs a list of the WWNs of the switches authorized to join the fabric, and a set of parameters that will be used to verify the identity of the other switches belonging to the fabric.

Manual configuration of such information within all the switches of the fabric is certainly possible, but not advisable in larger fabrics. And there is the need of a mechanism to manage and distribute information about authorization and authentication across the fabric.

Authorization database

The fabric authorization database is a list of the WWNs and associated information like domain-IDs of the switches that are authorized to join the fabric.

Authentication database

The fabric authentication database is a list of the set of parameters that allows the authentication of a switch within a fabric. An entry of the authentication database holds at least the switch WWN, authentication mechanism Identifier, and a list of appropriate authentication parameters.

Authentication mechanisms

In order to provide the equivalent security functions that are implemented in the LAN, the ANSI T11-group is considering a range of proposals for connection authentication and integrity which can be recognized as the FC adoption of the IP security standards. These standards propose to secure FC traffic between all FC ports and the domain controller. These are some of the methods that will be used:

- **FCPAP** refers to Secure Remote Password Protocol (SRP), RFC 2945.
- **DH-CHAP** refers to Challenge Handshake Authentication Protocol (CHAP), RFC 1994.
- **FCsec** refers to IP Security (IPsec), RFC 2406.

3.16.5 IP security

There are standards and products available originally developed for the LAN and already installed worldwide. These can easily be added into and used by SAN solutions.

Simple Network Management Protocol (SNMP) had been extended for security functions to SNMPv3. The SNMPv3 specifications were approved by the Internet Engineering Steering Group (IESG) as full Internet Standard in March 2002.

IPSec uses cryptographic techniques obtaining management data that can flow through an encrypted tunnel. Encryption makes sure that only the intended recipient can make use of it. (RFC 2401).

Other cryptographic protocols for network management are Secure Shell (SSH) and Transport Layer Security (TLS, RFC 2246). TLS was formerly known as Secure Sockets Layer (SSL). They help ensure secure remote login and other network services over insecure networks.

Remote Authentication Dial-In User Service (RADIUS) is a distributed security system developed by Lucent Technologies InterNetworking Systems. RADIUS is a common industry standard for user authentication, authorization, and accounting (RFC 2865). The RADIUS server is installed on a central computer at the customer's site. The RADIUS Network Access Server (NAS), which would be an IP-router or switch in LANs and a SAN switch in SANs, is responsible for passing user information to the RADIUS server, and then acting on the response which is returned to either permit or deny the access of a user or device.

A common method to build trusted areas in IP networks is the use of firewalls. A firewall is an agent which screens network traffic and blocks traffic it believes to be inappropriate or dangerous. You will use a firewall to filter out addresses and protocols you do not want to pass into your LAN. A firewall will protect the switches connected to the management LAN and allows only traffic from the management stations and certain protocols that you will define.

3.16.6 Virtualization

The technique of making something appear to be something different is called virtualization. The difference might be quite subtle, or there may be quite a significant change in functionality.

As an example, it is useful to be able to tell an application to access its data on one (logical) device, but to locate the data on several physical disk drives. This can help performance and depending on the method, will often increase resilience. In effect, the application is accessing a virtual device or volume and this is a form of virtualization.

IBM offers a product called the IBM TotalStorage SAN Volume Controller to provide storage virtualization. The IBM SAN Volume Controller is a hardware and software clustered node solution which allows the customer to pool storage volumes from IBM and non-IBM devices into a single reservoir of capacity with centralized management. The customer can allocate this storage as needed in logical virtual devices to hetergeneous servers. The SVC enables a tiered storage environment in which the cost of storage can be better matched to the value of data, and supports improved application availability by insulating host applications from changes to the physical storage infrastructure. The SAN Volume Controller includes a dynamic data-migration function thats helps administrators migrate storage from one device to another, without taking it offline. This helps administrators to reallocate and scale storage capacity without disrupting applications. The San Volume Controller makes possible virtual tape and replication and also the chance to use cheap storage as the backup for expensive but highly reliable storage like the ESS.

Another form of virtualization is to make a device appear to be something quite different. As an example, a particular system which sells in relatively small numbers may only support a limited number of peripheral devices. It may, perhaps, have only one particular type of tape drive that is supported and the price per megabyte of storage media may be very high. Due to the economies of scale, media for a different form of tape storage, commonly used on a platform that sells in much higher numbers than the former may be much cheaper. In this case, if there is a need to back up very large amounts of data, there is a high incentive to make the first platform be able to communicate with the second form of tape device. Sometimes a device driver can be written for the operating system to make the device work. It is often more practical to use the approach of virtualization. That is, to make the tape drive “look like” the kind of tape drive that the first platform supports.

Another type of virtualization is at the file system level. The IBM TotalStorage® SAN File System (based on IBM Storage Tank™ technology) is a network-based heterogeneous file system for data sharing and centralized policy-based storage management in an open environment. The SAN File System enables host systems to plug-in to a common SAN-wide file structure. Files and file systems are no longer managed by individual computers but managed as a centralized IT resource with a single point of administrative control.Data sharing and collaboration is enhanced through a single global name space across heterogeneous server platforms, with high performance and full locking support. The San File System reduces storage needs by pooling available and temporary filespace across heterogeneous servers and storage platforms and reduces the need to maintain duplicate data for sharing. The San File System can simplify and lower the cost of data backups by offering application server free backup and leveraging built in file-based FlashCopy Functions.

3.17 Solutions

The main support for Fibre Channel development came from the workstation market. While in the mainframe platform the I/O channels have evolved allowing storage attachment and sharing through multiple high speed channels, and fiber optic cabling was introduced with ESCON®, workstations have been using SCSI as the common interface for storage attachment.

For storage interconnections, a SCSI interface had been traditionally used, but as data volumes and performance requirements increased, SCSI limitations started to surface: bulky cables, shared bus architecture that limits performance due to bus arbitration, limited distance of up to 25 m, and limited addressing of up to 15 targets. The continual growth of storage capacity requirements, data sharing needs and performance issues, made clear that it was necessary to overcome SCSI limitations. IBM introduced the IBM Serial Storage Architecture, and the IBM 7133 Disk Storage that solved many of the limitations of SCSI devices.

Another solution was the Fibre Channel interface. Mapping SCSI over the Fibre Channel Protocol was the solution that allowed access to multiple storage devices, extended distances, reduce cable bulk, and sharing of devices. Initially Fibre Channel arbitrated loop (FC-AL) was implemented to connect disk devices to hosts, and provided many benefits like smaller cables and connectors, faster data transfers and longer distances. Today, the arbitrated loop solution may still work for a department or workgroup, but does not offer the performance and connectivity required by an enterprise SAN, so different vendors offer Fibre Channel HBAs which provide for point-to-point connection, as well as connections to a Fibre Channel switched fabric.

These are some of the many reasons to implement a SAN fabric implementation:

- **Storage consolidation:** Storage devices can be shared with more servers without increasing the number of ports in the device.
- **Clustering:** For high availability solutions, a SAN allows shared storage connections and provides for longer distances between devices.
- **LAN free backup:** The ability to consolidate tape drives and tape libraries and share them among several backup hosts provides the opportunity to optimize the utilization of the tape drives. The result is more data can be backed up with the same number of, or less, drives.

To expand the benefits of a SAN across longer distances and allow more companies to realize the benefits of a SAN, several projects are currently underway. One solution that may make use of Fibre Channel is the Internet Protocol (IP). It requires an upper layer protocol that takes care of sending IP packets as Fibre Channel Sequences. One project in the T11 committee deals

with Fibre Channel link encapsulation (FC-LE). As a result of this project's work, there is a new protocol known as FC-IP, or FCIP, that will allow for greater distances by IP encapsulating the Fibre Channel protocol.

A new protocol called iSCSI will enable SCSI commands to be packaged and sent over existing IP networks. This will allow companies that currently do not have the resources for a Fibre Channel network to build a SAN utilizing their existing IP network.

Example applications that exploit SANs

Today, we have information available in many different forms: text, images, audio, and video, which we usually refer to as multimedia. Given the storage capacity and performance levels that computer systems are able to provide today, and what can be expected in the future, it is becoming practical to store, distribute, and retrieve more information in digital form. This can dramatically increase the amount of data stored, the transmission throughput, and the sharing requirements.

SANs can fulfill many of the demands of multimedia applications. Some examples of these are described in the following sections.

Video editing

Digital video editing is usually performed in standard computers with specialized video boards. Due to the size of the video files they have to handle, workstations normally have big and fast local storage devices but are interconnected by a relatively low bandwidth LAN. Local copies must be made of the files to be worked on, and once again when the work is finished, either from a server through the LAN or using removable media.

Interconnecting all workstations and storage by means of a SAN has the following advantages:

- **Bandwidth:** A properly designed SAN can provide enough bandwidth for the workstations to access the data in shared storage eliminating the need and the time required for local copies.
- **Storage efficiency:** Consolidating storage in a single pull allows each workstation have the amount required for each task eliminating excess storage in each workstation.
- **Reliability:** Installing fault tolerant storage like RAID arrays and being able to perform simultaneous centralized backups and restores.
- **Workstation flexibility:** A task initiated in one workstation can be continued in another one. It is also possible to introduce a new platform and share existing data. This also eliminates the workstation as a single point of failure.

- **Location flexibility:** SAN distances allow workstations to be far apart from each other, making it possible to have some workstations in special locations, like soundproof rooms.
- **Simultaneous viewing:** The work in progress can be monitored without interruptions to the creative staff.
- **Separation of duties:** Creative staff can concentrate in their work leaving repetitive tasks like digitizing material and recording output to tape to other personnel. This can also help consolidating and saving additional equipment like VCRs.

In order to allow several users to share the storage and working copies simultaneously, some kind of management software is required.

Pre-press

Some of the characteristics of today's graphic industry are:

- **Multiple parties involved:** Printing facilities, advertising agencies, graphic designers, clients.
- **Gigantic size files:** Raster image processors (RIPs) and open pre-press interface (OPI) applications that feed computer to plate (CTP) devices often generate files in the gigabyte range, especially when dealing with large images at 1200 dpi resolution, making LANs too slow for reasonable file transfer times.
- **Mixed platforms:** High end graphic workstations running on Apple Macintosh workstations running Windows NT® or 2000, application and file servers running UNIX.

LANs have been successfully handling the mixed platform environment, but as file sizes have increased, LAN speed has become the bottleneck.

SANs can also handle the mixed platform environment and not only offer a greater bandwidth, but also reduce the processor overhead to move the larger data files. This makes it possible to connect the different workstations directly to shared storage and eliminates the need for making local copies of large files.

Video distribution

There is a lot of research in process related to multimedia systems. It deals with storing and transmitting large amounts of time critical data between storage systems and end users.

Some of the characteristics of a video server are:

- Huge amount of data stored
- Large number of potential concurrent users
- Real time requirements to allow all users receive a jitter free video signal

- User interaction, such as title selection, play, pause, stop
- Availability 24 hours x 365 days required
- Multiple server configurations to be able to handle all potential users
- Multiple storage devices to handle the amount of data stored

Without entering into the design considerations of such a system, SANs offer the capacity of handling the multiple server, multiple storage configuration with the bandwidth, the high availability, and the scalability characteristics required by these kinds of applications.

3.18 Emerging technologies

Many “hot” technologies emerge, some stay around and some go the way of the dinosaurs. Here is what is “hot” at the time of writing.

3.19 iSCSI

Internet SCSI (iSCSI) is a transport *protocol* that carries SCSI commands from an initiator to a target. It is a data storage networking protocol that transports standard Small Computer System Interface (SCSI) requests over the standard Transmission Control Protocol/Internet Protocol (TCP/IP) networking technology.

iSCSI enables the implementation of IP-based storage area networks (SANs), enabling customers to use the same networking technologies — from the box level to the Internet — for both storage and data networks. As it uses TCP/IP, iSCSI is also well suited to run over almost any physical network. By eliminating the need for a second network technology just for storage, iSCSI will lower the costs of deploying networked storage and increase its potential market.

One of the major advantages is that as iSCSI carries SCSI commands over existing IP networks, it has an innate and important ability to facilitate the transfer of data over both inter- and intra-nets, and to manage storage over long distances.

iSCSI, in simple terms, works in this way: when an end user or application sends a request, the operating system generates the appropriate SCSI commands and data request. These then go through encapsulation procedures. A packet header is added before the resulting IP packets are transmitted over a TCP/IP connection. At the receiving end when the packets are received they are unravelled and the SCSI commands are separated from the data request. The SCSI commands are sent on to the target storage controller and ultimately the SCSI storage device. iSCSI is a bi-directional protocol which means that it can also be used to return data in response if required.

3.20 iFCP

Internet Fibre Channel Protocol (iFCP) is a mechanism for transmitting data to and from Fibre Channel storage devices in a SAN, or on the Internet using TCP/IP.

FCP gives the ability to incorporate already existing SCSI and Fibre Channel networks into the Internet. iFCP is able to be used in tandem with existing Fibre Channel protocols, such as FCIP, or it can replace them. Whereas FCIP is a tunneled solution, iFCP is an FCP routed solution.

The appeal of iFCP is that for customers that have a wide range of FC devices, and who want to be able to connect these with the IP network, iFCP gives the ability to permit this. iFCP can interconnect FC SANs with IP networks, and also allows customers to use the TCP/IP network in place of the SAN.

iFCP is a gateway-to-gateway protocol, and does not simply encapsulate FC block data. Gateway devices are used as the medium between the FC initiators and targets. As these gateways can either replace or be used in tandem with existing FC fabrics, iFCP could be used to help migration from a Fibre Channel SAN to an IP SAN, or allow a combination of both.

3.21 FCIP

Fibre Channel over IP (FCIP) is also known as Fibre Channel tunneling or storage tunneling. It is a method for allowing the transmission of Fibre Channel information to be tunnelled through the IP network. Because most organizations already have an existing IP infrastructure, the attraction of being able to link geographically dispersed SANs, at relatively low cost, is enormous.

FCIP encapsulates Fibre Channel block data and subsequently transports it over a TCP socket. TCP/IP services are utilized to establish connectivity between remote SANs. Any congestion control and management, as well as data error and data loss recovery, is handled by TCP/IP services, and does not affect FC fabric services.

The major point with FCIP is that is does not replace FC with IP, it simply allows deployments of FC fabrics using IP tunnelling. The assumption that this might lead to is that the “industry” has decided that FC based SANs are more than appropriate, and that the only need for the IP connection is to facilitate any distance requirement that is beyond the current scope of an FCP SAN.

As a starting point for an IP storage discussion, the Storage Networking Industry Association's IP Storage Forum (IPS Forum) is a vendor-neutral environment for

end users to become informed on the current and future directions of IP-based storage technology:

http://www.snia.org/ipstorage/home/

4

Extending fiber over distance with DWDM

This chapter gives an overview of attaining higher throughput within a fibre network environment. We describe the technology and methods used to extend fibre networks over distance. Initially it is important to understand the distances and terminology used to describe the geography that we could extend over. We then describe the components that enable this today. Technologies that we include in this overview are traditional time-division multiplexing (TDM), wavelength division multiplexing (WDM) and dense wave division multiplexing (DWDM).

This chapter gives an understanding of what this means for the storage area networking (SAN) arena today.

© Copyright IBM Corp. 2004. All rights reserved.

4.1 Why we need fiber saving devices

It's the need for speed and more bandwidth in the same over-subscribed fibre (laying new fibre is very expensive), and driving existing fibre harder is an economically attractive solution (especially when you can mix protocols).

There are two ways of increasing the effective capacity of already deployed fibre. You can increase the bit rate of existing systems or increase the number of wavelengths being transmitted down the fibre. In this chapter we will discuss both methods here. We will also discuss both the architecture and resilience as these infrastructures are likely to be mission critical to your business.

By its very nature, today's storage area networking (SAN) environments continue to evolve. This continues to be driven by many diverse business factors. We will consider many of these driving factors and how each solution addresses these problems. We will demonstrate how specific vendors solutions integrate with IBM SAN solutions, and the business problems they address.

Graphically dispersed networks historically have been core to edge networking architectures that have evolved from LAN to MAN and/or WAN. This redbook is based around extending SAN over distance to facilitate geographically dispersed location. This book will be addressed from a storage center solution perspective, working from storage up to servers, not from servers down.

To facilitate this we will consider the core to edge architecture and how it fits in the distance market place. For the purpose of this book, SAN will be regarded as the center of the enterprise. This seems in line with current industry trends as enterprises find the value and size of data to be growing exponentially.

4.1.1 Storage Area Network

In this e-commerce economy, it is no secret that data is business. Because of the increased dependence on information, the storage needs of companies are growing exponentially. Research shows that storage needs for traditional bricks-and-mortar companies are doubling every year, and every 90 days for dot coms. This growth raises new concerns for the maintenance and protection of valuable data resources.

Historically in open system storage environments, physical interfaces to storage consisted of parallel SCSI channels supporting a small number of SCSI devices. Storage Area Networks use new technologies to connect greater numbers of servers and devices. Deployments of SANs today are exploiting the storage-focused capabilities of Fibre Channel. The Fibre Channel SAN consists of hardware components such as storage subsystems, storage devices, and servers that are attached to the SAN via interconnect entities (host-bus adapters,

bridges, hubs, switches). Another hardware element being deployed in SANs is commonly referred to as a SAN appliance or SAN server. These SAN appliances are computing elements attached directly to the SAN or installed in the storage data path. These SAN appliances are responsible for managing the Fibre Channel topology and additionally providing an abstraction of storage.

The heart of the storage management software is the virtualization. The term virtualization, when it pertains to disk storage, refers to the representation of a storage unit or data to the operating system and/or application running on an application server. The storage unit or data presented is decoupled from the actual physical storage where the information may be contained. There is some method for providing the translation between the logical and physical storage. Once the storage is abstracted, storage management tasks can be performed with a common set of tools from a centralized point, which will greatly reduce the cost of administration.

4.1.2 Metropolitan Area Network

Metropolitan Area Networks (MAN) will typically be a city wide solution, connecting systems across areas of 50 to 100 km. These networks will typically carry traffic between office locations within a metropolitan area. A MAN will often carry many diverse networking protocols, and these protocols will all have their own speed and performance characteristics. It can also be an intermediate media interfacing between a Storage Area Network (SAN) and a Wide Area Network (WAN). This begins to build a hierarchical model of the SAN to MAN to WAN. MANs have traditionally been based on Synchronous Optical Network (SONET) or Synchronous Digital Hierarchy (SDH) technology using point-to-point or ring topologies with add/drop multiplexers (Aims).

The MAN has two critical goals to achieve. It must meet the needs created by the dynamics of the ever-increasing bandwidth requirements. It must also address the growing connectivity requirements and access technologies that are resulting in demand for high-speed, customized data services.

4.1.3 Wide Area Networks

Wide Area Networks (WAN) are at the very core of a global network. Like the MAN, the WAN's purpose is as a transport medium, and this means they need to be resilient with a high level of capacity. These networks are often provisioned by SONET or SDH technology. Due to the high demand for bandwidth these solutions are experiencing large demands on their fibre. WAN’s are likely to be public (in the case of the Internet) or private in the case of a company’s self

deployed dedicated WAN. Some WANs are based on an agreed bandwidth within a public network, and this leads to virtual private networks. Most large corporate customers will deploy one of these last two options, or a combination of the two.

4.2 Fiber extending device concepts

There are three methods that are employed to extend the distance of fibre SAN solutions. The first we will describe is channel extenders. These translate incoming signals to another protocol that can be sent further, often this utilizes compression algorithms to help achieve speed at long distances. The second is known as Time Division Multiplexing or TDM. Multiple channels are transmitted on a single carrier by increasing the modulation rate and allotting a time slot to each channel. However, increasing the bit rate of a system requires more sophisticated high-speed electronics at the transmitting and receiving ends of the communications link. And as the bit rate increases, inherent modulation limiting characteristics of optical fibers become dominant. Chromatic and polarization mode dispersion cause pulse spreading, which affects the signal quality over longer transmission distances.

4.2.1 Channel extenders

Channel extenders are used to provide longer cable distances. Most optical interfaces are multimode cable. Extenders convert the multimode interface to single mode and boost the power on the laser. Typically, an extender will provide a single mode cable distance of 30 km or 18 miles.

Compression

Compression is the reduction in size of data in order to save space or transmission time. For data transmission, compression can be performed on just the data content or on the entire transmission unit (including header data) depending on a number of factors. Content compression can be as simple as removing all extra space characters, inserting a single repeat character to indicate a string of repeated characters, and substituting smaller bit strings for frequently occurring characters. This kind of compression can reduce a text file to 50 percent of its original size. Compression is performed by a program that uses a formula or algorithm to determine how to compress or decompress data.

The net result of compression is less data. The distance we can send data is often gaited by the amount of space or memory that it takes up, so less information or data to send often means longer distances are achievable.

4.2.2 Multiplexers

Multiplexing is the process of simultaneously transmitting multiple signals over the same physical connection. There are two common types of multiplexing used for fiber optic connections:

- Time Division Multiplexing (TDM)
- Wavelength Division Multiplexing (WDM)
- Dense Wavelength Division Multiplexing (DWDM)

Time-Division Multiplexing

Time-division multiplexing (TDM) was created by the telecommunication industry to maximize the amount of effective traffic that could be carried over an existing fibre network. Previously in the telephone industry, every phone call needed its own discrete physical link. This system was obviously very costly to operate and limited for growth. The multiplexing facility enabled many phone circuits to be sent over a single link. In summary TDM enables many signals to be concentrated into a single fibre and all within the same wavelength.

TDM can be thought of as highway traffic. Traffic that needs to traverse from one city to another will start out on minor routes, it will then make it to a main (single lane) highway where it will join into a slot between other vehicles. The traffic will be controlled so that it is fair to all minor routes. Once the vehicles arrive at their destination they will be taken off to minor routes again.

This method is used in synchronous TDM devices. It increases the capacity of the fibre by reducing time slots into smaller intervals. This enables more bits in the same fibre which increases the effective bandwidth of the fibre. Within the TDM the input sources are multiplexed in a fair time share manner. Each signal will therefore have a packet space allocated to it whether it has signal input to send or not. This can be inefficient because there can be unused packets within the frame. Some protocols will reduce the effect of this by keeping data flowing into the channel. This is more likely in an asynchronous architecture.

Figure 4-1 shows a TDM and its method of combining several slower speed data streams into a single high speed data stream. Data from multiple sources is broken into portions (bits or bit groups) and these portions are transmitted in a defined sequence. Each of the input data streams then becomes a “time slice” in the output stream. The transmission order must be maintained so that the input streams can be reassembled at the destination.

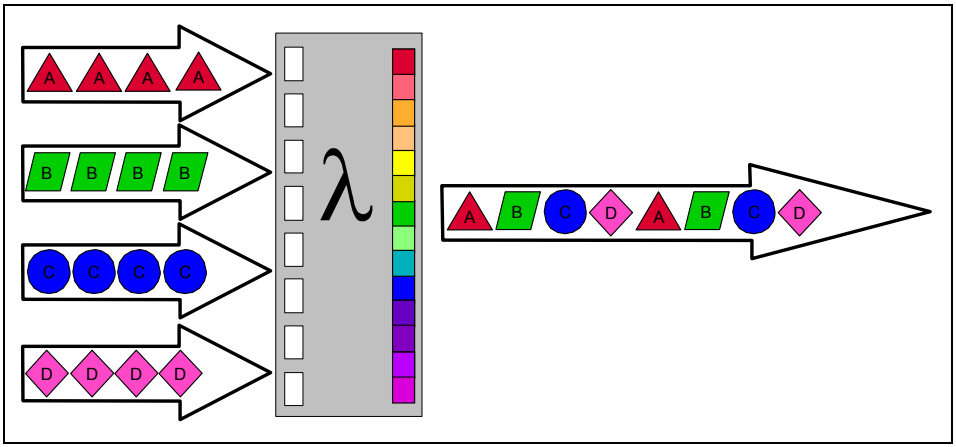


Figure 4-1 Time Division Multiplexer concepts

Wavelength Division Multiplexing

Wavelength Division Multiplexing (WDM) differs from TDM in that it does not use time to multiplex on — it uses the many wavelengths of light to multiplex on. WDM receives incoming optical signals from many sources (devices) which it converts to electrical signals, it then assigns them a specific wavelength (or lambdas or λ) of light and retransmits them on that wavelength. This method relies on the large number of wavelengths available within the light spectrum. You can think about WDM as though each channel is a different color of light; several channels then make up a “rainbow.” In summary WDM enables many signals to be concentrated into a single fibre all being sent at different wavelengths.

WDM allows the simultaneous transmission of a small number of data streams over the same physical fiber, each using a different optical wavelength. The advantages of WDM over TDM are that transmission order does not need to be maintained and that the information streams can use different protocols and bit rates. The WDM is sometimes described as the coarse wave division multiplexer.

Figure 4-2 shows each of the input signals coming into the WDM being multiplexed at different wavelengths on the output.

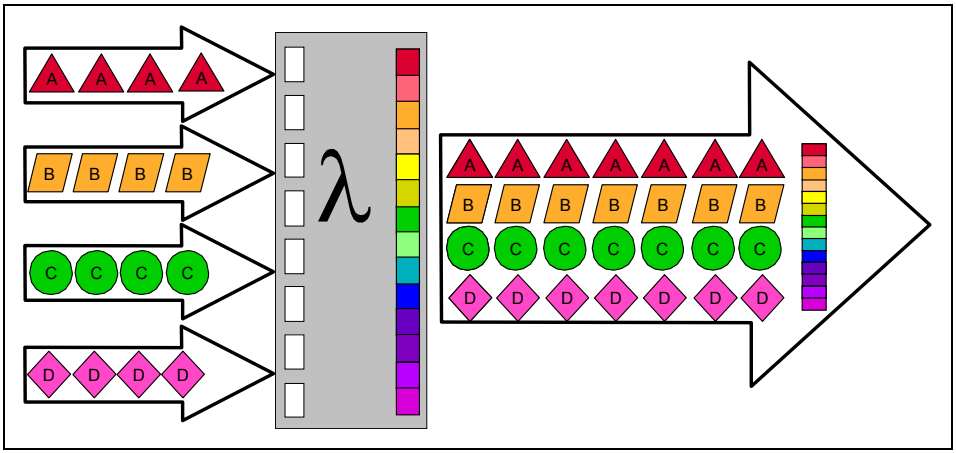


Figure 4-2 Wave Division Multiplexer concepts

This output is all multiplexed into the fiber. At the other end of the fiber, the signals are demultiplexed by the receiving WDM. This differs from the TDM method where time is displaced. Within the WDM environment each signal has its own wavelength. This means that they can all co-exist within the same fibre at the same time. This results in all channels having full bandwidth.

Dense Wave Division Multiplexer

Dense Wave Division Multiplexer (DWDM) uses the same design principles as the WDM, it can simply handle a much larger number of wavelengths. By reducing the spacing of these wavelengths more channels can be accommodated. As each channel maps to its own individual wavelength, the more wavelengths we have, the greater the total capacity or bandwidth of the device. We currently see DWDM devices that are capable of driving 256 discrete wavelengths along one single mode fibre.

DWDM is an approach to opening up the conventional optical fiber bandwidth by breaking it up into many channels, each at a different optical wavelength (a different color of light). Each wavelength can carry a signal at any bit rate less than an upper limit defined by the electronics, typically up to several gigabits per second. DWDM has all the advantages of WDM but with the added benefit of supporting far more independent transmissions over the same fiber. Due to the nature of these boxes, they are often considered transparent to protocol and bit rate.

Note: Some DWDM solutions use TDM within a channel (wavelength) for some low data rate protocols. This allows more efficient use of the available bandwidth.

The business drivers for DWDM are clear and deliver:

- **Bandwidth**

  If you currently have a pair of fibre installed that you are using for a single channel then by employing DWDM's that could bring you 255 extra channels per fibre.

- **Protocol independence**

  The DWDM is deployed as part of the physical layer. It is therefore independent of protocol, simply passing signal information in the format it is received. Examples of the protocols it can support are ATM, Gigabit Ethernet, ESCON, and Fibre Channel.

- **Growth on demand**

  A DWDM solution can preserve investment in already deployed fibre infrastructure, and be easily expanded to meet growing capacity.

- **Speed to market**

  Once a DWDM solution is deployed it can be used quickly and efficiently for new application, transparent of application and protocol. This enhances your ability to react to platform changes at an enterprise level. This can never be understated in the new e-business world.

DWDM summary

Figure 4-3 shows a generic overview of the components within a DWDM. The incoming signals form inputs are varied, we have simply shown SAN, ESCON and other protocols. Other Protocols could include, Gigabit Ethernet, SONET (OC-3, OC-12, OC-48), SDH (STM-1, STM-4, STM-16), Fibre Channel (1 Gbps), ESCON, FICON, and more.

These are often vendor specific as to what is supported and should be checked at the product level. We show the translation within the Optical to Electrical to Optical converter or Transponder and this takes the input signal on a specific wavelength and converts it to electrical signal which is then re-modulated on the new frequency that it will use during transit within the dark fibre media. These new wavelengths should adhere to the ITU-T grid, however different vendors will often use different channels from this grid and may even skip channel.

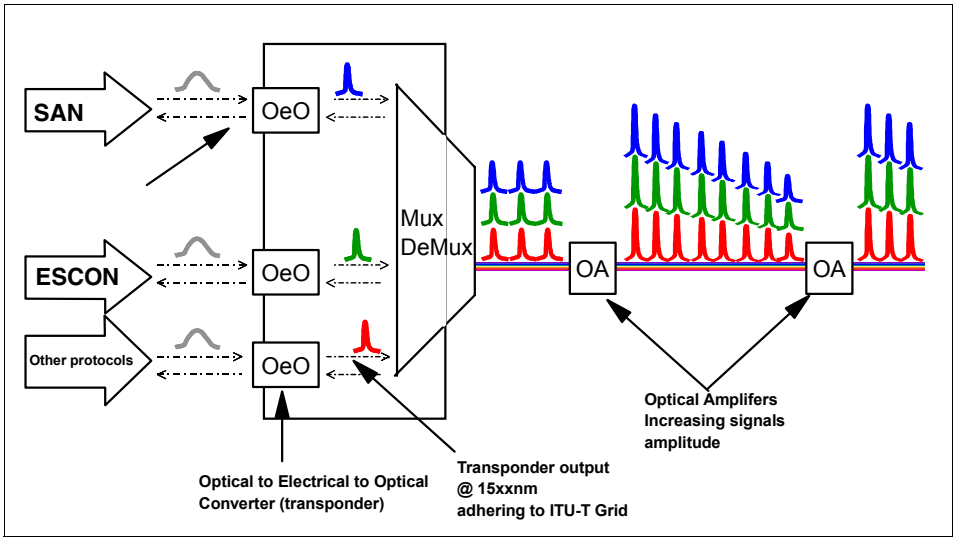


Figure 4-3 Generic example of DWDM architecture

Some of the differentiations within the DWDM market place are:

- Optical filters
  - Thin-film
  - Laser-welded
  - Fibre Bragg Grating
  - Planar-waveguide
- Channel expansion methods
  - Interleaving
  - Reduced channel sizes greater bandwidth
- Modulation techniques
  - Direct
  - Indirect

Many of these features have price and performance trade offs, however the comparison of these components within the DWDM industry is beyond the scope of this redbook.

4.2.3 Multiplexers and demultiplexers

The DWDM needs internal components that are capable of taking the many input signals and aligning them to the wavelength that they occupy within the fibre during transfer. This work is performed by the multiplexer, it does it by taking the signal wavelengths from the input fibres and converges them into one beam of light that is comprised of many wavelengths. This band of light then gets sent across the fibre to the receiving DWDM. At the receiving end, the opposite operation is performed, the signal is taken and split out at an optical level. This is then sent to the appropriate receiving photo detector. DWDM architecture components include DWDM filter modules, transmitters, receivers, (DWDM-capable) optical amplifiers, integrated optoelectronics, tunable filters used to add or drop specific frequencies.

In the topics that follow we will describe some of the internal components.

Lasers

There are two types of light emitting devices that are used in optical transmission, light-emitting diodes (LEDs) and laser diodes. The LED's are typically used for slower designs, and are often found in multimode implementations with speeds up to 1 Gb/s. LED's are relatively inexpensive devices. Lasers diodes are more expensive but lend themselves better to single mode devices. Two types of lasers diodes are widely used, monolithic Fabry-Perot lasers, and distributed feedback (DFB) lasers. The latter type is particularly well suited for DWDM applications, as it emits a nearly monochromatic light, is capable of high speeds, has a favorable signal-to-noise ratio, and has superior linearity. DFB lasers also have center frequencies in the region around 1310 nm, and from 1520 to 1565 nm. The latter wavelength range is compatible with EDFAs.

There are many other types and subtypes of lasers. Narrow spectrum tunable lasers are available, but their tuning range is limited to approximately 100-200 GHz. Under development are wider spectrum tunable lasers which will be important in dynamically switched optical networks. Cooled DFB lasers are available in precisely selected wavelengths.

Photo detectors

The photo detector is necessary to recover the signals transmitted at different wavelengths on the fiber. Photo detectors are wideband devices and cannot identify which band it is detecting. This means that the optical signals have to be demultiplexed before they reach the detector. The industry today tends to use two types of photodetector, the positive-intrinsic-negative (PIN) photodiode and the avalanche photodiode (APD). PIN photodiodes work in a similar fashion to LED's except in reverse — light is absorbed rather than emitted and photons are

converted to electrons which are transmitted as electrical signals. APDs are similar to PIN photodiodes, however they amplify the signal. This result in one photon releasing many electrons (that is to say, one-to-many). While PINs are cheaper and more reliable APDs have more sensitivity and accuracy.

In Figure 4-4 we show a multiplexer to demultiplexer.

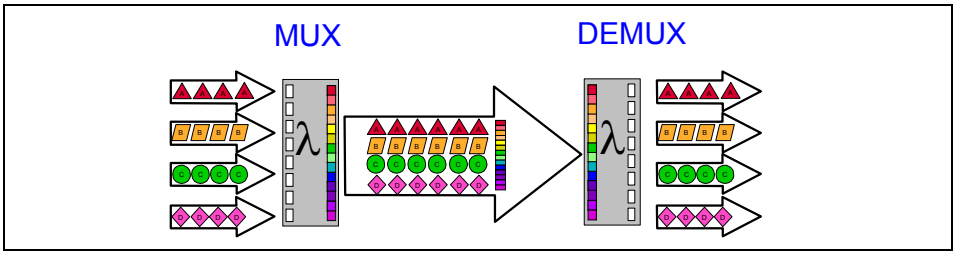


Figure 4-4 Multiplexer to demultiplexer

Different wavelengths are utilized to send traffic in opposing directions on the same fibre. This is achieved by using different wavelengths for each direction. In these instances there is a mux/demux function in the device at both ends of the fibre. This is shown in Figure 4-5.

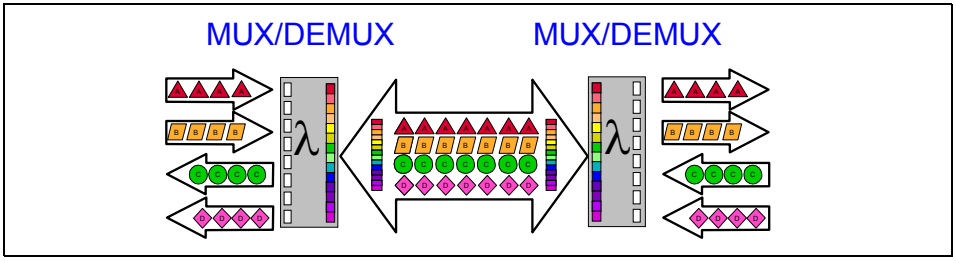


Figure 4-5 Both multiplexer and demultiplexer

Optical amplifiers can also be used to boost signal power after multiplexing or before the demultiplexer function.

4.2.4 Optical add/drop multiplexers

Another component that offers functionality in this arena is the optical add/drop multiplexers (OADM). These devices can be used at interim points between DWDM mux/demux units to split off or to inject signal wavelengths from or into the fibre. These units can be static devices or active.

The static first generation of OADMs have been built and configured to strip off (drop) or inject (add) specific wavelengths. The dynamic second generation of OADMs can be dynamically reconfigured to add or drop specific wavelengths to or from the wavelengths present in the fibre. Wavelengths that are not to be added or dropped from the traffic continue unchanged and are sometimes referred to as express channels for the purpose of that OADM.

In Figure 4-6 we show light being stripped out.

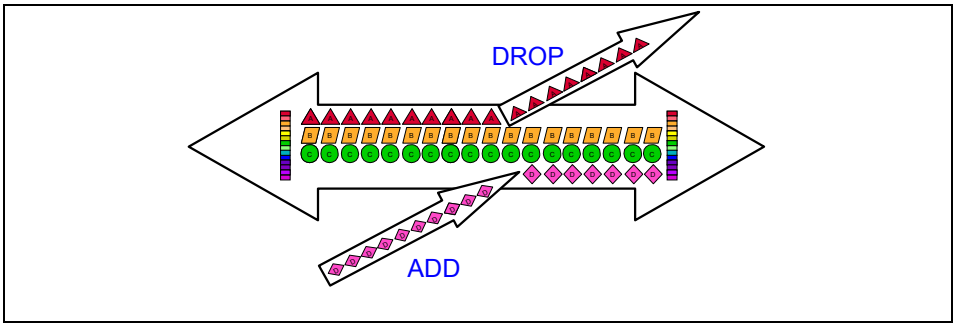


Figure 4-6 Light dropped and added

An OADM is used at intermediate stations. It removes (drops) a channel from a combined DWDM signal or adds a channel to a combined DWDM signal without interfering with the other channels on the fiber. After a channel has been dropped, the wavelength then becomes available to be reused by a different signal.

Figure 4-7 shows an OADM multiplexer/demultiplexer device. It has a crystal of transparent material with parallel sides on which a dielectric filter is deposited. The filter allows a single wavelength to be transmitted, reflecting all others. Therefore, a ray of light entering the device through a Graded Index (GRIN) lens will have one wavelength separated or demultiplexed from it. The device will operate in reverse as a DWDM multiplexer.

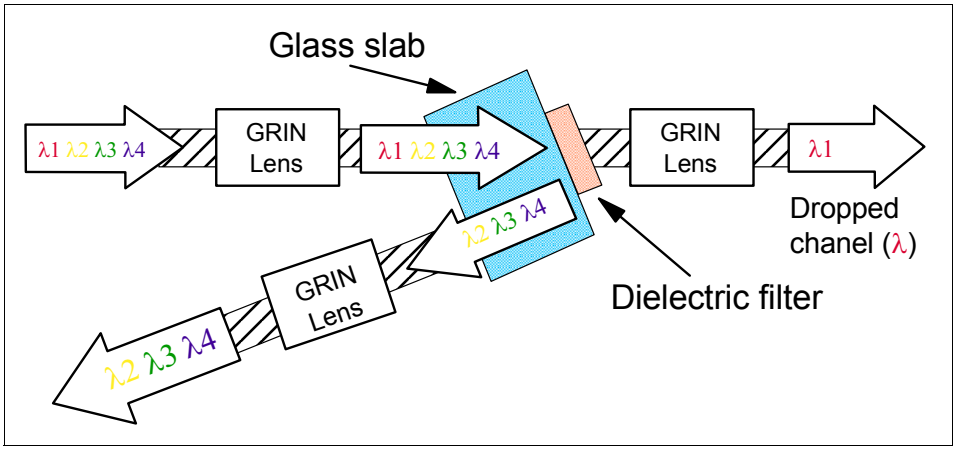


Figure 4-7 Example of OADM using dielectric filter

4.2.5 Optical amplifiers

Due to attenuation, the distance the signal on a fibre can propagate without loss of integrity is limited. We overcome this by using optical amplifiers to increase the signal strength. In the past we needed a repeater for each channel or signal transmitted, however this is now not true. The optical amplifier will amplify all the signals (wavelengths) as light and they will not need to be converted to from optical to electrical to optical in this process. Typically signals travel for up to 120 km between amplifiers, and longer distances are attainable, but the signal must be regenerated.

4.2.6 Regenerative repeaters

A regenerative repeater is a device that regenerates optical signals by converting incoming optical pulses to electrical pulses. It will clean up the electrical signal to eliminate noise, and reconvert them to optical pulses for output. This gives the ability to extend over very long distances.

4.3 DWDM topologies

DWDM can be implemented in more than one way and we describe these topologies:

- Point-to-point
- Linear
- Ring

– Hubbed ring
– Mesh ring

4.3.1 Point-to-point

This is the simplest of the three implementations. A point-to-point topology is a connection between two DWDM connections across a pair of single fibres. This is implemented with two Fibre Channels, one of which is considered an east link, and the other a west link.

Figure 4-8 shows a simple point-to-point connection.

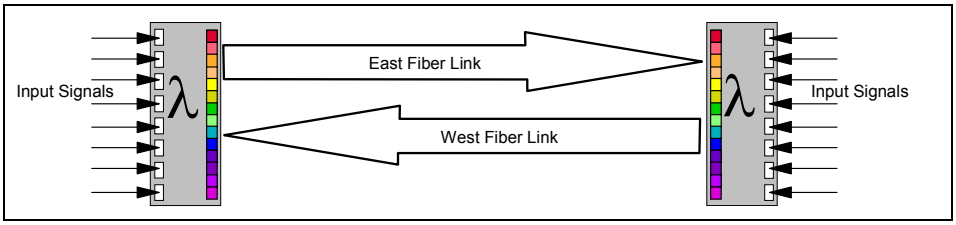


Figure 4-8 Point-to-point topology

4.3.2 Linear

A linear topology is a logical progression from the point-to-point architecture. It is a connection between DWDMs that are set out in a linear fashion. This is implemented with two Fibre Channels with one considered an east link and one a west link, between each DWDM station.

Figure 4-9 shows a simple linear connection.

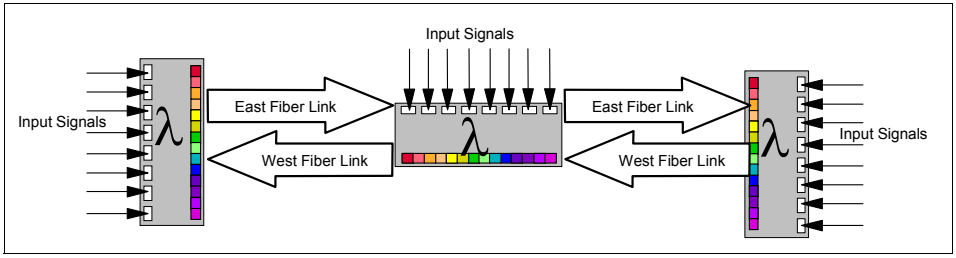


Figure 4-9 Linear topology between three locations

4.3.3 Ring

This topology is typically implemented where many geographically dispersed locations need to be connected. We will show a ring here with four points of presence. This solution could be implemented with only one or more DWDMs or could comprise of many components including OADM devices and Hubs. Channels can be dropped and added at one or more nodes on a ring. Rings have many common applications, including providing extended access to SANs where increasing the capacity of existing fiber is desirable.

Hubbed ring

A hubbed ring is composed of a hub node and two or more add/drop or satellite nodes. All channels on the ring originate and terminate on the hub node. At the add/drop node certain channels are terminated (dropped and added back) while the channels that are not being dropped (express channels) are passed through optically, without being electrically regenerated. This is shown in Figure 4-10.

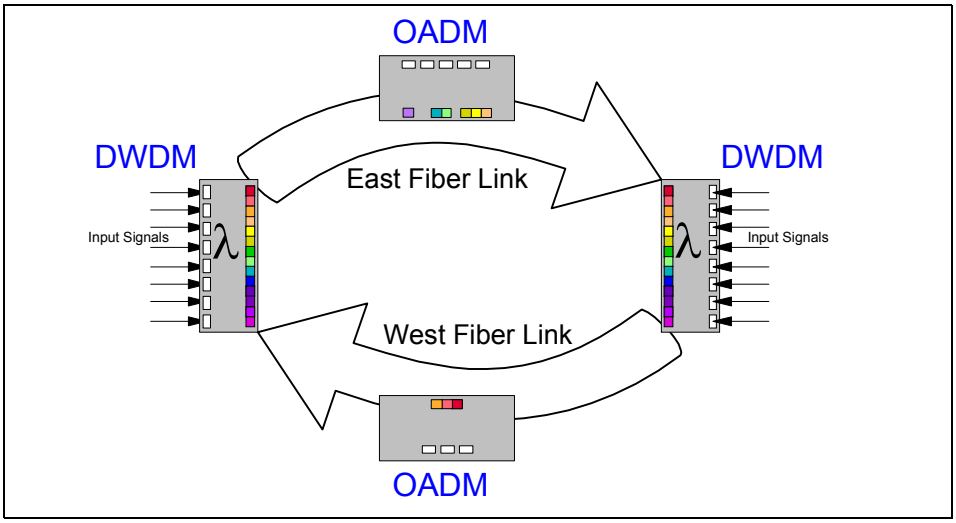


Figure 4-10 Ring topology using two DWDM and two OADM

Meshed ring

A meshed ring is a physical ring that has the logical characteristics of a mesh. While traffic travels on a physical ring, the logical connections between individual nodes are meshed.

Figure 4-11 shows a ring topology, and logical connections between some of the nodes can be thought of as meshed.

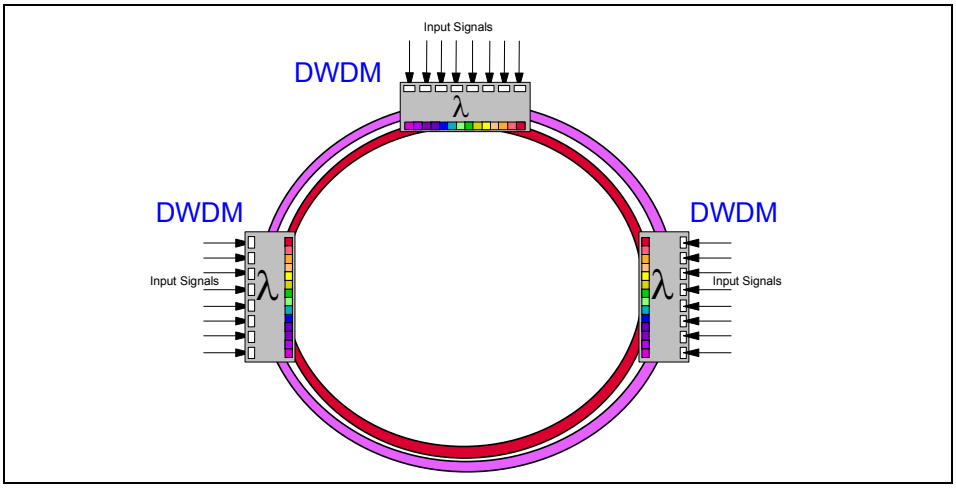


Figure 4-11 Ring topology with three DWDM

East and west

DWDM's are often composed of shelves each of which operate in a specified wavelength band, which is determined by the wavelength band of the optical modules installed in the shelf.

Figure 4-12 shows a DWDM shelf. Each shelf is divided into two parts, the west side and the east side. The two sides must have the same wavelength band.

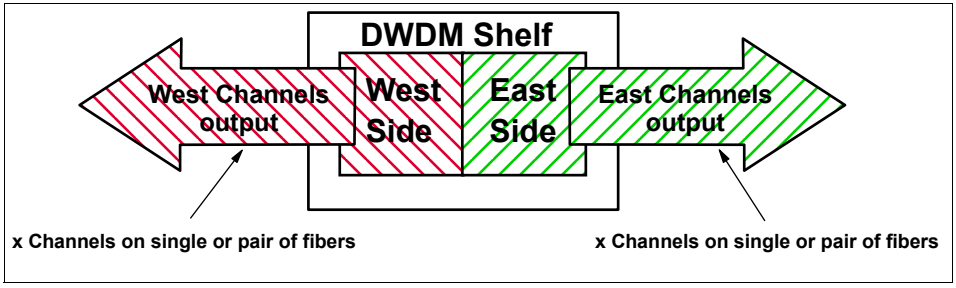


Figure 4-12 DWDM module showing east and west

A side can have many channels, each operating on a different wavelength within the shelf's wavelength band. Each west side channel has the same wavelength as the corresponding east side channel.

Figure 4-13 shows a DWDM comprising of four shelves, each with its own band.

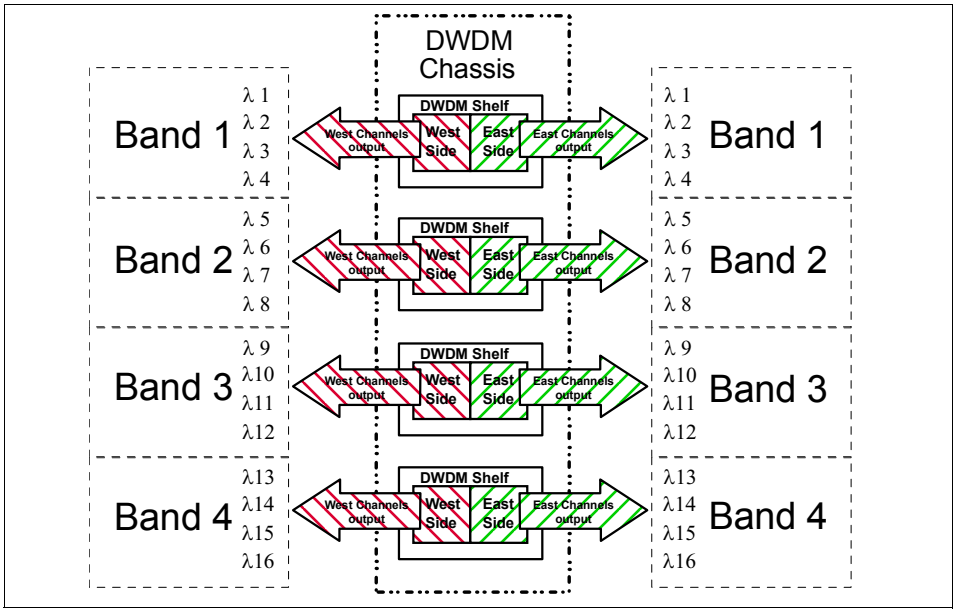


Figure 4-13 East and west must have same wavelengths within the same band

We also introduce the concept of those bands having many wavelengths. In this example we have four wavelengths (λ) per band. As described above, both east and west sides of the optical module must operate on the same wavelengths.

Protection

Protection within the DWDM devices comes in two flavors:

- **Internal protection**

  Internal protection is the ability to configure redundant components internal to the DWDM chassis. This protection is often at component level. Components that will often have internal protection include CPU modules, power supplies, fan assemblies.

- **External protection**

  External protection is the ability to configure redundant fibres external to the DWDM device. This protection is aimed at surviving fibre failure. Implementing protected circuits will often halve the number of channels that are available to you. This protection can be implemented by deploying redundant fibres. Many DWDMs support this option and will automatically sense and reroute their signal upon loss of the primary link.

  Redundant paths can be implemented using line protection or path protection.

- Line Protection is when you have two lasers transmitting — one transmits east, one transmits west.
- Path Protection is when you have one laser and the signal is optically split to the east and the west.

4.4 Factors that will affect distance

Fiber channel distances depend on many factors and include:

- Type of laser used, longwave or shortwave
- Type of fiber optic cable, multi-mode or single-mode
- Quality of the cabling infrastructure in terms of dB loss — connectors, cables, and even bends and loops in the cable can result in dB signal loss
- Native shortwave FC transmitters have a maximum distance of 500 m with 50 micron diameter, multi-mode, optical fiber
- Although a 62.5 micron, multi-mode fiber can be used, the larger core diameter has a greater dB loss and maximum distances are shortened to 300 meters.
- Native longwave FC transmitters have a maximum distance of 10 km when used with 9 micron diameter single-mode optical fiber.

Link extenders will provide a signal boost that can potentially extend distances up to about 100 km. These link extenders simply act as a very big, fast pipe. Data transfer speeds over link extenders depend on the number of buffer credits and efficiency of buffer credit management in the FC nodes at either end of this fast pipe. Buffer credits are designed into the hardware for each FC port.

FC provides flow control that protects against collisions. This is extremely important for storage devices, which do not handle dropped or out-of-sequence records. When two FC ports begin a conversation they exchange information on their buffer capacities. An FC port will send only the number of buffer frames for which the receiving port has given credit. This not only avoids overruns, but also provides a way to maintain performance over distance by filling the “pipe” with in-flight frames or buffers.

The maximum distance that can be achieved at full performance is dependent on the capabilities of the FC node that is attached at either end of the link extenders. This is very vendor specific. There should be a match between the buffer credit capability of the nodes at either end of the extenders. A host bus adapter (HBA) with a buffer credit of 64 communicating with a switch port with only eight buffer

credits would be able to read at full performance over greater distance than it would be able to write. This is because on the writes, the HBA can send a maximum of only eight buffers to the switch port, while on the reads, the switch can send up to 64 buffers to the HBA.

Until recently, a rule of thumb has been to allot one buffer credit for every 2km in order to maintain full performance.

4.4.1 Light or link budget

It is important to understand the link budget terminology. The decibel (dB) is a convenient way of expressing an amount of signal loss or gain within a system or the amount of loss or gain caused by some component of a system. When signal power is lost, you never lose a fixed amount of power. The rate at which you lose power is not linear. Instead you lose a portion of power — one half, one quarter, and so on. This makes it difficult to add up the lost power along a signal's path through the network if measuring signal loss in watts.

For example, a signal loses half its power through a bad connection, then it loses another quarter of its power later on through a bent cable. You cannot add 1/2 plus 1/4 to find the total loss. You must multiply 1/2 by 1/4. This makes calculating large network dB loss time-consuming and difficult.

Decibels, though, are logarithmic. This allows us to easily calculate the total loss/gain characteristics of a system just by adding them up! Keep in mind that they scale logarithmically. If your signal gains 3dB, the signal doubles in power. If your signal loses 3dB, the signal halves in power.

It is important to remember that the decibel is a ratio of signal powers. You must have a reference point. For example, you can say, "there is a 5dB drop over that connection." But you cannot say, "the signal is 5dB at the connection." A decibel is not a measure of signal strength, but a measure of signal power loss or gain. A decibel milliwatt (dBm) is a measure of signal strength. People often confuse dBm with dB. Do not fall into this trap! A dBm is the signal power in relation to one milliwatt. A signal power of 0 dBm is one milliwatt, a signal power of 3 dBm is 2 milliwatts, 6 dBm is 4 milliwatts, and so on. Also, do not be misled by minus signs. It has nothing to do with signal direction. The more negative the dBm goes, the closer the power level gets to zero.

For example, -3 dBm is 0.5 milliwatts, -6 dBm is 0.25 milliwatts, and -9 dBm is 0.125 milliwatts. So a signal of -30 dBm is very weak.

4.4.2 Buffer credits

Buffer credits within the switches and directors have a large part to play in the distance equation. The buffer credits in the sending and receiving nodes heavily influence the throughput that is attained within the Fibre Channel. Fibre Channel architecture is based on a flow control that ensures a constant stream of data to fill the available pipe. A rule-of-thumb says that to maintain acceptable performance one buffer credit is required for every 2 km distance covered.

Buffers

Ports need memory, or “buffers”, to temporarily store frames as they arrive and until they are assembled in sequence, and delivered to the upper layer protocol. The number of buffers, that is the number of frames a port can store, is called its “Buffer Credit”.

BB_Credit

During login, N_Ports and F_Ports at both ends of a link establish its Buffer to Buffer Credit (BB_Credit).

EE_Credit

In the same way during login all N_Ports establish End to End Credit (EE_Credit) with each other.

During data transmission a port should not send more frames than the buffer of the receiving port can handle before getting an indication from the receiving port that it has processed a previously sent frame. Two counters are used for that. BB_Credit_CNT and EE_Credit_CNT, and both are initialized to 0 during login.

Each time a port sends a frame it increments BB_Credit_CNT and EE_Credit_CNT by 1. When it receives R_RDY from the adjacent port it decrements BB_Credit_CNT by 1, when it receives ACK from the destination port it decrements EE_Credit_CNT by 1. Should at any time BB_Credit_CNT become equal to the BB_Credit or EE_Credit_CNT equal to the EE_Credit of the receiving port, the transmitting port has to stop sending frames until the respective count is decremented.

The previous statements are true for Class 2 service. Class 1 is a dedicated connection, so it does not need to care about BB_Credit and only EE_Credit is used (EE Flow Control). Class 3 on the other hand is an unacknowledged service, so it only uses BB_Credit (BB Flow Control), but the mechanism is the same on all cases.

Here we can see the importance that the number of buffers has in overall performance. We need enough buffers to make sure the transmitting port can continue sending frames without stopping in order to use the full bandwidth.

This is particularly true with distance. At 1 Gb/s a frame occupies 4 km of fiber. In a 100 km link we can send 25 frames before the first one reaches destination. We need an ACK (acknowledgment) back to start replenishing EE_Credit. We will be able to send another 25 before we receive the first ACK. We need at least 50 buffers to allow for non stop transmission at 100 km distance.

4.4.3 Quality of fiber

The optical properties of the fibre will influence the distance that can be supported. There is a decrease in signal strength along a fiber. As the signal travels over the fibre, it is attenuated, and this is caused by both absorption and scattering; and this is usually expressed in decibels per kilometer (dB/km). Some early deployed fibre was designed to support the telephone network and this is sometimes of insufficient specification for the new multiplexed environments. If you are being supplied dark fibre by another party you will normally specify that they must not allow more than xdB loss in total.

4.4.4 Cable types

Light is sent into the fibre by the emitter and it propagates in optical pulses and is detected by the detector. In Figure 4-14 we show this.

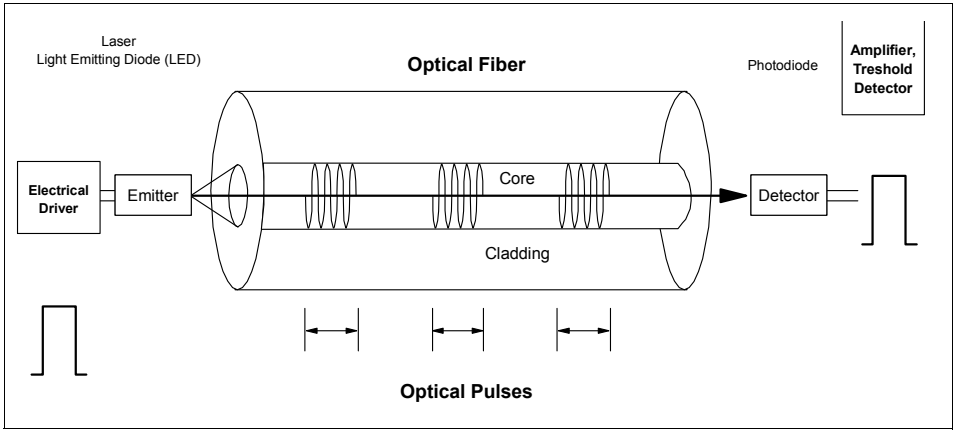


Figure 4-14 Light propagation through fiber

Fibre cables within the SAN environment break down into two categories:

- Single mode fibre
- Multimode fibre

Note: DWDM applications use single mode fibre on the transport or dark fibre side.

In most cases, it is impossible to distinguish between single-mode and multi-mode fiber with the naked eye (unless the manufacturer follows the color coding schemes specified by the Fibre Channel physical layer working subcommittee of orange for multi-mode and yellow for single-mode). There may be no difference in outward appearance, only in core size.

Both fiber types act as a transmission medium for light, but they operate in different ways, have different characteristics, and serve different applications. We show the light propagation characteristic of single mode fibre (Figure 4-15) and in multi mode fibre (Figure 4-16).

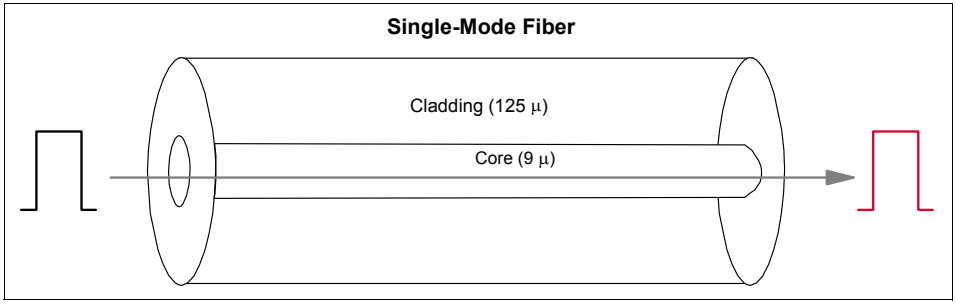


Figure 4-15 Light propagation in single-mode fiber

Single-mode (SM) fiber allows for only one pathway, or mode, of light to travel within the fiber. The core size is typically 8.3 µm. Single-mode fibers are used in applications where low signal loss and high data rates are required, such as on long spans between two system or network devices, where repeater/amplifier spacing needs to be maximized.

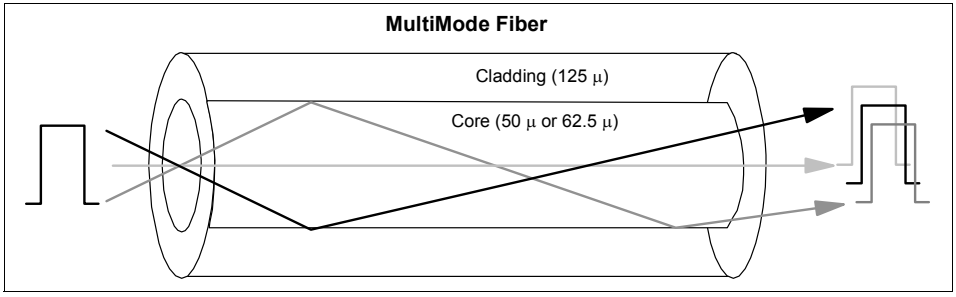


Figure 4-16 Light propagation in multi-mode fiber

Multi-mode (MM) fiber allows more than one mode of light. Common MM core sizes are 50 µm and 62.5 µm. Multi-mode fiber is better suited for shorter distance applications. Where costly electronics are heavily concentrated, the primary cost of the system does not lie with the cable. In such a case, MM fiber is more economical because it can be used with inexpensive connectors and laser devices, thereby reducing the total system cost. This makes multi-mode fiber the ideal choice for short distance under 500 meters from transmitter to receiver (or the reverse).

4.4.5 Droop

Droop will effect performance and it is experienced when a critical distance is exceeded. Factors that affect it are the laws of physics, the speed of light in fibre, the link data rate and the available buffering within the sending and receiving devices. Droop begins when a links distance reaches a point where the time the light takes to make one round trip on the link equals the time it takes to transmit the number of bytes that fit in the receiver's buffer.

In Figure 4-17 we show an ESCON estimate of how the effective data rate decreases as the path length increases. At a distance of 9 km, performance begins to decrease precipitously. This data point is referred to as the distance data rate droop point.

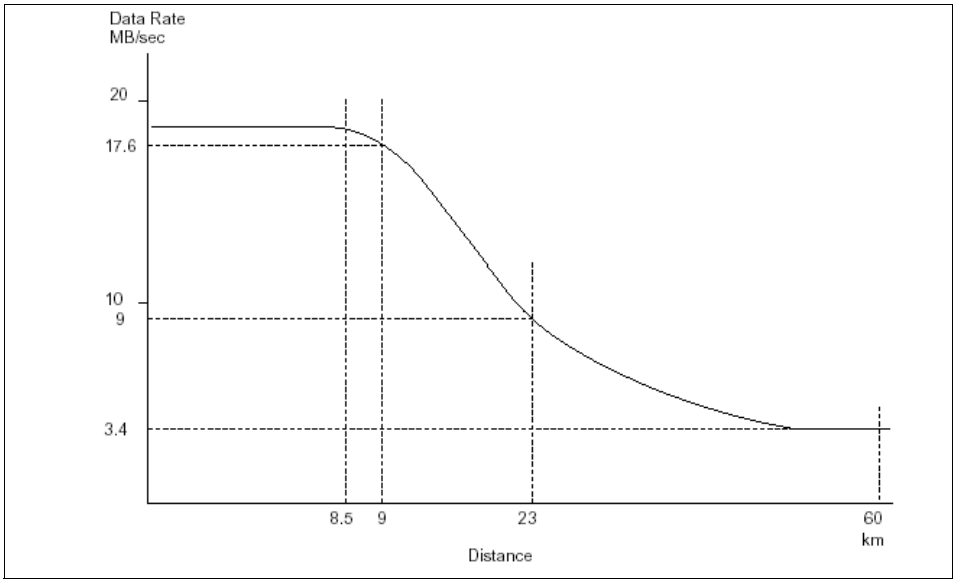


Figure 4-17 ESCON droop example

4.4.6 Latency

Within the SAN arena, longer distances introduce other factors to consider in the SAN design, one of which is latency. A key contributor to latency is distance, latency increases with distance since the time for the signal to travel the longer links has to be added to the normal latency introduced by switches and/or directors. This discussion leads to time-outs and buffer credits, and these should allow for increased travel times. Latency may or may not be a problem, it is that ability of the fabrics components to tolerate it that will cause issues. The fabric components, applications and operating systems ability to tolerate it will determine its effects.

4.4.7 Sizing

Sizing will be site specific based on the number of channels that need to be sent from location to location. The number of fibres needed will be influenced by whether protection is needed and the topology deployed.

4.4.8 Hops

The hop count as such is not increased by the DWDM architecture, the DWDM components are transparent to the number of hops. The hop count limit within a fabric is set by the fabric devices (switch or director) operating system and it is used to derive a frame hold-time value for each fabric device. This hold-time value is the maximum amount of time that a frame can be held in a switch before it is dropped or the fabric is busy condition is returned. For example, a frame would be held if its destination port is not available. The holdtime is derived from a formula using the error detect time-out value and the resource allocation time-out value.

The discussion on these fabric values is beyond the scope of this book, however further information can be found in *Designing an IBM Storage Area Network*, SG24-5758-00*,* or *IBM SAN Survival Guide*, SG24-6143-00.

If these times become excessive, the fabric will experience undesirable time outs. It is considered that every extra hop adds about 1.2 microseconds latency to the transmission.

We are therefore saying that the number of hops that can be achieved in a SAN fabric is a function of time. DWDM solutions introduce long fibre distances and increased latencies, and this may well affect the number of hops that can be supported. The amount of data that can be in transit is related to buffer credits (we describe these in “Buffer credits” on page 202). As a working rule of thumb we can say that we should allow 1 buffer credit per 2 km of fibre.

4.4.9 Physical location of repeaters

In the past it has been necessary to deploy repeaters at regular intervals to attain long distances. These would have been in the order of 40 km in many examples. DWDM's make it possible to extend over longer distances with less repeaters or optical amplifiers (distances in the order of 120 km). This means that you do not need convenient points en route to accommodate repeater or optical amplifying kits. Remember it may not be as simple as setting up a repeater every 40 km. It will depend on many factors.

4.4.10 Standards

The International Telecommunication Union (ITU), headquartered in Geneva, Switzerland is an international organization within which governments and the private sector coordinate global telecom networks and services. The ITU Telecommunication Standardization Sector (ITU-T) is one of the three Sectors of the International Telecommunication Union and its mission is to ensure an efficient and on-time production of high quality standards covering all fields of telecommunications. The ITU has standards for communications equipment based on the specific wavelength that lasers operate at (the ITU recommendations for optical interfaces for multichannel systems with optical amplifiers G.692 (10/98)). Their Web site can be found at:

http://www.itu.int

4.4.11 Terminology

In this section we will introduce some of the commonly encountered terminology.

Nanometer

The word "nano" means 10^{-9}, so a nanometer is one billionth of a meter.

A nanometer is a unit of spatial measurement that is 10^{-9} meter, or one millionth of a millimeter. It is commonly used in nanotechnology which is the building of extremely small machines.

The wavelengths of light are measured in nanometers.

Wavelength

The technical definition of a wavelength is the distance measured in the direction of propagation of a light wave between two successive points in the wave that are characterized by the same phase of oscillation. This is expressed as nanometer (nm), where one nm is the equivalent of one one-millionth of a millimeter between the two successive points.

In non-technical terms, lower wavelength values have a longer range, are brighter, and generally have a larger spot size than a higher wavelength value.

For example, a 635 nm laser product is brighter with greater range than a 650 nm laser product when both have the same the output power (as defined below). Combined with the appropriate output power, wavelength is an important consideration when using a laser product outdoors or in brightly lit interior environments.

Bandwidth

Bandwidth (the width of a band of electromagnetic frequencies) is used to mean how fast data flows on a given transmission path and, somewhat more technically, the width of the range of frequencies that a signal occupies on a given transmission medium. Any digital or analog signal has a bandwidth. In optical networks bandwidth is defined as the range of frequencies within which a fiber optic waveguide or terminal device can transmit data or information.

Channel

In telecommunications in general, a channel is a separate path through which signals can flow. In optical fiber transmission using dense wavelength-division multiplexing (DWDM), a channel is a separate wavelength of light within a combined, multiplexed light stream.

In IBM mainframe systems, a channel is a high bandwidth connection between a processor and other processors, workstations, printers, and storage devices within a relatively close proximity. It is also called a local connection as opposed to a remote (or telecommunication) connection.

DWDM

Dense wavelength division multiplexing (DWDM) is a technology that puts data from different sources together on an optical fiber, with each signal carried at the same time on its own separate light wavelength. Using DWDM, up to 80 (and theoretically more) separate wavelengths or channels of data can be multiplexed into a light stream transmitted on a single optical fiber. Each channel carries a time division multiplexed (TDM) signal. In a system with each channel carrying 2.5 Gbps (billion bits per second), up to 200 billion bits can be delivered a second by the optical fiber. DWDM is also sometimes called wave division multiplexing (WDM).

Since each channel is demultiplexed at the end of the transmission back into the original source, different data formats being transmitted at different data rates can be transmitted together. Specifically, Internet (IP) data, Synchronous Optical Network data (SONET), and asynchronous transfer mode (ATM) data can all be traveling at the same time within the optical fiber.

4.4.12 Protocol definitions

In this section we describe some of the commonly encountered protocols.

ATM

ATM (asynchronous transfer mode) is a dedicated-connection switching technology that organizes digital data into 53-byte cell units and transmits them over a physical medium using digital signal technology. Individually, a cell is processed asynchronously relative to other related cells and is queued before being multiplexed over the transmission path. Because ATM is designed to be easily implemented by hardware (rather than software), faster processing and switch speeds are possible. The pre-specified bit rates are either 155.520 Mbps or 622.080 Mbps. Speeds on ATM networks can reach 10 Gbps. Along with Synchronous Optical Network (SONET) and several other technologies, ATM is a key component of broadband ISDN (BISDN).

Gigabit Ethernet

The Ethernet protocol is the worlds most popular LAN (local area networking) protocol. This standard has evolved from the original shared 10 megabit per second technology, developed in the 1970's, to the recently completed Gigabit Ethernet standard, (the first Gigabit Ethernet standard (802.3z) was ratified by the IEEE 802.3 Committee in 1998). Gigabit Ethernet is the newest version of Ethernet and it supports data transfer rates of 1 Gigabit (1,000 megabits) per second.

SONET (OC-3, OC-12, OC-48) and SDH (STM-1, STM-4, STM-16)

SONET and SDH are a set of related standards for synchronous data transmission over fiber optic networks. SONET is short for Synchronous Optical NETwork and SDH is an acronym for Synchronous Digital Hierarchy.

SONET is the United States version of the standard published by the American National Standards Institute (ANSI). SDH is the international version of the standard published by the International Telecommunications Union (ITU).

Fibre Channel

Fibre Channel is a technology for transmitting data between computer devices at a data rate of up to 2 Gbps, or one billion bits per second. A data rate of 10 Gbps has been proposed by the Fibre Channel Industry Association.

Fibre Channel is especially suited for connecting computer servers to shared storage devices and for interconnecting storage controllers and drives. Since Fibre Channel is three times as fast, it has begun to replace the Small Computer System Interface (SCSI) as the transmission interface between servers and clustered storage devices. Fibre channel is more flexible; devices can be as far

as ten kilometers (about six miles) apart if optical fiber is used as the physical medium. Optical fiber is not required for shorter distances, however, because Fibre Channel also works using coaxial cable and ordinary telephone twisted pair.

Fibre Channel offers point-to-point, switched, and loop interfaces. It is designed to interoperate with SCSI, the Internet Protocol (IP) and other protocols, but has been criticized for its lack of compatibility — primarily because (like in the early days of SCSI technology) manufacturers sometimes interpret specifications differently and vary their implementations. Standards for Fibre Channel are specified by the Fibre Channel Physical and Signalling standard, and the ANSI X3.230-1994, which is also ISO 14165-1.

ESCON

ESCON is a 200 Mbps unidirectional serial bit transmission protocol used to dynamically connect mainframes with their various control units. ESCON provides non blocking access through either point-to-point connections or high speed switches, called ESCON Directors. ESCON performance is seriously affected if the distance spanned is greater than approximately 8 km. For instance, measurements have shown that ESCON performance at 20 km is roughly 50 percent of maximum performance. Performance degradation continues as distance is further increased.

FICON

FICON is the next generation bidirectional channel protocol used to connect mainframes directly with control units or ESCON aggregation switches (ESCON Directors with a bridge card). FICON runs over Fibre Channel at a data rate of 1.062 Gbps. One of the main advantages of FICON is the lack of performance degradation over distance that is seen with ESCON. FICON can reach a distance of 100 km before experiencing any significant drop in data throughput.

5

SAN disciplines

One of the key elements of a successful SAN installation is the physical location of the equipment and the disciplines introduced to manage those elements. Typically a new SAN installation is born based on an individual requirement at the time. The new SAN usually starts small and simple, but will grow very rapidly. Disciplines which were not an issue with the small SAN become major management problems as the SAN develops. It is very difficult to introduce standards to an established SAN, so careful consideration at the conceptual phase will be rewarded.

In this chapter we look at some of the SAN disciplines that should be considered prior to implementing a SAN, and look at the potential effects of not implementing these disciplines.

We will start from the floor up and make general observations and recommendations along the way to building the SAN. In the following sections we consider the pre-planning activity and compare the pros and cons of some of the options.

Remember, simply connecting the SAN components together is not a challenge, but development of a decent SAN design is one.

© Copyright IBM Corp. 2004. All rights reserved.

5.1 Floor plan

In comparison to a traditional open server environment based on SCSI technology, with Fibre Channel Protocol (FCP) we are no longer faced with short distance limitations and are able to spread our SAN over thousands of kilometers. This has its benefits for things like Disaster Recovery, but the more you distribute the SAN, the higher the cost and management overhead.

5.1.1 SAN inventory

Prior to establishing a floor plan, it is good practice to establish a high level inventory list of the SAN components that already exist, and those that will be added to the SAN. This list, which can include logical and physical components, can be used to plan the quantity and location of the SAN fabric cabinets and will feed into a more detailed list that will help design the SAN layout.

The list should include the following:

- Server type (vendor, machine type and model number)
- Switch/director type (vendor, machine type and model number)
- Storage type (vendor, machine type and model number)
- Fibre Channel protocols that devices support and cannot support
- Device (server, storage, SAN components) names and description
- Distances between devices (maximum and minimum)
- Location of admin consoles or management servers
- Storage partitioning
- Location of SCSI drives (no more than 25 m away)
- Fabric names
- Zone names
- IP addresses
- Naming conventions employed
- Passwords and userids
- Current cabinet address
- Operating systems, maintenance level and firmware levels
- Quantity and type of adapters installed
- List of WWNs and WWPNs
- If devices will have single or multiple attachments in the SAN
- Cabling cabinets
- Labels for cables
- Cable routing mapped
- Current connections
- Current configurations

5.1.2 Cable types and cable routing

There are a number of different types of cable that can be used when designing a SAN. The type of cable and route it will take all need consideration. The following section details various types of cable and issues related to the cable route.

Distance

The Fibre Channel cabling environment has many similarities to telco environments and is a major expansion over scsi open systems environments. The increase in flexibility and adaptability in the placement of the electronic network components is similar to the LAN/WAN environment, and a significant improvement over previous data center storage solutions.

Single-mode or multi-mode

Every data communications fiber belongs to one of two categories:

- Single-mode
- Multi-mode

In most cases, it is impossible to distinguish between single-mode and multi-mode fiber with the naked eye unless the manufacturer follows the color coding schemes specified by the FC-PH (see 3.13.10, “List of evolved Fibre Channel standards” on page 141) working subcommittee (typically orange for multi-mode and yellow for single-mode). There may be no difference in outward appearance, only in core size. Both fiber-optic types act as a transmission medium for light, but they have different diameters and different demands for the spectral width of the light sources:

- **Single-mode (SM):** This mode, also called mono-mode fiber, or single-mode fiber, allows for only one pathway, or propagation mode, of light to travel within the fiber. The core size is typically 8.3 - 10 µm. SM fibers are used in applications where low signal loss and high data rates are required, such as on long spans between two system or network devices, where repeater/amplifier spacing needs to be maximized. SM fiber links use longwave laser at 1270-1300 nm wavelength.
- **Multi-mode (MM):** This mode, also called multi-mode fiber, allows more than one mode of light. Common MM core sizes are 50 µm and 62.5 µm. MM fiber links can either use a shortwave (SW) laser operating at 780-860 nm, or a longwave (LW) laser at 13270-1300 nm wavelength. The low-cost shortwave laser is based on the laser diode developed for the CD players and benefits from the high volume production with that market. That makes shortwave/MM- equipment more economical. MM fiber is therefore the ideal choice for short distance applications between Fibre Channel devices.

For the supported distances of 1 Gb/s and 2 Gb/s links, refer to 2.2.1, “Small Form Factor Pluggable Media” on page 20, and 2.2.2, “Gigabit Interface Converters” on page 23.

Host bus adapter (HBA) and switch vendors are expected to start offering 4Gbps Fibre Channel products in late 2004 or early 2005. This is expected to make 4Gbps an interim speed before the industry takes on the major transition to 10Gbps Fibre Channel. 4Gbps Fibre Channel utilizes the same technology as 2Gbps but is twice as fast. The higher fibre channel speeds will still utilize the same basic fiber cables, but distance limitations may change.

In Figure 5-1 we show the differences in single-mode and multi-mode fiber routes through the fiber-optic cable.

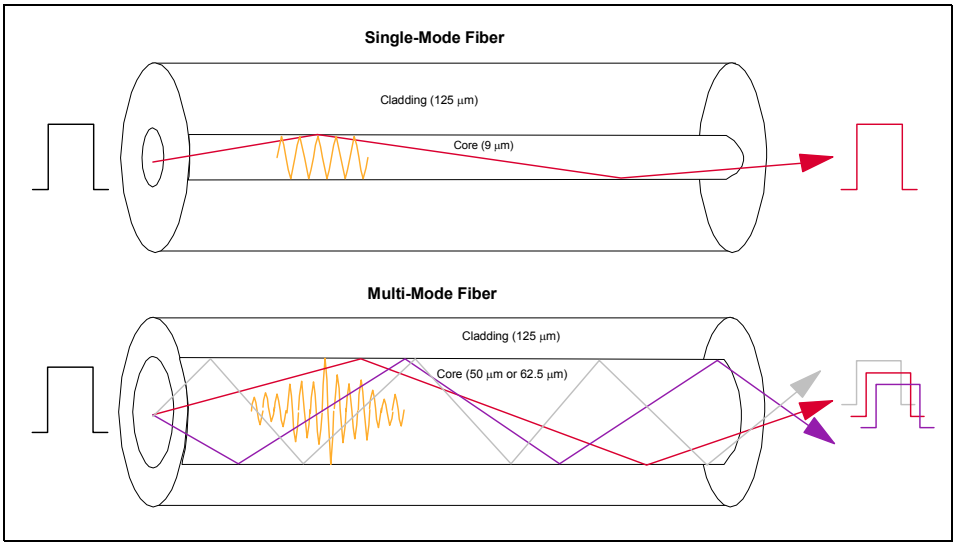


Figure 5-1 Mode differences through the fiber-optic cable

Propagation mode: The pathway of light is illustrative in defining a mode. According to electromagnetic wave theory, a mode consists of both an electric and a magnetic wave mode, which propagates through a waveguide. To transport a maximum of light, we need to have a total internal reflection on the boundary of core and cladding. With the total reflection, there comes a phase shift of the wave.

We look for modes that have the same wave amplitude and phase at each reflection to interfere constructively by wave superposition. With help of the mode equation we will find modes for a given electromagnetic wave (the light with its wavelength) propagated through a given waveguide (a fiber with its geometry). We call a waveguide *mono-mode* when only the lowest order bound mode (fundamental mode of that waveguide) can propagate.

There is exhaustive technical and scientific material about fiber and optics in 'Optics2001.com', the free Optical Community at the Web site:

http://www.optics2001.com/Optical-directory.php

Fiber optic cable

Fiber optic cable for telecommunications consists of three components:

- Core
- Cladding
- Coating

Core

The core is the central region of an optical fiber through which light is transmitted. In general, the telecommunications industry uses sizes from 8.3 µm to 62.5 µm. As already discussed, the standard telecommunications core sizes in use today are 8.3 (9) µm, 50 µm and 62.5 µm.

Note: Microns or micrometers (µm)? A micron is 0.0000394 (approximately 1/25,000th) of an inch, or one millionth of a meter. In industrial applications, the measurement is frequently used in precision machining. In the technology arena, however, microns are most often seen as a measurement for fiber-optic cable (which has a diameter expressed in microns), and a unit of measure in the production of microchips.

Micrometer is another name for a micron, but it is more commonly used for an instrument that measures microns in a wide variety of applications, from machine calibration to the apparent diameter of celestial objects.

Cladding

The diameter of the cladding surrounding each of these cores is 125 µm. Core sizes of 85 µm and 100 µm have been used in early applications, but are not typically used today. The core and cladding are manufactured together as a single piece of silica glass with slightly different compositions, and cannot be separated from one another.

Coating

The third section of an optical fiber is the outer protective coating. This coating is typically an ultraviolet (UV) light-cured acrylate applied during the manufacturing process to provide physical and environmental protection for the fiber. During the installation process, this coating is stripped away from the cladding to allow proper termination to an optical transmission system. The coating size can vary, but the standard sizes are 250 µm or 900 µm. The 250 µm coating takes less space in larger outdoor cables. The 900 µm coating is larger and more suitable for smaller indoor cables.

The 62.5 µm multi-mode fiber was included within the standard to accommodate older installations which had already implemented this type of fiber-optic cable. Due to the increased modal dispersion and the corresponding distance reduction of 62.5 µm multi-mode fiber, 50 µm multi-mode fiber is the preferred type for new installations. It is recommended to check with any SAN component vendor to see if 62.5 µm is supported.

For more details about fiber-optic cables, visit the “American National Standard for Telecommunications” Glossary at the Web site:

http://www.atis.org/tg2k/t1g2k.html

Structured and non-structured cables

In this topic we look at two types of cables: non-structured and structured.

Non-structured cables

Non-structured cables consist of a pair of optical fibers that provide two unidirectional serial bit transmission lines; they are commonly referred to as a jumper cable. Jumper cables are typically used for short links within the same room. They can be easily replaced if damaged, so they are most suited to connecting SAN components that may require regular cabling alterations.

Multi-jumper cables are available with more than one pair of fibres. They are typically used to connect more than one pair of Fibre Channel ports.

As individual cables can become easily tangled and difficult to locate, they are best avoided for longer under-floor runs.

Structured cables

Structured cables consist of multiple fiber optic cables wrapped as a single cable that have a protective member and outside jacket, and these cables are commonly referred to as trunk cables. Trunk cables are normally terminated at each end into the bottom of a patch panel. Jumper cables are then used from the top of the patch panel to the SAN fabric.

Typically trunk cables are used for longer runs between server and SAN fabric cabinets, as the trunk cable terminates at a patch panel, normally there is no requirement to make future cable alterations.

In Table 5-1 we compare the advantages and disadvantages of using non-structured and structured cabling practices for server cabinet to SAN fabric cabinet connections.

Table 5-1 Comparison between structured and non-structured cables

Non-structured cables	Structured cables
Unknown cable routing	Known cable route
No cable documentation system	Defined cable documentation
Unpredictable impact of moves, adds and changes	Reliable outcome of moves, adds and changes
Every under floor activity is a risk	Under floor activity can be planned to minimize risk
No waste, only the cables required are run	Initially not all fibers will be used; spare cables will be run for growth

There are a number of companies that provide fiber cabling service options. IBM provide this service offering with Fibre Transport Service Cabling System (FTS).

When planning the use of trunk cable over a longer distance it is important to consider the potential light loss. Every time a joint is made in a fiber cable there will be a slight light loss, with the termination at the patch panel there will be considerably more light loss.

The fibre installation provider should be able to calculate potential light loss to ensure the trunk cable run is within acceptable light loss limits.

Patch panels

Patch panels are commonly used to connect trunk cables, particularly between floors and buildings. They provide the flexibility to enable repatching of fibres but generally the panel configuration will remain the same after installation. Pay

attention not to mix fibre cables with different diameters when crossing patch panels.

5.1.3 Planning considerations and recommendations

Many miscellaneous considerations are needed to successfully install fiber-optic links for any protocol. However, the higher data rate and lower optical link budgets of Fibre Channel lends itself to more conservative approaches to link design. Some of the key elements to consider are:

- All links must use the currently predominant “physical contact” connectors for smaller losses, better back reflectance, and more repeatable performance.
- The use of either fusion or mechanical splices is left to the designer to determine the desired losses weighed against the cost of installation.
- Multi-mode links cannot contain mixed fiber diameters (62.5 and 50 micron) in the same link. The losses due to the mismatch may be as much as 4.8 dB with a variance of 0.12 dB. This would more than exceed the small power budgets available by this standard.
- The use of high quality factory terminated jumper cables is also recommended to ensure consistent performance and loss characteristics throughout the installation.
- The use of a structured cabling system is strongly recommended even for small installations.
- A structured cabling system provides a protected solution that serves current requirements as well as allows for easy expansion.
- The designer of a structured system should consider component variance affects on the link if applicable.

Much of the discussion so far has been centered around single floor or single room installation. Unlike earlier FDDI or ESCON installations that had sufficient multi-mode link budgets to span significant distances, Fibre Channel multi-mode solutions for the most part do not. Though the Fibre Channel standard allows for extended distance links and handles distance timing issues in the protocol the link budgets are the limiting factor.

Therefore, installations that need to span between floors or buildings will need any proposed link to be evaluated for its link budget closely. Degradation over time, environmental effects on cables run in unconditioned spaces, as well as variations introduced by multiple installers need to be closely scrutinized. The choice between single-mode and multi-mode devices may need to be made for many more links. Repeating the signal may also provide a cost effective solution if intermediary conditioned space can be found.

Since Fibre Channel provides a built in mirroring capability to SAN, in addition to its 10 km link distances using single-mode fiber, there will be more consideration for off-campus or across city links. In these cases, right-of-way issues, leasing of “dark” fiber (no powered devices provided by the lessors) issues, service level agreements, and other factors associated with leaving the client owned premises needs to be planned for and negotiated with local providers. The industry has also announced interest in providing wide area network (WAN) interfaces similar to those employed in the networking world of today. When these devices are made available, then connections to these devices will need to be included in the designs as well.

5.1.4 Structured cabling

Because of access to the Internet, the data centers of today are changing rapidly. Both e-business and e-commerce are placing increasing demands on access to and reliance on the data center. No longer is the data center insulated from the rest of the company and just used to perform batch processing.

Now, access and processing is a 24x7 necessity for both the company and its customers. The cabling that connects servers to the data storage devices has become a vital part of corporate success. Few companies can function without a computer installation supported by an efficiently structured and managed cabling system.

There are many important factors to consider when planning and implementing a computer data center. Often, the actual physical cabling is not given enough planning and is considered only when the equipment arrives. The result of this poor planning is cabling that is hard to manage when it comes to future moves, adds, and changes due to equipment growth and changes.

Planning a manageable cabling system requires knowledge about the equipment being connected, the floor layout of the data center(s), and, most importantly, how the system requirements will change. Questions that should be considered include:

- Will the data center grow every year?
- Will you need to move the equipment around the floor(s)?
- Will you upgrade the equipment?
- Will you add new equipment?
- What type of cabling do you require?
- How will you run the cables?
- How will you label the cables?
- Can you easily trace the cables if there is a problem?
- Will you need to backup the data at a remote site?

Answers to these important questions should be obtained as part of the early planning for the cabling installation.

5.1.5 Data center fiber cabling options

The first data center connectivity environment that used fiber cabling was IBM's ESCON architecture. The same structured fiber cabling principles can be applied in the SAN environment, and to other fiber connectivity environments such as IBM's Fiber Connection (FICON), Parallel Sysplex, and Open Systems Adapters (OSA). The examples throughout this chapter apply to structured fiber optic cabling systems designed to support multiple fiber-optic connectivity environments.

The need for data center fiber cabling implementation arises from the following three scenarios:

- Establishing a new data center
- Upgrading an existing data center by replacing the cabling
- Adding new equipment to an existing data center

IBM can help you design and implement a network that leverages existing investments, avoids costly downtime, and saves time and money when moving to performance enhancing technologies.

IBM Network Integration and Deployment Services

IBM Network Integration and Deployment Services helps businesses integrate and deploy a complex network infrastructure that leverages multivendor technologies. IBM will analyze existing networks, protocols, wired and wireless configurations to identify performance, interoperability, and connectivity requirements. Implementation planning, detailed logical and physical network design, rapid deployment and network rollouts, product installation and customizing and operational services for network and cabling infrastructures is provided.

This enables business to securely converge data, voice, and video networks, enables intelligent network infrastructures, and deploy mobility solutions by exploiting technologies such as virtual private networking (VPN), video and voice over IP (VoIP), fiber optic networking, content delivery networks, storage networking and wireless.

More information is available at Web site:

http://www-1.ibm.com/services/networking/integration/index.html

IBM Cabling Services and the Advanced Connectivity System

Today's telecommunications infrastructure, both copper and optical fiber, supports data rates that were undreamed of even a decade ago. The cabling infrastructure is at the core of every voice, data and multimedia network.

Integrating multivendor equipment has become challenging, time-consuming and increasingly dependent on how IT systems are physically connected. Proper planning, configuration and installation for connectivity is critical. Our professionals can analyze your existing network, protocols, wiring configurations and cabling infrastructure, identify system interoperability and connectivity requirements, and help you connect and integrate your cabling systems.

IBM has a wide array of premises, data center, server and storage networking solutions to help select and install the right cabling solution for the e-business infrastructure.

Cabling solutions using the IBM Advanced Connectivity System include copper and fiber solutions for any building or premises, and each is designed to be intermixed and adapted to different topologies. There is a choice of several grades of cabling infrastructures that meet or significantly exceed today's technical standards and performance demands.

http://www-1.ibm.com/services/networking/integration/acs.html

Metropolitan Area Network cables

Metropolitan Area Network (MAN) cables are typically used for business continuance between two sites. MANs used for business continuance normally consist of a diverse route of a primary and alternate cable. The alternate route is normally only used when the primary route is not available. To ensure we introduce no single points of failure it is critical these cables enter and leave the building at separate locations and at no point share the same cable run or equipment.

It is important to have a detailed intersite cable route plan to highlight any single points of failure and to determine the exact distance of both routes.

If the primary site is several kilometers shorter than the secondary route, there may be latency issues to consider when using the secondary route. It will only introduce problems, when there are parallel MAN links used in a shared manner, and the skew which comes with the different latencies cannot be handled by the protocol.

As the MAN cables enter the buildings, the routes of the primary and alternate cables should be clearly marked on the floor plan.

Ethernet cables

The majority of SAN products require IP addresses to enable remote software management. To ease the administration of SAN management it is common to place the Ethernet ports within the same LAN or VLAN and choose IP addresses from the same IP subnet.

Ethernet cables will need to be laid from the site Ethernet switch to the SAN fabric cabinet, and these cable routes should be detailed on the plan.

Future growth

Future technologies and design issues are mostly affected by length and attenuation due to increased speeds. Since future technologies are unknown, most organizations are pulling single-mode fiber along with the new multi-mode fiber while keeping the proposed distance limitations in mind when designing the cable plant.

One option is to leave the single-mode fiber un-terminated and dark for future technologies. Connectors, panels, and the labor to terminate, test, and install these items could be a significant cost, so leaving these cables un-terminated and dark can save money in the short term.

Another option is to proceed with the termination in anticipation of rapid technological developments. Indications point to a reduction in the cost of LW lasers. This would drive down the price of LW technology and LW equipment applications, influencing the adoption of single-mode usage as well.

5.1.6 Cabinets

For IBM SAN fabric components that do not come with an associated cabinet, you will have a choice of rack mount or non-rack mount feature codes. In most cases it is advisable to select the rack mount option for the following reasons.

- **Security:** To prevent unauthorized actions on the SAN fabric components, the cabinet can be locked and key access restricted to selected personnel.
- **Audit trail:** Hardware changes can be tracked by recording who has requested the cabinet key, and at what time.
- **Cable Management:** When there are large numbers of fibre cables hanging from the SAN fabric, it can be very difficult to locate and alter cables. You also have an associated risk that when you are making SAN cable alterations, you may damage other cables in the SAN fabric. The use of cable supports, cable ties or velcro strips will enable cables to be tied back along the cabinet edges, reducing the risk of accidently damage and enabling cables to be easily identified.

- **Hardware replacement:** When customer engineers need to repair or replace a part of the SAN fabric, it is important they have easy access to the component. The use of racks will guarantee they are able to access the device without disturbing any other SAN components.
- **Power outlets:** Most cabinets have a default number of power outlets, this number can be used to plan current and future SAN fabric power requirements.
- **Component location:** Each cabinet should be clearly labelled, these labels can be incorporated within the SAN fabric components naming standards. In the event of a problem or change the correct component can be easily identified.

In Figure 5-2 we show an example of a SAN fabric that has not been racked, and where only one cable has been labelled. It is easy to see how cables could become damaged and mistakes could occur.

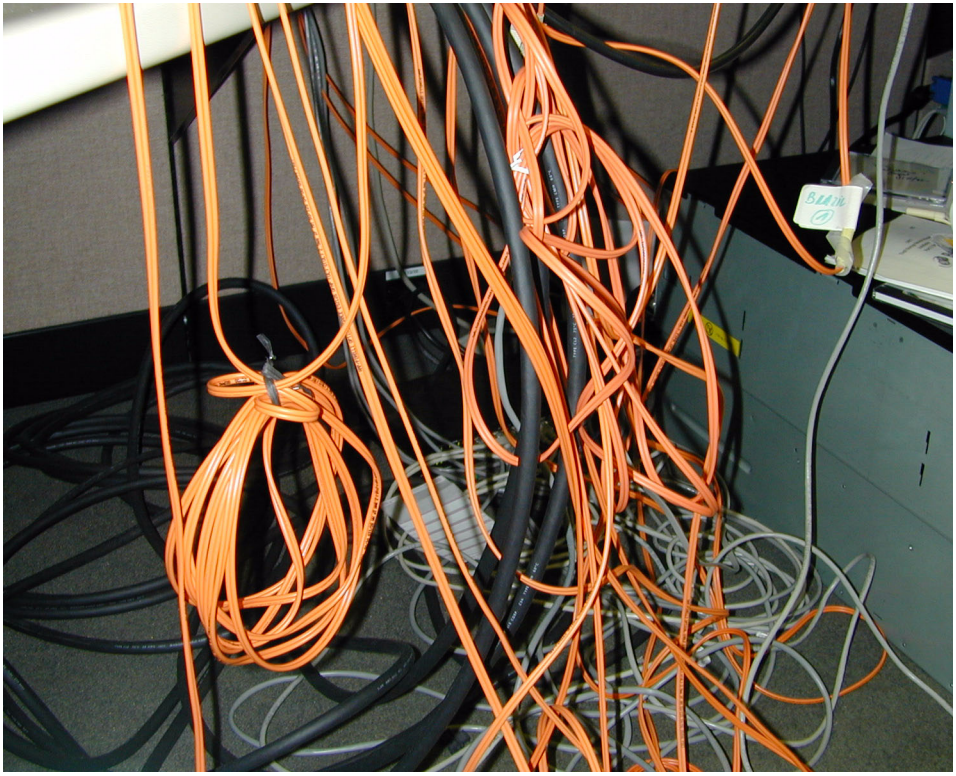


Figure 5-2 Messy cabling, no cabinet, and no cable labels

5.1.7 Phone sockets

Most of the larger SAN devices will have dial home facilities which will require a phone line. Although phone lines can be shared between devices, sufficient phone sockets need to be provided to prevent phone line bottlenecks. It is also wise to think about ensuring that spare phone lines are available should one fail, or be in use for a long period of time for any reason.

As an example, if a phone line was shared between an IBM ESS and a fabric component, and if log information had to be extracted from the ESS, the phone line could be busy for over one hour — and any potential error on the fabric component may go unnoticed.

5.1.8 Environmental considerations

In this section we consider some basic requirements for power sources and heat dissipation.

Power

The majority of SAN components have the option of single or dual power supplies. To realize the benefits of two power supplies, it is essential the power source supplying the devices is from two independent supplies.

In a cabinet full of smaller SAN fabric devices with dual power supplies, it is very easy to exceed the available number of power sockets.

You need to ensure the SAN fabric cabinet has sufficient power sockets to satisfy the SAN fabrics power requirement from both day one and a potentially full cabinet.

Heat

Several small switches in a SAN cabinet will generate quite a lot of heat. To avoid heat damage it is important the cabinet is located in a room that has temperature control facilities.

5.1.9 Location

Typically the location of the SAN fabric cabinets, servers and storage (disk and tape) will be dictated by available space and power supply.

The typical placement for components is in clusters around the periphery of the work area. This minimizes the length of cables and their exposure to points of failure. An alternative is to locate components in a central cluster. However, the smaller the area that the components are gathered in, the more potential exists

for a burst pipe, for example, taking out the whole fabric. It may be wise to distance some components apart from each other.

The further that the two components are apart, the less likely it is that a single disaster will render both of them unusable.

5.1.10 Sequence for design

Assuming you are cabling a facility with existing components, the usual sequence is to do the following:

- Base it on the server inventory and detail the current components accurately and completely.
- Determine what new components will be added and their location.
- Verify that the type of cable is appropriate for each connection.
- Calculate loss and attenuation for each connection and for the total system.
- Modify the design as needed.

A detailed floorplan should be drawn with cabinet and slot locations of all components of the SAN. The floorplan should include:

- Servers
- SAN fabric
- Storage devices
- Cable routes
- Cable type
- Cable entry and exit points
- Power points
- Power source
- Phone lines
- Ethernet cable routes
- Location of SAN ethernet hub and any required ethernet switches

If two buildings are connected using a MAN or similar, the cable routes, the total distance of both the primary and secondary routes, and the entry and exit points into the building need to be detailed.

On completion of the floor plan, the checklist displayed in Table 5-2 in should be performed to validate the proposed layout.

Table 5-2 Checklist for proposed layout

Check	Validate	Successful
Location of SCSI devices	Within 25 m of SAN bridge device	
Multi-mode, shortwave devices	Within 550 m (1 Gb/s) and 300 m (2 Gb/s)	
	Cable route uses 50/125 multi-mode fiber	
Single-mode, shortwave devices	Within 275 m (1 Gb/s) and 150 m (2 Gb/s)	
	Cable route uses 62.5/125 multi-mode fiber	
Long-wave devices no extender	Within 10 km	
Long-wave devices with extenders	Within 100 km	
Power source	Independent supply	
Number of power sockets	Sufficient number, including units with dual power supply	
Location of SAN devices	In a lockable cabinet	
	Sufficient space in SAN cabinets	
LAN	Sufficient free IP addresses and Ethernet ports for SAN devices	
Capacity	No physical constraints for growth - cables, full cabinets, and so on	
Phone sockets	Location and quantity of phone sockets that require dial home	

5.2 Naming conventions

Use of descriptive naming conventions is one of the most important factors in a successful SAN. Good naming standards will improve problem diagnostics, reduce human error, allow for the creation of detailed documentation and reduce the dependency on individuals.

5.2.1 Servers

Typically, servers will already have some form of naming standard in place. If a server name does exist, it should have been captured during inventory, as described in 5.1.1, “SAN inventory” on page 212.

The local server name is typically used as the host name defined to the disk system. For the ESS you would normally use the server name in the server description field. The same local server name can be used within the switched fabric as an alias for zone settings, and whenever possible the use of the server name should be consistent throughout the SAN.

5.2.2 Cabinets

SAN fabric cabinets should be labelled to adhere with local site standards.

5.2.3 SAN fabric components

A good naming convention for the SAN fabric component should be able to tell you the physical location, component type, have a unique identifier and give a description of what it connects to. The following are some descriptor fields that may be considered when designing a fabric naming convention. If your SAN only has one vendor type or only one cabinet the name could be a lot simpler.

Component description

This should describe the fabric component and the product vendor (for mixed vendor environments) which will help you locate the management interface and the component number within the SAN. For example, to give it a unique identifier you may want to use something similar to the following:

- Type — Switch (S) Director (D) Gateway (G) Hub (H) Router (R)
- Vendor — Brocade (B) CNT)(I) McDATA (M) Cisco (C)
- Number — 1 - 99

For example, the third Brocade Switch in cabinet one would be:

- S3 B

Connection description

This should detail what the component is connecting to. For highly available devices such as the ESS, it is important to understand which cluster side of the device the component is connected to. This will help prevent potential mistakes in the SAN design. For devices used to expand the SAN that do not connect to disk or tape, we will simply identify them as cascade.

- Connection — Disk (D (for ESS either cluster A or B)), Tape (T), Cascade (C)
- Number — 1 - 99

To continue our example, the third Brocade Switch in cabinet one connecting to ESS3 Cluster A would be

- S3 B D3A

Physical location

This may be the cabinet descriptor field and, for example, SAN cabinet one could be C1. For our example this would give us:

- S3 B D3A C1

We show how our name is developed in Figure 5-3.

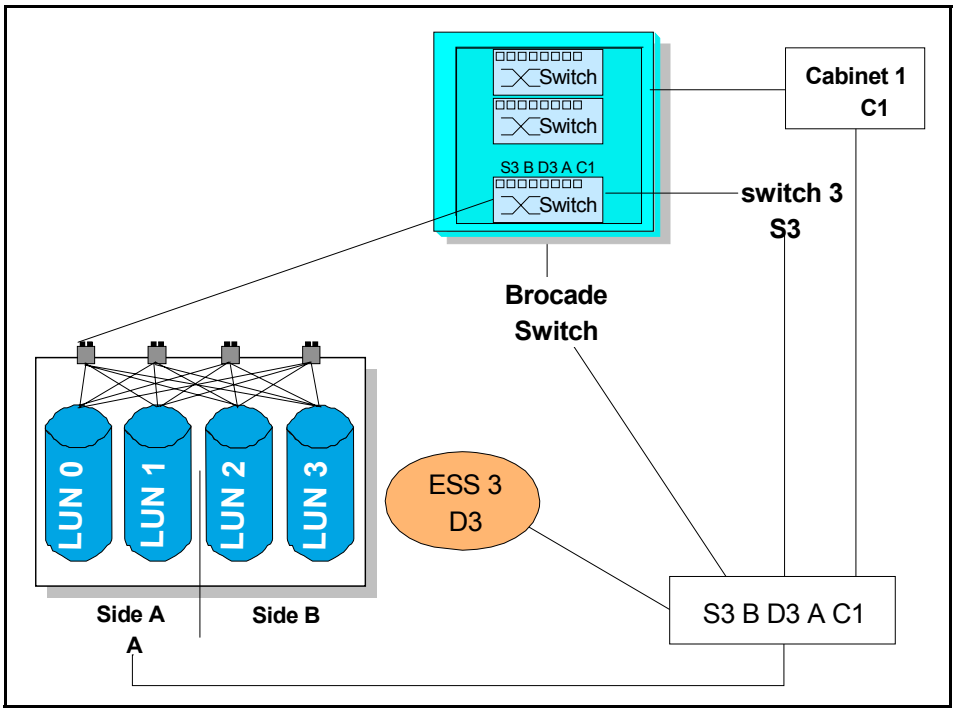


Figure 5-3 Naming convention development

5.2.4 Cables

Modifications to a SAN that does not have sufficient labelling in place could lead to the incorrect selection and reconfiguration of the SAN with potentially disastrous effects.

When determining the exact cable identification tag, try to avoid using device specific names. Such names do not take into account adding or subtracting devices, or devices being renamed, which are fundamental parts of a SAN.

The chosen cable tag naming standard should be incorporated in a detailed SAN fabric port layout plan. The port layout plan will enable you to identify the exact devices to which the cable is connected.

Any SAN cable reconfigurations should have an associated change record. Part of that change process should include a pointer to update the cable tag and port layout plan. Adhering to this plan will ensure that the document is always kept up-to-date.

5.2.5 Zones

Understanding software zones defined by another SAN administrator, when no naming standards have been defined, can be very difficult. Researching the zone setting can be time consuming and can cause problems if activated incorrectly. The introduction of site standard software zone naming standards will minimize this risk.

A good zone naming standard should consist of a meaningful description of the servers effected by the zone and what you are hoping to achieve.

For example, if we assume that the WWNs of servers ABC_FIN1 and ABC_FIN2 (the company production finance server) have been zoned to only see the WWN of disk ESS3.

If we called this zone1, the creator would probably understand what the zone was doing. If, at a later stage, the finance department requested the ability to access archive files that resided on ESS1, a modification to zone1 would be required. However, as zone1 gives us no descriptive information as to the contents of the zone itself, for anyone other than its creator, this change would be very difficult and may involve investigating every zone to see its contents.

If, however, we had called the zone PRODFINESS3, locating and modifying the zone would be much simpler.

5.3 Documentation

There are a number of software tools (such as Tivoli's Storage Area Network Manager) that are able to provide detailed information and documentation about the SAN. This includes connection diagrams, server utilization reports and status monitors and more besides.

These products, although very good at giving you an overall picture of the SAN, do not have sufficient detail to be the only source of information in order to manage the SAN.

Data that needs to be collected and recorded in SAN documents include the following:

- **Floorplan:** The floor plans of all SAN machine rooms
- **Server Inventory:** A list of servers connected to the SAN, type of Host Bus Adapters (HBAs), World Wide Name of HBAs
- **List of fabric components:** The naming convention and list of all fabric components
- **Space allocated:** A list of LUNs allocated to servers
- **Space available:** A list of free space in the disk device
- **Fabric Connection:** A detailed wiring diagram of the SAN fabric
- **Fabric Port layout:** A port usage plan detailing what ports are currently used, which ports are spare
- **Zone Information:** Both hard and soft zoning in place
- **IP addresses:** A list of IP addresses for all fabric components, as well as a list of spare ones
- **Fabric model Serial numbers:** The IBM product serial numbers (used when raising a call with the IBM call center).
- **Micro code versions:** The level of micro code installed on the disk devices (used when raising a call with IBM call center)
- **Firmware version:** The level of firmware running on the SAN fabric (used when raising a call with IBM call center)
- **Procedures:** A step-by-step how to perform a SAN function guide

In addition to the information documented for the primary site there will also be a requirement for a similar level of documentation for the disaster site.

5.4 Power-on sequence

After a site power-down, it is important to stagger the power-up of the servers connected to the SAN fabric. The reason for this is that during boot-up some operating systems will scan all the switch ports and will look up other HBA ports.

With some combinations of HBA cards this can have an adverse effect on other servers in the SAN. Symptoms can be unpredictable, ranging from clusters being brought down, to NT losing SAN access and requiring a reboot.

The use of soft zoning to separate vendor cards would prevent the risk of this occurring.

5.5 Security

Consolidating storage onto central devices has many benefits, but can also increase the risks to your business. With large amounts of critical data in one location, it is important to ensure that you are providing the maximum protection of your data. That topic was discussed from a general point of view in 3.16, “Security” on page 168.

5.5.1 General

All SAN software management tools come with a default userid and password which typically has the highest level of authority. Obtaining unauthorized access to these IDs would enable a user to alter zone information and give servers access to data that would otherwise be protected.

Generally, SAN software products do not police their userids and passwords and will not request them to be changed. It is common to find default IDs remaining on the system months after the SAN has been installed. The userids and passwords need to be changed as part of the installation, and passwords should be altered at regular intervals from then on.

Operational controls

In 3.16, “Security” on page 168, operational controls were described as part of a whole security concept. Numerous attempts are required to structure and define single steps and tasks in order to archive the highest possible security level in the IT environment. Such tasks as backup and recovery, physical security, and so on, are defined in policies and grouped in operational controls.

The Acceptable Use Policy defines acceptable use of equipment and computing services, and the appropriate employee security measures to protect the organization's corporate resources and proprietary information.

A security policy can start simply as an Acceptable Use Policy for network resources, and may grow to large documents as a complete set of laws, rules, and practices that regulate how an organization manages, protects, and distributes sensitive information. In such a policy, you can state which administration group will have access to which components: Does your SAN administration manage your servers too? (see 3.16, "Security" on page 168 about that topic). By building up your security policy, you will define and publish your security rules. RFC 2196 suitably defines a security policy in a 73-page *Site Security Handbook*.

RFC 2196: "A security policy is a formal statement of the rules by which people who are given access to an organization's technology and information assets must abide."

The full text of RFC 2196 is available under this URL:

http://www.ietf.org/rfc/rfc2196.txt

5.5.2 Physical access

Physical security is an absolutely essential component of any comprehensive security plan. Even with excellent software controls in place, physical access to enterprise elements opens the door to a whole range of security issues. To ensure physical security, fabric devices should reside in environments where physical access controls provide adequate protection.

Secure machine room

With the flexibility of a SAN there is the temptation to distribute the SAN fabric in the location of the servers. This should be avoided if the locations cannot be adequately protected.

Cabinet protection

As detailed in 5.1.6, "Cabinets" on page 222, fabric cabinets should be lockable with restricted access to the key.

Switch protection

SAN switches usually provide RS-232 and Ethernet connections. Access to either of these interfaces must only be given to trusted persons, as all of the vital data of switches and fabric can be monitored and changed from here.

Cable protection

Damage to a fiber optic cable can result in performance degradation or a complete loss of access to the data. Fibers should be laid in cable trays or trunks with rodent control measures in place.

5.5.3 Remote access

There are a variety of ways to obtain information from fabric switches. Common management access methods involve the use of telnet for command line functionality, HTTP for Web-based access, in-band Fiber Channel for management server access, and console access for direct switch connectivity. Common to all of these applications is that they need IP connectivity — and the IT community has been alarmed for years about how many ways there are to break into IP hosts. Each of the possible access methods has its associated security issues.

Telnet: The essential problem with telnet access is that it transmits uninterrupted the username, password, and all data going between the management system and the switch. Any user with a promiscuous network interface card and data-sniffing programs can capture the whole data transfer back and forth, including account and password. Telnet is not recommended.

HTTP: Similar to the telnet issues mentioned above, when a system uses a Web-based application like Web Tools to logon (authenticate) to the switch in order to run privileged commands, it passes the login information not encrypted.

Management Server: This remote management method uses an in-band Fibre Channel connection to administer or obtain information from the fabric switches. By default, it grants access to any device. However, it is possible to create an access control list to limit the WWNs of devices that can connect to the switch using this method. The management server may use an SNMP agent for remote switch management using IP over the Ethernet and Fibre Channel interfaces. Within the SNMP model, a manageable network consists of one or more manager systems (or network management stations), and a collection of agent systems (or network elements).

Console Access: Although not usually thought of for remote access, it is possible to adapt console connections to remote use through the use of terminal server devices. Thus, an organization can use telnet, secure shell (SSH), or some similar application to connect to the terminal server, which then in turn connects to the selected device through the console interface. This solution has the potential to provide additional security through the use of third-party products.

Secure Shell (SSH): is a client-server network application. SSH provides additional security by encrypting data, userids, and passwords. The switch acts as the SSH server in this relationship. The SSH client provides a secure environment in which to connect to a remote machine. It uses the principles of public and private keys for authentication. SSH keys are generated by the SSH software. This includes a public key, which is uploaded and maintained by the cluster and a private key that is kept private to the host that is running the SSH client. These keys authorize specific users to access the administration and service functions on the cluster. This scheme is based on public-key cryptography, using a scheme known commonly as RSA. The encryption and decryption is done using separate keys. This means it is not possible to derive the decryption key from the encryption key. Physical possession of the private key allows access to the cluster, so it must be kept in a protected place, such as the .ssh directory on the AIX host, with restricted access permissions. When an SSH client (A) attempts to connect to an SSH server (B), the key pair is needed to authenticate the connection. The key consists of two halves: the public and private keys. The SSH client public key is put onto the SSH Server (B) using some means outside of the SSH session. When the SSH client (A) tries to connect, the private key on the SSH client (A) is able to authenticate with its public half on the SSH server

You'll find customers who lack a firm security policy and so leave userids and passwords as default, and apply public network addresses to their switches — and then there is the other extreme: customers who disconnect the switches from the IP network. To satisfy the needs of security *and* manageability, a SAN needs the IP-connectivity *and* the strongest possible security features available.

5.6 Education

It is important that the educational requirements of all those involved in implementing and maintaining the SAN are considered in order to gain the maximum benefit from the SAN — and minimize the room of remaining human error. Therefore, the right skills have to be defined and a way to validate these skills.

5.6.1 SAN administrators

The SAN administrator is commonly responsible for effective utilization of the SAN resource, resource protection, balancing traffic, performance monitoring, utilization trending, and error diagnostics, in addition to many maintenance functions. The SAN administrator must be identified as the focal point for any additions, deletions or modifications of the SAN environment.

Management of the SAN is usually performed using the software interface that comes with each of the SAN fabric components. There are a number of software products that enable all components of the SAN to be monitored and managed from a central point. Most SAN software management tools have facilities to create different levels of access and these range from view through to full administration.

5.6.2 Skills

As already stated in 3.16, “Security” on page 168, the Fibre Channel fabric and its components are considered shared resources. This accessibility, plus the increase in both quantity of data and users of that data, has heightened the awareness of the IT community to security exposures. As such, Fibre Channel fabrics and their components are prone to the same types of security breaches once associated only with IP networks. These concerns of the SAN community reflect the same concerns of the IT community.

But not only is security a major link between SANs and LANs — the nature of connectivity is the same and it is likely to become even closer as technology matures.

Good networking skills are needed to implement and operate SANs — the kind of skills which have typically been developed in LAN environments. Good operating system and diverse platform skills are also required with this increase in connectivity — so both mainframe and open systems skills may now be needed in the same person.

The SCSI heritage in the SAN is enormous. Good storage skills must not be overlooked. The combination for SAN success in implementation and services will likely be a merged skill of networking, storage, system, and security skills.

5.6.3 Certification

Although not a part of education, certification is a good indicator as to the core competency and ability of an individual. There are a number of programs available.

IBM Professional Certification Program

The IBM Professional Certification Program is designed to validate technical skills to network administrators and integrators, systems integrators, solution architects and developers, resellers, technical coordinators, sales representative, or educational trainers.

The Program has developed certification roles in various technical specialties to guide the participants in their professional development.The Professional Certification Program from IBM provides with a structured program leading to an internationally recognized qualification. The certifications are continually being updated and changed based on the latest technological and market changes.

The IBM Professional Certification Program offers Storage programs among other programs such as AIX, Linux, DB2®, WebSphere®, Lotus®, and Tivoli., The IBM Professional Certification Program provides TotalStorage certifications by offering IBM IT Certified Specialist roles in High End Tape Solutions, High End Disk Solutions, Open System Storage Solutions, Storage Sales, and TotalStorage Networking and Virtualization Architecture.

IBM TotalStorage Networking and Virtualization Architecture

The IBM Certified Specialist designs IBM TotalStorage end-to-end storage networking solutions to meet customer needs. This individual provides comprehensive storage networking solutions that include servers, storage networking, storage devices, storage virtualization, management software, and services. This specialist has detailed knowledge of SAN technologies, storage devices, storage virtualization, and the corresponding management software. He or she has broad knowledge of IBM storage products and their features and functions, and can describe in detail the storage networking strategy and solutions, industry, competition, and business trends.

To learn more about the IBM Professional Certification Program, visit the Web site:

`http://www-1.ibm.com/certify/index.shtml`

SNIA Storage Networking Certification Program

The Storage Networking Industry Association (SNIA) is introducing the industry's first vendor-independent certification program for storage networking called Storage Networking Certification Program (SNCP). The program was developed in response to demand from enterprise customers worldwide in order to provide standards for measuring the storage networking expertise of IT professionals.

The SNIA has identified the technologies that are integral for IT professionals to understand and deploy storage area networks. The first modules of the SNIA SNCP, developed for the SNIA by the industry-leading training company Infinity I/O, included certification exams that test the candidates' knowledge of Fibre Channel SANs.

The certifications have been enhanced to reflect the advancement and growth of storage networking technologies over the past few years, and to provide for expanded offerings in the future. The SNIA SNCP establishes a uniform standard by which individual knowledge and skill sets can be judged.

The SNCP certification exams were updated in 2004 to offer four levels of certification:

- Level 1 - SNIA Certified Professional (SCP)
- Level 2 - SNIA Certified Systems Engineer (SCSE)
- Level 3 - SNIA Certified Architect (SCA)
- Level 4 - SNIA Certified Storage Networking Expert (SCSN-E)

The SNCP structure now consists of four certification domains: Concepts, Standards, Solutions and Products. The higher certifications require additional exams in more certification domains.

To learn more about SNCP and the depth of the different certification levels, visit the Web site:

http://www.snia.org/education/certification/

6

Host Bus Adapters

The IBM supported SAN environments contain a growing selection of server Fibre Channel Host Bus Adapters (HBAs), each with their own functions and features. For the majority of open systems platforms, this presents us with the opportunity to select the most suitable card to meet the requirements of the SAN design. IBM utilizes HBAs from Emulex, JNI and QLogic. As the porfolio is extremely rich in features and function, we will not cover each available HBA in this redbook; instead we will give pointers to where more information can be found.

© Copyright IBM Corp. 2004. All rights reserved.

6.1 Selection criterion

In this section we look at a some points that should be considered when selecting the right HBA.

6.1.1 IBM supported HBAs

The first and most important factor to consider, when selecting a Fibre Channel HBA, is whether it is supported by IBM for the server make and model, and also the manner in which you intend to implement the server. For example, an HBA may be supported for the required server, but if you require dual pathing or the server to be clustered, the same HBA may not be supported.

To ensure that the HBA is supported by IBM in the configuration you require, refer to this interactive HBA search tool:

http://knowledge.storage.ibm.com/HBA/HBASearchTool

For IBMers only, for an HBA that is not detailed as supported for a specific platform, support can be requested using the Request Product Quotation (RPQ) process.

6.1.2 Special features

Any special functions you require from your SAN need to be considered, as not all HBAs may support the function. These functions could include:

- Dual connection
- Performing an external server boot
- Connection to mixed storage vendors
- Fault diagnostics

6.1.3 Quantity of servers

Another factor to consider is the number of servers in your environment that will require Fibre Channel HBAs. Having a common set of HBAs throughout your SAN environment has a number of advantages:

- It is easier to maintain the same level of firmware for all HBAs.
- The process for downloading and updating firmware will be consistent.
- Firmware and device driver can be a site standard.
- Any special BIOS settings can be a site standard.
- Fault diagnostics will be consistent.
- Error support will be from a single vendor.

6.1.4 HBA parameter settings

For some suggestions as to the best HBA parameter settings refer to:

http://knowledge.storage.ibm.com/HBA/PDF/HBASettings.pdf

7

SAN design considerations

In this chapter, we go into the general details associated with satisfying the business and technology goals of your organization by showing how the SAN building blocks can be implemented.

© Copyright IBM Corp. 2004. All rights reserved.

7.1 What do you want to achieve with a SAN?

Before starting to design your SAN, you need to define what the application type is that you will be using. The type of application will help act as a guide in the design of your SAN environment.

For meeting the different needs for your application environment it is likely that you will need to utilize different types of products, software, and services. In the following list you can see some of the typical applications of a SAN:

- Storage consolidation
- HA solutions (Clusters, Web farms, and so on)
- LAN-free backup
- Server-free backup
- Server-less backup
- Disaster recovery

We will explain what we mean by them as these are key to understanding some of the general SAN designs that we cover within this chapter.

Storage consolidation

With storage consolidation you are logically connecting all your storage resources (disk, tapes) into one large group which can be accessed by anyone who wants to utilize those resources. With this type of consolidation you can achieve better utilization of your storage resources. A SAN is just one of the technologies which can help you in storage consolidation. For example, storage consolidation of disk resources can also be achieved at the level of the storage device. This type of consolidation has its limits, which can be overcome with a well planned design and usage of a SAN.

High availability solutions

SANs can also be utilized in various high availability solutions. They offer shared access to the storage devices across virtually unlimited distances on a high speed dedicated network. This type of resource sharing gives you the ability to, for example, build cluster solutions, which can be geographically dispersed. These clusters can be either highly available clusters or high performance clusters.

LAN-free backup

By providing high speed data transfer, SANs are the ideal platform for LAN-free backup. The idea of LAN-free backup is to share the same backup devices across multiple backup clients directly over the SAN. With such a setup the client can perform backup directly to the backup device over a high speed SAN, compared to LAN backup, where all the traffic goes through one backup server. Utilizing LAN-free backup allows you to shorten backup windows, thus freeing up more time to perform other more important tasks.

Server-free backup

As in the LAN-free backup environment, SANs provide the infrastructure for server-free backup. Server-free backup is only possible in cases where the storage devices, such as disks and tapes, can talk to each other over a dedicated network.

This is only possible by utilizing a SAN, which gives you any to any access at the storage level. In the server-free backup scenario, the client will initiate the SCSI outboard command to the storage device that effectively copies the volume with data to a particular backup device. This backup device does not need to be tape.

Server-less backup

By utilizing the any to any storage design, the SAN can be used to implement server-less backup. The SAN allows us to implement high speed data sharing across the various clients. With this we can share the data which has to backed up among the production servers, and the servers performing backup tasks. In such a setup the backup server can backup the production data, off-loading the processor usage from the production server, which is usually needed to perform backup tasks.

Disaster recovery

One of the characteristics of a SAN is that it will allow you to implement disaster recovery protection for your storage. By allowing you to spread the storage across longer distances than other storage technologies, you can, for example, copy your data to another location. Because of the SAN high speed data transfers, the data replication can be implemented seamlessly into existing operations. The data replication can be achieved at the storage device level, such as Remote Copy functions; or at the level of the server operating system such as Remote Mirroring across the SAN. In both cases, we are utilizing the distance capabilities and the speed of the SAN.

The type of SAN solution chosen will also be determined by taking into account the following factors:

- Flexibility
- Goals to achieve
- Benefits expected
- TCO/ROI
- Investment protection (for example, reuse of the components)

Flexibility

Flexibility will play a big role in designing the SAN. By this we mean how flexible is your design to adopt new requirements. For example, adding new resources, changing the connectivity type, changing the topology, and so on. The design has to be flexible to sustain sudden, unplanned for changes.

Goals to achieve

The design will be affected by the technological goals that you want to achieve. An example of these goals could be the other technologies that you wish to integrate with, the management of the design, interoperability with existing equipment, the throughput you want to achieve, and so on. With this in mind, the design has to be carved out to meet all the goals you want to achieve.

Benefits expected

With every new technology incorporated into your environment, you are probably expecting to realize some benefits for your operations. To fulfill the technology or business expectations, you need to adapt your design taking into account the factors that we described in Chapter 1, “Introduction” on page 1.

TCO/ROI

The TCO and ROI are two important considerations in your SAN design.

Investment protection

When you are designing a SAN, it makes sound business sense to take into account investment protection. For example, the components used in the design should be upgradeable, or at least prepared to adopt any future standards and technologies. If this is the case, you will not need to replace your equipment every time you wish to introduce new technology.

7.2 Existing resources needs and planned growth

There is some important data which has to be collected before starting to outline the design of your SAN. For example, it is not just enough to count the number of server and storage devices and plan the equivalent number of ports for your SAN. This would be a good example of a poor approach to a SAN design.

7.2.1 Collecting the data about existing resources

Before selecting a SAN design you will need to understand the nature of the estimated traffic. Which servers and storage devices will generate data movements. Which are the sources, and which are the targets? Will data flow between servers as well as from servers to storage? If you plan to implement LAN-free or server-free data movement, what are the implications? How much data will flow directly from storage device to storage device, such as disk to tape, and tape to disk? What is the protocol? For instance, is this standard SCSI, or are you including digital video or audio?

What are the sizes of data objects sent by differing applications? Are there any overheads which are incurred by differing Fibre Channel frames? What Fibre Channel class of service needs to be applied to the various applications. Which departments or user groups generate the traffic? Where are they located, what applications do each community use, and how many in the user group? This information may point to opportunities for physical storage consolidation. It will also help you to calculate the number of Fibre Channel nodes required, the sum of all the data traffic which could be in transit at any time, and potential peaks and bottlenecks.

Can you identify any latent demand for applications, which are not carried out today because of constraints of the existing infrastructure? If you introduce high speed backup and recovery capabilities across a SAN, could this lead to an increase in the frequency of backup activity by user groups? Perhaps today they are deterred by the slow speed of backups across the LAN? Could the current weekly backup cycle move to a daily cycle as a result of the improved service? If so, what would this do to SAN bandwidth requirements?

With all this information in hand, you should be able to identify all of the following important connection parameters for your SAN design:

- The number of SAN ports needed today (this can be derived from the number of devices which will participate in the SAN and with number of ports per device. It is not necessary that all devices will have the same number of ports).
- The desired throughput from servers to the storage devices (disks or tapes). This data can increase the number of ports on the server and on the storage

side of your SAN. This will also affect the number of Inter Switch Links (ISL) in a fabric.

- The minimal throughput in case of a failure of redundant components in your SAN. In a worst case scenario, this can double the number of ports needed.
- The minimal throughput in the case of upgrading components. Depending on the core components you will use, this can also increase the number of ports needed in your SAN.
- You should also identify the type of port connectivity, for example is it a fabric or loop. This will help you in identifying the number of different port types needed in the SAN design.

After you have collected the existing data which covers all your requirements, you should have a rough picture of the capacity needed to build your SAN.

But it does not end here. We recommend that you include any future needs into your design.

7.2.2 Planning for future needs

One of the very important areas which has to be considered in the SAN design is future growth. In this section we will limit ourselves only to the growth of the IT resources connected to our SAN. Of course, the growth of those resources is tightly related to business growth.

To plan for your future needs you need to identify the following parameters:

- What are the planned connectivity upgrades?
- What are the planned performance upgrades and how will they affect the port count? For example, you can upgrade the performance of the server, but you do not need to upgrade the connectivity to the SAN.
- What is the predicted growth of the business, and how will it affect the needs for more storage capacity and performance?
- What are your plans regarding any future change of the technology, introducing new technology, upgrade policy and other changes? For example, how fast will you migrate to higher speed SANs?
- What are your plans for the maintenance policy and how is this incorporated into your operations plan?
- Do you plan to introduce any disaster recovery implementations into your environment?
- What is the impact to application changes? For example, if I change the server platform, how will this impact my storage setup?

7.2.3 Platforms and storage

How many servers and what are the operating platforms that will be attached to the SAN? The majority of early SAN adopters have tended to implement homogeneous installations (that is, supporting a single operating platform type such as all Netfinity, all HP, or all Sun servers). As SANs are maturing, the trend is towards larger scale networks, supporting multiple heterogeneous operating platforms (such as AIX, Linux, Windows NT/2000, and so on).

Fibre Channel capable servers require Fibre Channel HBAs to attach to the SAN fabric. The choice of HBA is probably already decided by the server vendor.

Before you decide how many HBAs you require in your host to achieve optimal performance, you need to evaluate the performance of the server. Fibre Channel HBAs today transfer data at 200 MB/s. Can the system bus provide data at the same or higher speed? If not, the HBA will not be fully utilized. The most common system bus in use today is the Peripheral Component Interconnect bus (PCI), available in various generations, ranging from 33 MHz at 32-bit (132 MB/s) to PCI-X running at 133 MHz with 64-bits (~1 GB/s). The Sun SBus operates at 50 MB/s, and the HP HSC at only 40 MB/s.

If the system bus delivers 132 MB/s or less, you will only need to attach one Fibre Channel HBA to the bus to potentially overrun the bus speed. If you attach a second HBA, it should only be for redundancy purposes. A PCI-X bus, should be able to cope with four or five HBAs, without becoming saturated.

Our recommendation is to distribute adapters across the available system buses, without overloading any of them. For pSeries, see the *Adapter Placement Reference for AIX*, SA38-0538 manual.

Another major component of your current assets are the storage systems. You may have a variety of internally attached disk devices, which will not be relevant in a SAN operation. Also, you may have externally attached JBODs or RAID disk subsystems, and tape drives or libraries, which can be utilized within the SAN. These current assets have implications for the selection of interconnections to the SAN. You may wish to support existing hardware which are SCSI or SSA compatible, and which will need to be provided with router or gateway connections for protocol conversion to Fibre Channel.

Armed with all this data, we can make a close estimate of the port count needed in our SAN. We have also outlined the performance requirements for our SAN, which means that we have established the paths from servers to storage devices or from storage device to device, and we have identified the bandwidth needed. We have also identified what impact we can afford in the case of maintenance or upgrades.

Now we can start to outline our core storage area network design.

7.3 Select the core design for your environment

In this section, we will cover the core design of a SAN. We will show what types of designs suit different application needs.

7.3.1 Selecting the topology

The most fundamental choice in the design of your SAN is the selection of the most appropriate topology. This selection may be colored by the overall approach to SAN planning that your organization wishes to adopt.

The question is whether to have a top-down or bottom up design? In other words, should you try to design a corporate strategy, with a view to implementing an enterprise wide SAN; or should you address the problem from the perspective of individual departments or user groups, and implement multiple SANlets. Maybe these small SANs will later merge into an enterprise wide solution.

This is a difficult question to answer. Probably it will be answered differently depending on the size of the organization, the IT management philosophy, the politics of the organization, and the business objectives. It is also shaped by the degree of risk which you associate with the implementation of an enterprise wide SAN today. Not all server platforms can easily participate in Fibre Channel configurations, and the rate of change is extremely rapid.

The majority of early SANs were relatively small, point solutions. By this we mean that they were designed to address a specific “pain” or problem. Many users have implemented simple point to point Fibre Channel solutions to solve distance or performance issues. Many others have installed small clustered server solutions, or shared storage capacity by means of FC-Arbitrated Loops because this provides improved connectivity and better utilization of storage resources. Others have designed switched fabric solutions for departments; or have used FC directors to facilitate large scale storage consolidation in campus locations.

In the past then, the bottom up approach was the most pragmatic: solve specific application needs to deliver value to your organization. Today a larger scale view should be taken. You should establish some common guidelines or standards regarding the purchase of equipment within the enterprise. This will facilitate inter-operability, and avoid dead end investments which cannot be integrated in a larger SAN environment as you expand.

You may decide that there are a number of discrete and independent operating environments within your organization, and these will not need to be inter linked in the future. If so, you may choose to establish SAN islands which are configured with different topologies and components in which cross island inter-operability is not required.

Current and future SANs should be switched fabric environments. This is because a fabric offers the greatest flexibility and scope for the future. Remember that FC-AL is not designed for high performance, and it does not scale. As you add more devices on a loop, performance will tend to reduce because of the shared bandwidth, and arbitration overheads. A maximum of two servers is advisable on a loop. If one server fails and has to reboot, causing a new LIP, it will momentarily interrupt the whole loop. Availability is therefore also a serious consideration.

FC-SW, on the other hand, scales performance as you add nodes. This does not apply to Inter Switch Links (ISLs) which do not add bandwidth between end nodes because ISLs reduce the number of end to end connections. Secure zones and masked LUNs can be created so that only authorized servers can access specific information. FC-SW provides comprehensive, flexible growth options.

7.3.2 Scalability

Any design of a SAN should also address the scalability issue. This means that the design you are putting in place must be able to grow if necessary. You should not lock your SAN design into a closed environment, impacting your ability to expand. You should be able to expand on demand if needed, for example by adding additional port capacity just for a period of high peaks.

7.3.3 Performance

We must address performance issues in our design. There are several important points to consider as performance factors:

- **Hop count:** In SAN designs, a hop is counted when a frame travels from switch to switch in the fabric. For example, if you have two switches connected in the fabric, this is one hop for the frame traveling from the device connected to one switch to the device connected on the other switch. More hops means you have more latency in the SAN.
- **Latency:** When a frame travels through a switch, there is a very small delay in travelling from the entry port to the exit port. This is the latency of the switch, and is usually in the order of a couple of microseconds. If the frame has to hop from one switch to another, then the extra switch will add additional latency.

- **Over-Subscription:** With this term, we mean the number of devices which want to talk to the same device. For example, the number of servers talking to the storage device. If, for example, your storage device can handle up to 100 MB/s on one port and four servers with a speed of 60 MB/s would like to get the data from it, then you have 4 x 60 MB/s = 240 MB/s to 100 MB/s; this is a 2.4:1 over-subscription ratio.
- **ISL Over-Subscription:** With this we mean the ratio of possible switch ports traffic going over the ISL to another switch. This is typically worse in a meshed fabric than a core-edge fabric, since large meshed fabrics usually require more hops. The worst case scenario would be connecting to `n` port switches over one ISL. In this case, we would have `n` to `1` over-subscription ratio, which will become worse if the number of ports on the switch increases.

  Note: If all ports on the switches are operating with the same speed, it is fairly simple to calculate the ISL over-subscription ratio. In cases where some of the ports are capable of sustaining higher speeds, for example 2 Gb/s against 1 Gb/s, this calculation can become quite complex.

- **Congestion:** Congestion is when over-subscription is excessive. For example, too many servers are trying to talk to the same storage device over a common link. In this case, multiple servers are contending for bandwidth. Because the link has limited bandwidth, the servers will be throttled down to the available bandwidth.
- **Blocking:** Blocking means that the data does not get to the destination as opposed to congestion where data will be delivered albeit with a delay. If we take again our example of more servers talking to the same storage device over the same ISL, then in a blocking environment whenever the new server tries to use the ISL (which is already in use) it will be denied access and it will have to wait until the ISL is free.
- **Fabric Shortest Path First (FSPF):** FSPF is defined in the Fibre Channel standards and is used in fabric switches to discover fabric topology and route frames correctly in order.

  Note: FSPF provides load sharing among equal cost links. Do not confuse this with load balancing.

- **Fan-out:** This is the ratio of server ports to a single storage port. This is important in the SAN design, because, for example, if your storage device has only one connection and six servers are connecting to it, you will have a ratio of 6:1. In such an example, it is sometimes reasonable to have less ISLs than the full capacity of the servers, because the storage end can only handle limited bandwidth.

Note: Fan-out differs from over-subscription in that it represents a ratio based on connections rather than throughput.

- **Fan-in:** This is the ratio of storage ports to a single server port. This information is also important in the SAN design. For example, by reallocating the storage ports across the fabric you could overcome a bad ISL over-subscription ratio.

7.3.4 Redundancy and resiliency

When designing the SAN, you should also provide redundancy and resilience in your design. Redundancy is the duplication of components up to and including the entire fabric to prevent failure of the total SAN solution. Resiliency is the capability of a fabric topology to withstand failures. We can group SAN designs into four types:

Single fabric non resilient design

All fabric components are connected together in a single fabric, and there is at least one single point-of-failure.

We show an example of such a fabric in Figure 7-1.

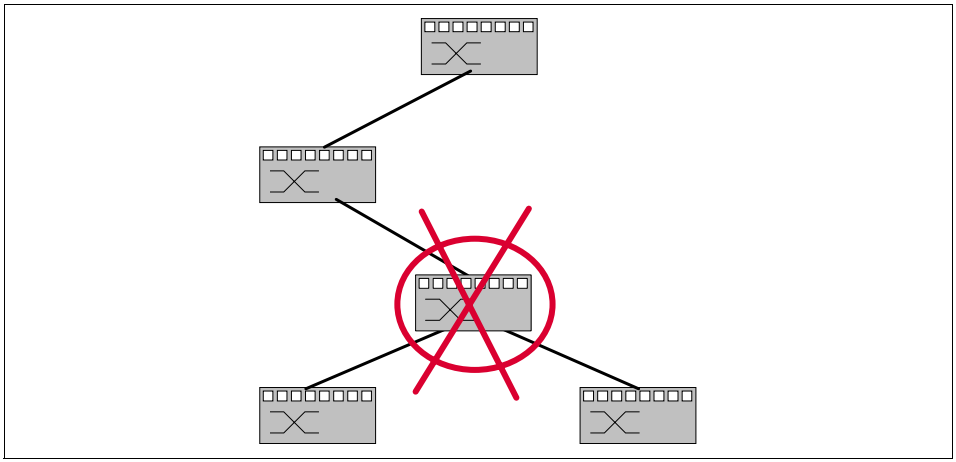


Figure 7-1 Single fabric non resilient

Here we can see that if one switch in the single fabric of the SAN fails, we will lose connection from the top to the bottom of the fabric. So if, for example, we have a server connected to the top switch which wants to access the data on the storage device connected to the bottom switch, this will not be possible in this example after the failure. We have introduced a single point-of-failure in our SAN design.

Single fabric resilient

All fabric components are connected together in a single fabric, but we do not have any single point-of-failure.

We show an example of this in Figure 7-2.

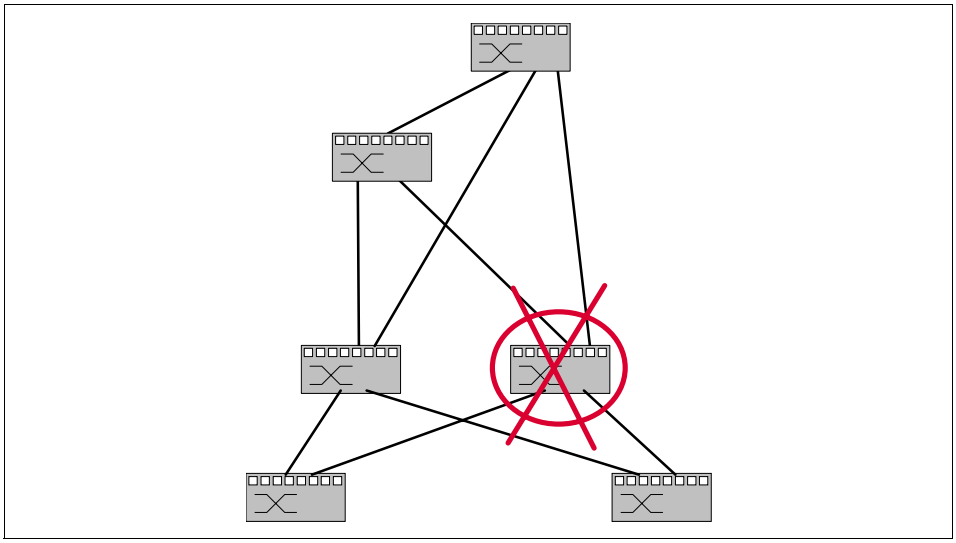


Figure 7-2 Single fabric resilient

As we can see in the example, even if one of the switches in the single fabric SAN fails, we can still access the storage devices connected to the bottom tier switches from the servers connected to the top tier of switches.

Redundant fabric non resilient

The components in the SAN are duplicated into two independent fabrics. But we still have a single point-of-failure in at least one of them. This type of design can be used in combination with dual attached servers and storage devices. This will keep the solution running even if one fabric fails.

We show an example of this in Figure 7-3.

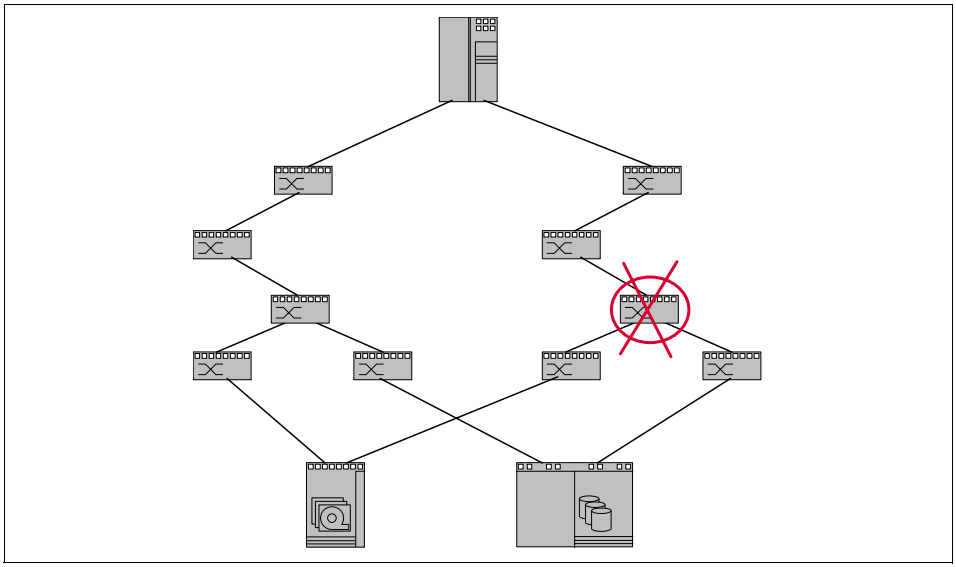


Figure 7-3 Redundant fabric non resilient

Even if one of the switches in the SAN fabric failed, we can still access the storage device in the bottom tier from the server at the top tier. Even though the fabric itself is not resilient, the data path availability is ensured through the redundant fabric.

Redundant fabric resilient

The components in the SAN are duplicated into two independent fabrics. There is no single point-of-failure in either one of them. This type of design can be used in combination with dual attached servers and storage devices. This will keep the solution running even if one complete fabric was to fail.

We show an example of this in Figure 7-4.

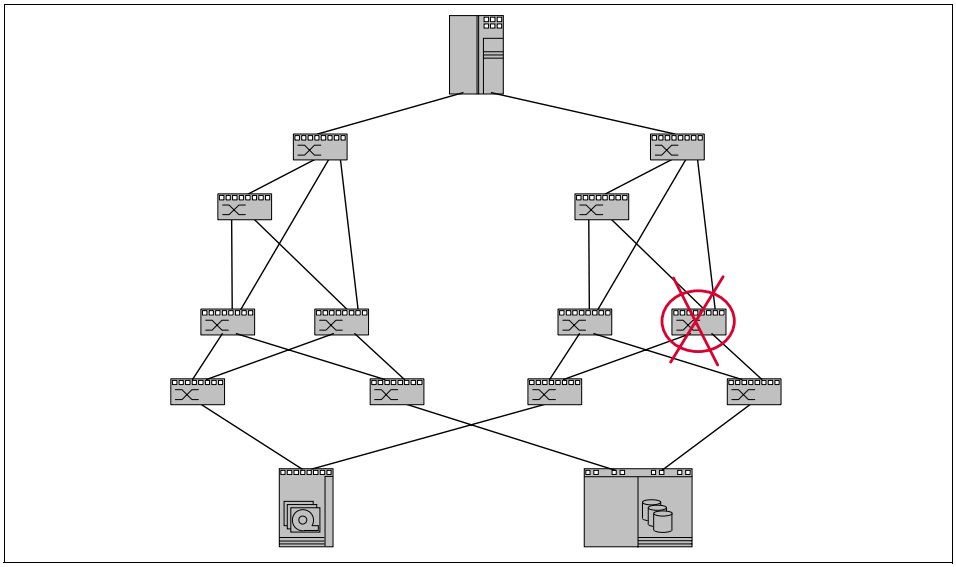


Figure 7-4 Redundant fabric resilient

Even if one of the switches in the SAN fabric failed, we can still access the storage device on the bottom tier from the server at the top tier. With this type of design we are basically protecting at two levels. First, we are protecting against switch failure and secondly, we are protecting against a failure of the whole fabric.

7.4 Host connectivity and Host Bus Adapters

IBM's supported SAN environments contain a growing selection of server Fibre Channel Host Bus Adapters (HBAs), each with their own functions and features. For the majority of open systems platforms, this presents us with the opportunity to select the most suitable card to meet the requirements for the SAN design.

For some open systems platforms, the supported HBA is provided by the vendor. In most cases, the HBAs used by the vendor are manufactured by one of the main HBA providers detailed in this section. For example, the HBA FC 6239 supported for IBM pSeries servers is supplied by Emulex.

7.4.1 Selection criterion

There are a number of points that need to be considered when selecting the right HBA to meet your requirements. In this section we look at a number of points that should be considered.

IBM supported HBAs

The first, and most important factor to consider when selecting a Fibre Channel HBA, is if it is supported by IBM for the server make and model and also the manner in which you intend to implement the server. For example, an HBA may be supported for the required server, but if you require dual pathing or the server to be clustered, the same HBA may no longer be supported.

To ensure the HBA is supported by IBM in the configuration you require, refer to:

http://www.storage.ibm.com/hardsoft/products/ess/supserver.htm#6

For IBMers only, an HBA not detailed as supported for a specific platform, support can be requested using the Request for Price Quotation (RPQ) process.

Special features

Any special functions you require from your SAN need to be considered, as not all HBAs may support the function. These functions could include:

- Dual connection
- Performing an external server boot
- Connection to mixed storage vendors
- Fault diagnostics
- Persistent binding, especially for platforms which does not support this in operating system (e.g. Windows)

Quantity of servers

Another factor to consider is the number of servers in your environment that will require Fibre Channel HBAs. Having a common set of HBAs throughout your SAN environment has a number of advantages.

- Easier to maintain the same level of firmware for all HBAs
- The process for downloading and updating firmware will be consistent
- Firmware and device driver can be a site standard
- Any special BIOS settings can be site standard
- Fault diagnostics will be consistent
- Error support will be from a single vendor

7.4.2 Multipathing software

Another important component in any SAN design is multipathing software. As the tendency is to design SANs to be redundant in functionality, we come across multiple paths to the same storage space.

For example, if we have more HBAs in the server or more storage ports connected to the fabric in our SAN, the server will see the same LUN in storage devices multiple times.

We show an example of this in Figure 7-5.

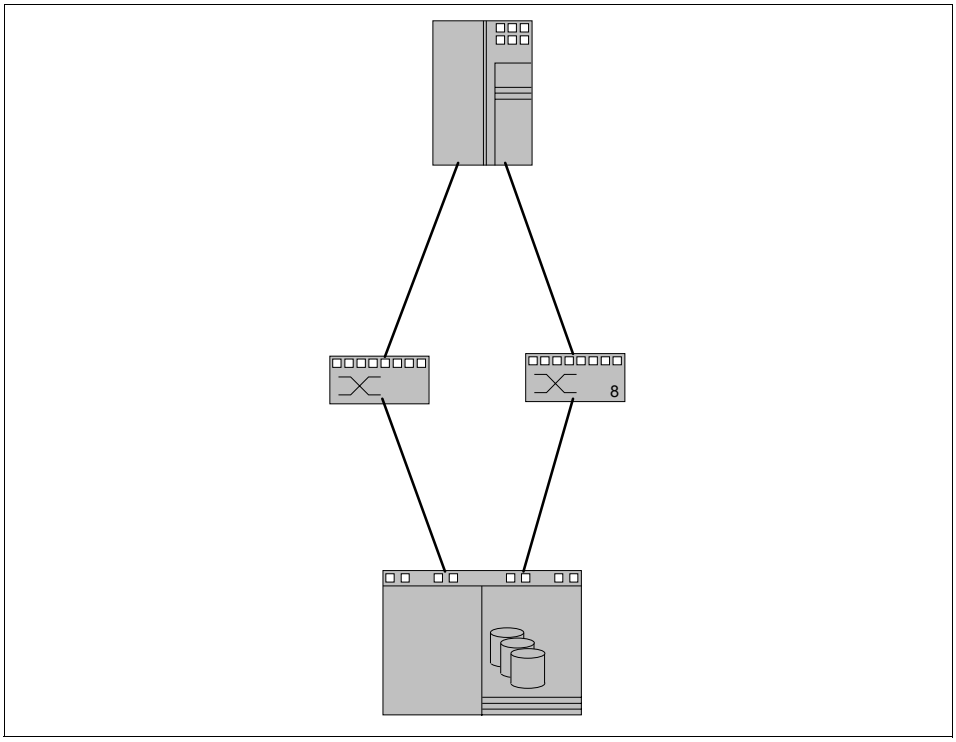


Figure 7-5 Multiple paths to the same LUN

As shown in Figure 7-5, we have the potential to arrive at the storage device using two paths.

Note: Usually the storage devices with Fibre Channel ports are setup in such a manner that all LUNs are accessible through all ports. This setup gives the highest possible redundancy, because in the case of a port failure, LUNs are accessible through other ports.

So, in our example we have two paths, because we are using two fabrics in our SAN.

However, you will also get two paths in the example shown in Figure 7-6.

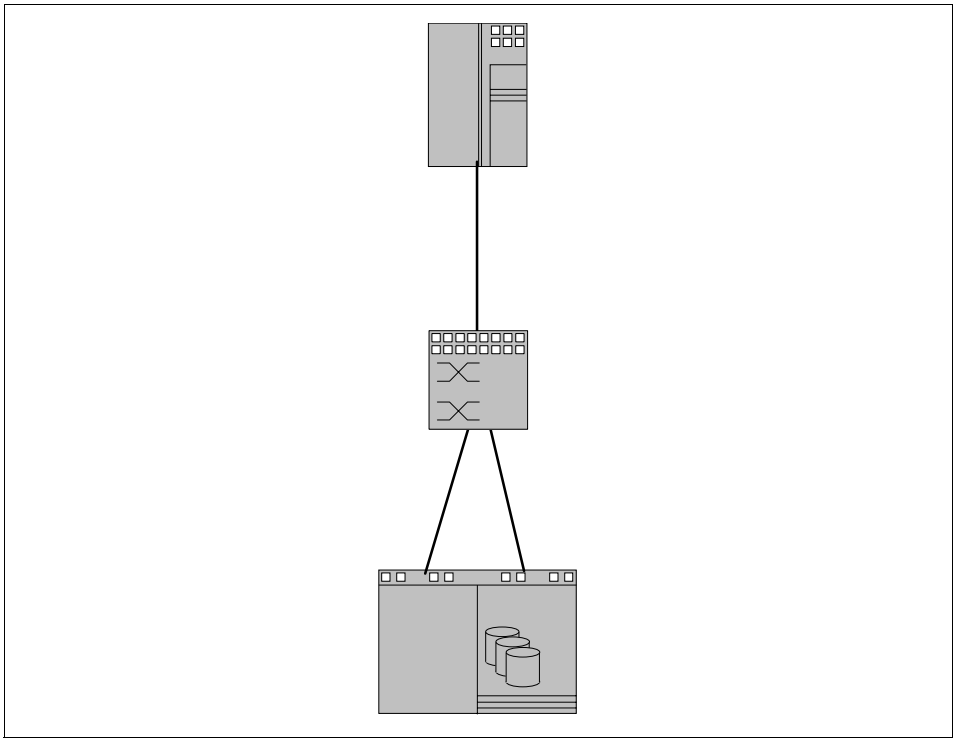


Figure 7-6 Multipath in single fabric SAN

So, if we have multiple paths to the storage, is that enough? Maybe not. We should ask ourselves these questions: Is this really enough? Is the operating system aware of multiple paths, and will it load balance and failover between the paths?

Almost all current operating systems do not have the ability to handle this at a system level. Because of this, additional software is needed.

Tip: AIX 5.2 with ML2 introduced native multipathing support with MPIO.

When you are selecting multipathing software you need to ascertain if the software supports your environment. This means it has to be supported with your server's operating systems and storage devices. Some Logical Volume Managers (LVM) already include this feature. Also, some storage and HBA vendors provide this software along with their products.

IBM provide Subsystem Device Driver (SDD), available for free with IBM TotalStorage Enterprise Storage Server and SAN Volume Controller, and supported on multiple platforms:

- AIX
- AIX with MPIO
- Windows NT/2000/2003
- Linux
- Solaris
- HP-UX
- NetWare

SDD can be downloaded from:

http://www-1.ibm.com/servers/storage/support/software/sdd.html

7.4.3 Storage sizing

SANs are primarily used to access and share storage resources from servers and also for storage device to device communication. Two of the most important characteristics of SANs are its speed and the ability to share. As we can simply design the SAN to allow device sharing, it is harder to design a SAN which will also offer high speed access to the data.

On the other hand, it is not recommended to create a high speed SAN which is then connected to a storage device which cannot fulfill the speed which is delivered to it. Because of this reason your storage device must be correctly sized for the SAN. This means having enough storage ports for data access, but again there is no sense in having more ports than the capacity of the storage device itself. Sometimes it is recommended to have more ports than the theoretical value, because it is a known fact that ports are not necessarily achieving 100% of their theoretical bandwidth.

7.4.4 Management software

When designing your SAN you should also consider how you will provide for a management solution for your design. It is important that when you are building a storage network infrastructure that you also implement management disciplines. Usually all fabric component vendors provide management software for their components. This software also offers you the ability of managing more than one device in the same package. With this you have a single interface if you are using components from the same vendor. Those packaged tools do not usually provide some of the more advanced functions such as topology views and asset management though.

For this type of management you should consider specialized tools. Some of these software suites also offer additional functionality to SAN components such as:

- LUN masking on the server side
- Storage allocation
- Policy based management

7.5 Director class or switch technology

For the purpose of starting this discussion, we will assume that three SAN experts have been discussing the *perfect* SAN solution if you needed 16 ports in a redundant fabric.

Using their combined intelligence and many years of wisdom they came up with the proposed solutions as shown in Figure 7-7.

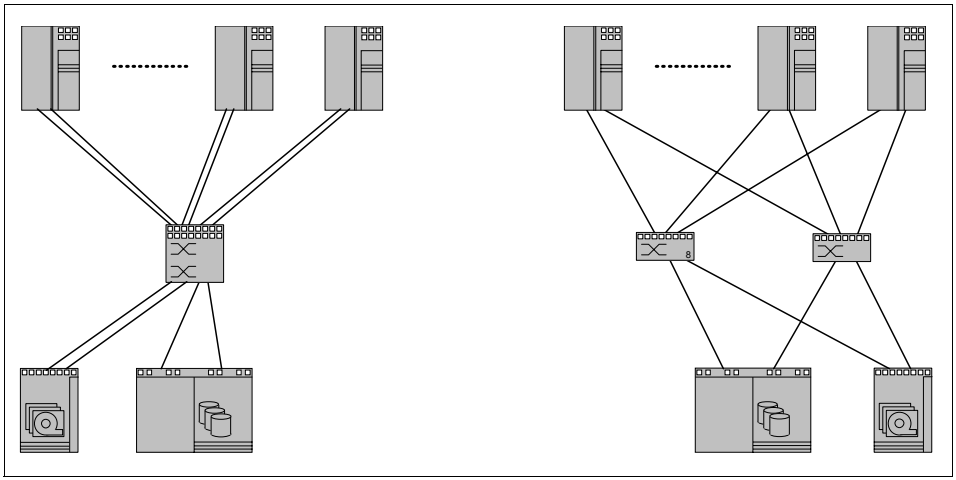


Figure 7-7 Director class or switch dilemma

In this simple scenario, we will try to outline the important parameters which can affect the SAN design you will ultimately choose. Let us assume that we identified the amount and the type of fabric ports as described in Chapter 7.2, “Existing resources needs and planned growth” on page 247, and that you know the bandwidth needed for each of the servers. In our example, we assume that we have up to sixteen fabric ports in the near future, each device with redundant connections.

In Figure 7-7, we show two sample designs that answer these requirements. On the left side we show a design with a director class product, and on the right side we show the same solution with switch technology.

Let us now try to outline the important factors which can be used to determine which solution should we use:

1. A director class product is a highly available, redundant fabric component. It has duplicated almost all its components. Switches typically only duplicate power supplies and fans. By definition, the director class product has 99.999% availability. On the other hand the switch has 99.9% availability.

   Note: Be aware that we are making a distinction between a *switch* and a *core switch* in this redbook. A core switch, such as the M12 provides 99.999% availability.

2. All director class products available today have a single backplane, which represents a single point-of-failure in the SAN or fabric.

3. By using switches as a design component in our design we are building a redundant SAN, thus avoiding a single point-of-failure. Even though we have *only* 99.9% availability, the whole SAN availability might be higher because we do not have a single point-of-failure. With such an approach you can build a 99.999% available SAN using this redundant fabric approach.
4. Director class products, and some switches, offer non-disruptive upgrades of their firmware, which means that you are not losing the connection due to a firmware upgrade. The traffic will only be blocked for a short time, and is not causing any disconnection from the device side.
5. In the scenario we used for the switch solution we can upgrade the firmware of each switch separately. If we need to restart the switch so that new firmware will become active, we will lose half of the connections. This restart can take up to one minute. This could cause time-outs, but as we are using redundant paths only performance will be degraded. If you manage the SAN correctly, you can avoid time-outs during firmware upgrades. But on the other hand, with such a design, you have the option to test new firmware on one switch, and if something goes wrong with the new firmware, you still have a working switch with the old firmware. Some of the director class products with partitioning support allow you to upgrade their firmware at the level of a single partition.
6. With the director class product in our example, we have only one physical device to manage, compared to managing two switches in the second example. For example, if the connectivity failure is in the outband management network, we need to connect to each switch separately on the serial port to perform tasks. This delivers an unwanted management overhead. In the case of the director class product, we would only need to do this once.
7. Some vendors use an embedded Web server as an option for managing the components, as opposed to specialized management tools from other vendors. Usually it takes more time to download the management applet from the fabric component than just to collect the data and process this data locally with the dedicated management software. This is especially noticeable when you have a lot of devices as you will need to open a lot of browser sessions in this case.
8. Some director and switch products can only be managed with a remote management station (that is to say, an additional Personal Computer with management software). In a case where this component fails you will be unable to perform management tasks. As opposed to other implementations where the management software is built into the device, thus allowing you to manage the device without any additional hardware components.

We show an example of managing the fabric components via external management hardware in Figure 7-8.

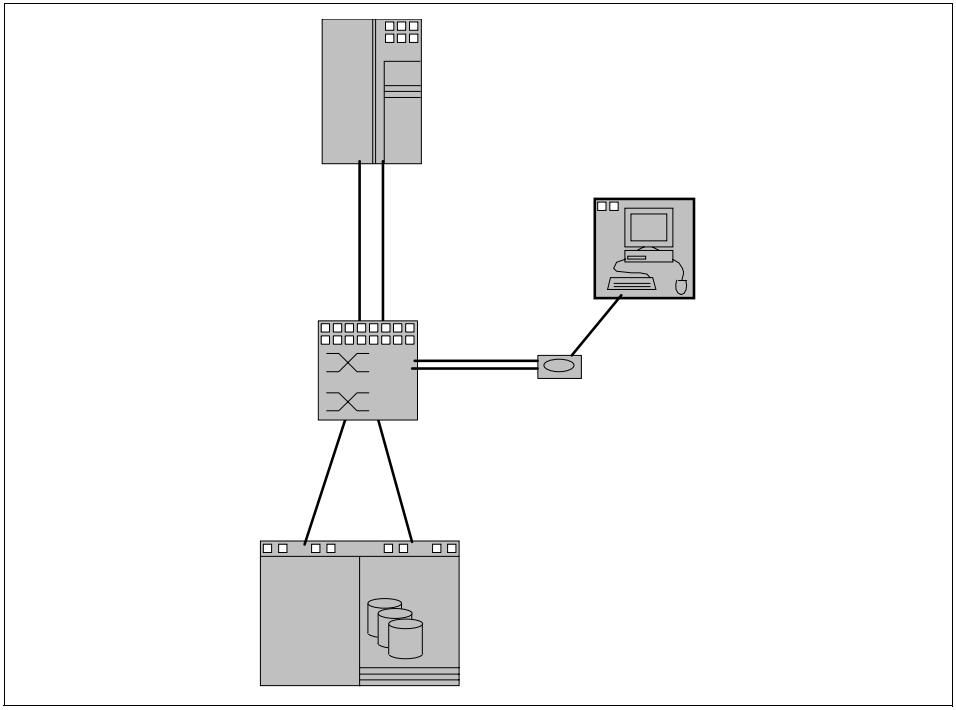


Figure 7-8 External managing of director class product

9. For products that use the ethernet network for management, in a case that this interface or connection to the component fails, the component will still perform its functions. You can still manage the fabric components via the serial port interface.

10. Director class products are usually designed to have what is often referred to as *port blades*. This means that the ports are grouped together on the same blade and because of this we need to be careful how we connect the devices to the ports. We should always distribute the connection from the same device to different port blades. With such a setup, the device will still have connection to other devices via an alternate path in a case that the blade fails.

We show an example of a good cabling practice in Figure 7-9.

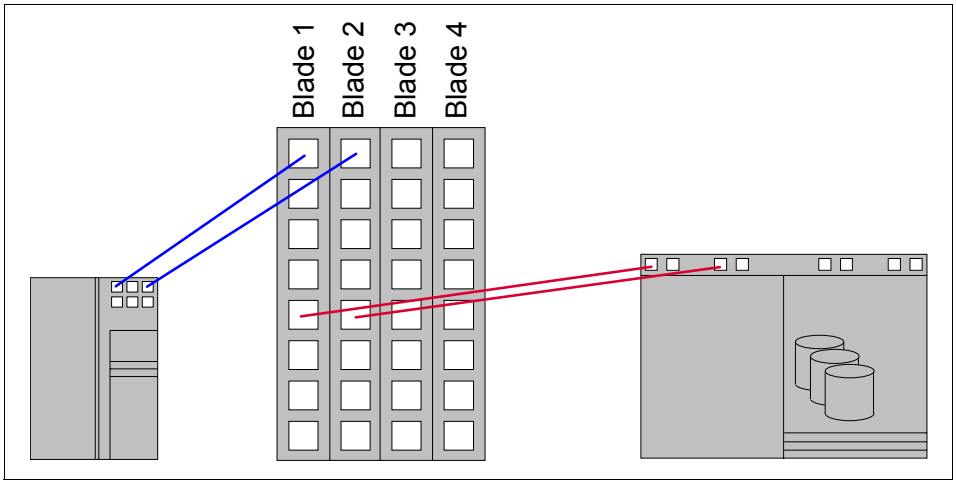


Figure 7-9 Ports on different blades

11. When you are using director class products in a similar manner as that shown in Figure 7-9, each server to storage connection has more paths to the storage device. In this example, we have two paths per connection, which means four paths per server.

    We show the possible routes in Figure 7-10.

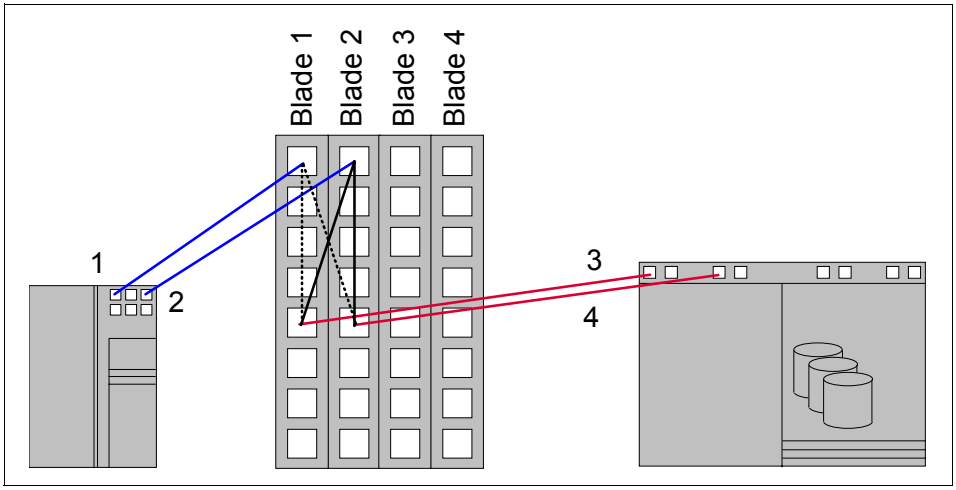


Figure 7-10 Routes in director class product

For a situation like this you would use some kind of multipathing software on the server side. By using multipathing software in this case, you are relying on it to ensure that the load will be balanced correctly between paths. For example, it could happen that multipathing software will use path 1 to 4 and path 2 to 4 at the same time, and this could cause congestion on link 4.

One of the possible solutions for this case (not necessarily the best one) would be to zone at the blade level as shown in Figure 7-11.

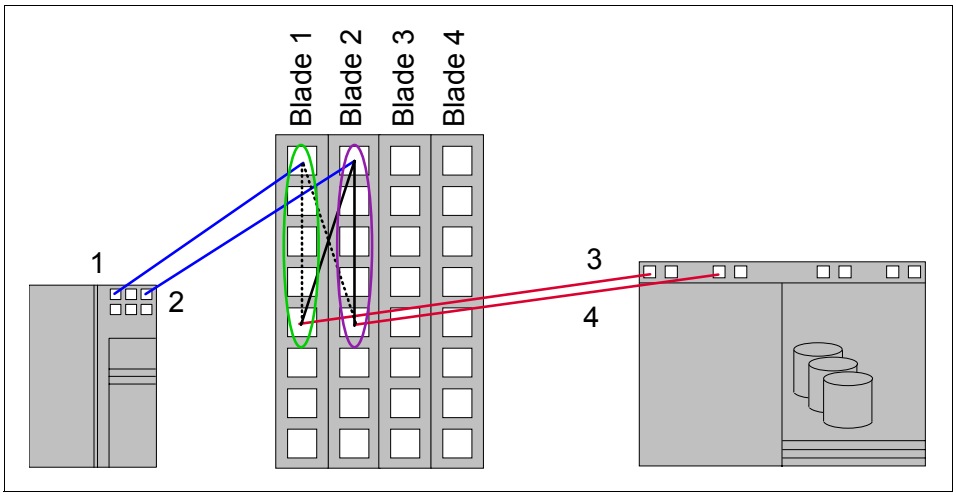


Figure 7-11 Blade zoning

With this type of zoning you are basically creating separate switches inside the director class product. With this zoning implemented, we now have only two paths from the server to the storage device. But in the case that path 4 and 2 fails at the same time, you will lose connection *because* of the zoning. Also with zoning you are introducing overhead on the communication path. This would be a reason not to use zoning in such a situation.

If the multipathing software is not capable of handling such a situation correctly, the better solution would be to use two separate fabric devices instead of using zoning inside the director class product, because you are avoiding a single point-of-failure, which exists because director class products have a single backplane.

12. Director class products usually have an option for *in chassis* expansion to a high number of ports (up to 256 ports today). This means that in case more fabric ports are needed, you are not introducing any additional components into the SAN.

In this case, you are also not introducing any new management objects into your SAN. With all the ports in one device you are also not introducing ISL links and possibly over-subscription. This point is valid only if the director class product truly has a non-blocking design.

Because of port expansion options those products are usually higher priced, so initial investment will be much higher than investment in switches. If you want to build the SAN with a high port count using switches, which usually have up to 32 ports, you would introduce a lot of other considerations:

- Cabling
- More IP addresses to manage the SAN
- More usernames/password (you can also use one for all of them, but this may raise security issues; more usernames and passwords may cause administrative problems)
- A complex to understand topology
- More power and heat
- More devices to monitor
- In the case of Web management, you can suddenly have a lot of open browser windows.
- You need to maintain more devices (firmware upgrades for example)
- Ultimately, limited growth capability

In Figure 7-12, we show the comparison between a 64 port director class product and a 72 port full meshed fabric built from 16 port switches.

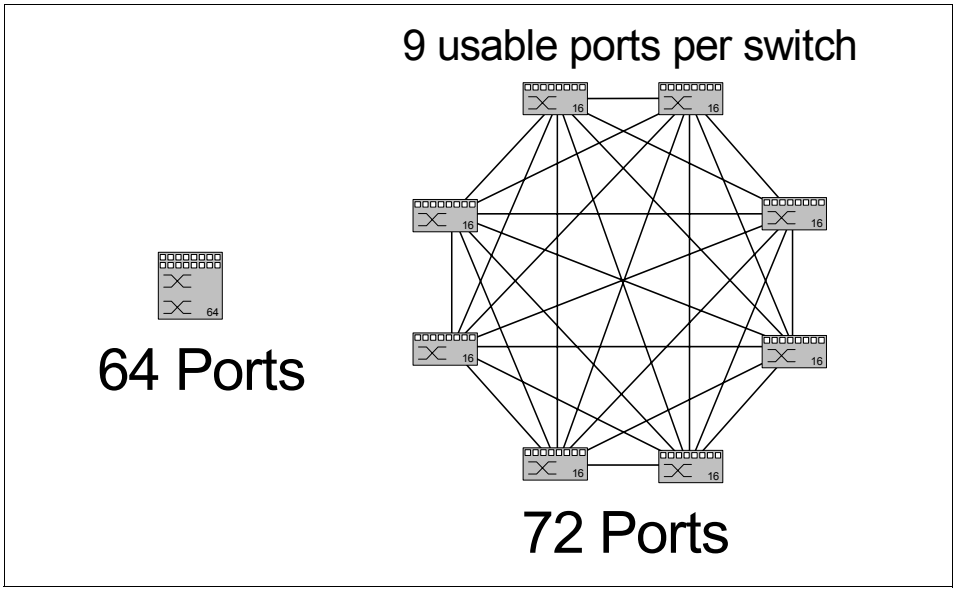


Figure 7-12 Director class product versus full meshed switch fabric

As you can see in Figure 7-12, we have a 9:7 over-subscription ratio on each switch. The other important issue in such a setup will arise when you want to have communication from 9 ports on one switch to nine ports on the other switch. Because of the FSPF behavior explained in Chapter 3.7.2, “Fabric Shortest Path First” on page 89, all communication will go through one ISL, and this gives a 9:1 over-subscription ratio.

With the director, there are no ISLs and hence no over subscription within the fabric itself.

In Figure 7-13, we show an example of how to provide 64 ports using six 16-port switches.

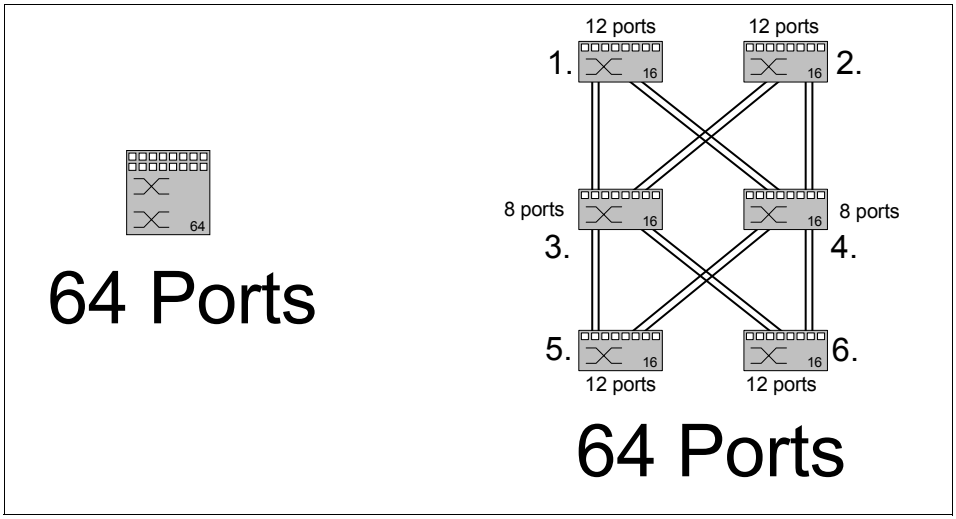


Figure 7-13 Director class 64 ports against 64 ports switch fabric

In this example, if all ports are 100% utilized, we have a 12:4 (3:1) over-subscription ratio on each of the outer switches and no over-subscription on the inner switches. We have these possible communication channels:

- From devices on switch 1 and/or 2 to devices on switch 5 and/or 6 we have 3:1 over-subscription. As we can see from the picture, the traffic is going across the middle switches. Because of this we are introducing two hops as opposed to port to port transfers in the director class product. Each hop has up to 2 microseconds delay. So, in our case, we have up to 4 microseconds delay in communication.
- From devices on Switch 1 and/or 5 to devices on Switch 2 and/or 6, we have the same situation as in the previous case.
- From devices on Switch 1 and/or 5 to Switch 3 we have a 12:2 (6: 1) over-subscription ratio and we have one hop, or approximately 2 microseconds delay in communication.
- From devices on switch 2 and/or 6 to Switch 4 we have the same situation as in the previous case.
- From devices on Switch 3 to devices on Switch 4 we do not have over-subscription, but we have a two hop delay as in the first case.

As we can see from these two simple examples, it requires more design and planning to replace a director class product with any type of meshed fabric. And by introducing ISLs, we are also increasing the price per available device port.

Note: When creating a meshed fabric be aware that the number of available ports will not be the number of all ports from all switches. It will be smaller by the number of ISLs multiplied by two.

13. Some of the director class products and switches do not support FC-AL connections. In this case we need to introduce “edge switches” which allow FC-AL connections and are then connected to the components which do not support FC-AL connections. With this we are introducing another layer of complexity. So, in some cases the need for FC-AL connections would suggest that we use director class products or switches which are capable of supporting this type of connection.

    In Figure 7-14, we show such a solution with the introduction of loop switches to support FC-AL storage devices.

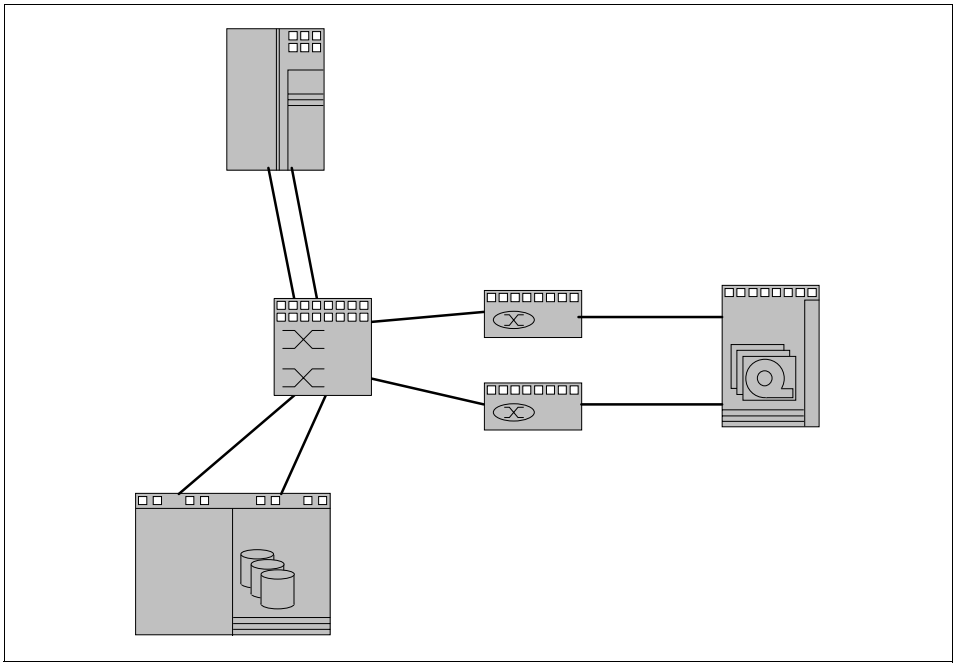


Figure 7-14 Adding tapes to the director class SAN

In the example in Figure 7-14, we are assuming that the tape device is only FC-AL capable. Most native tape attachments are FC-AL, although newer tape devices such as IBM 3592 are FC-SW capable, If you connect legacy SCSI tape devices over a Fibre Channel gateway you have the opportunity to configure it as a FC-SW type of fibre device. In such a case, you do not need the loop switch to attach tape drives.

The solution with the switches supporting FC-AL is much more elegant in such a case.

We show this solution in Figure 7-15.

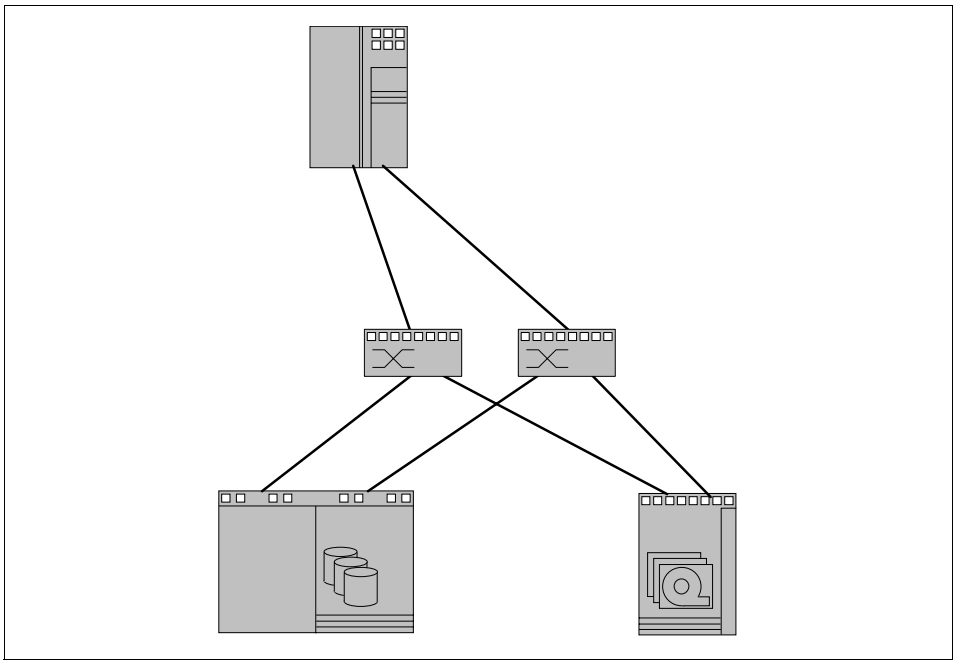


Figure 7-15 Switches with loop support

There is also another reason where you would want to use separate edge switches for connecting FC-AL devices. Storage FC-AL devices are usually low bandwidth type of devices. So, when you are implementing director class solutions, it makes sense to consolidate more of such storage devices on the edge switch, which is lower priced per port, and it only consumes one higher priced director class port.

14. Director class products also offer *call home* and *call in* functions which allows the device, in event of failure, to automatically call the vendor support center. Then, remote access can be used to fix the problem if it is possible, or if not, other actions can be taken without any end user intervention.
15. Director class products have almost all their components hot pluggable. This means that they can be replaced “on the fly” during operation. This is not valid just for fans, power supplies, but also for the blades with switch ports on it. As opposed to switches where usually only the power supplies are hot pluggable.
16. In Figure 7-16, we show an example of implementing an ISL between two redundant fabrics.

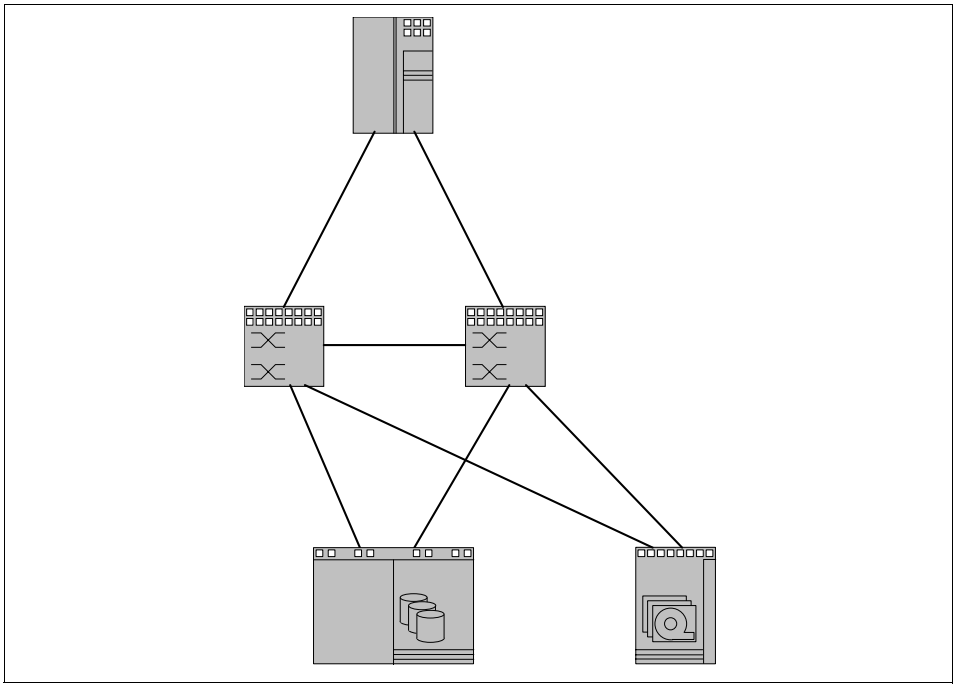


Figure 7-16 ISL between two redundant fabrics

There are some considerations to be taken into account in this scenario. In this case you will have four paths from server to storage side and because your multipathing software will not know that two paths are slower, it could happen that it will “round robin” the data across the ISLs. The only difference in a similar case in the director class product is that here it is safe to zone the paths to the disk device and leave paths to the tape device open for additional reliability.

So let us now review the possible solutions for the example we gave at the beginning of the chapter.

Design with two director class products supporting FC-AL

We show this design in Figure 7-17.

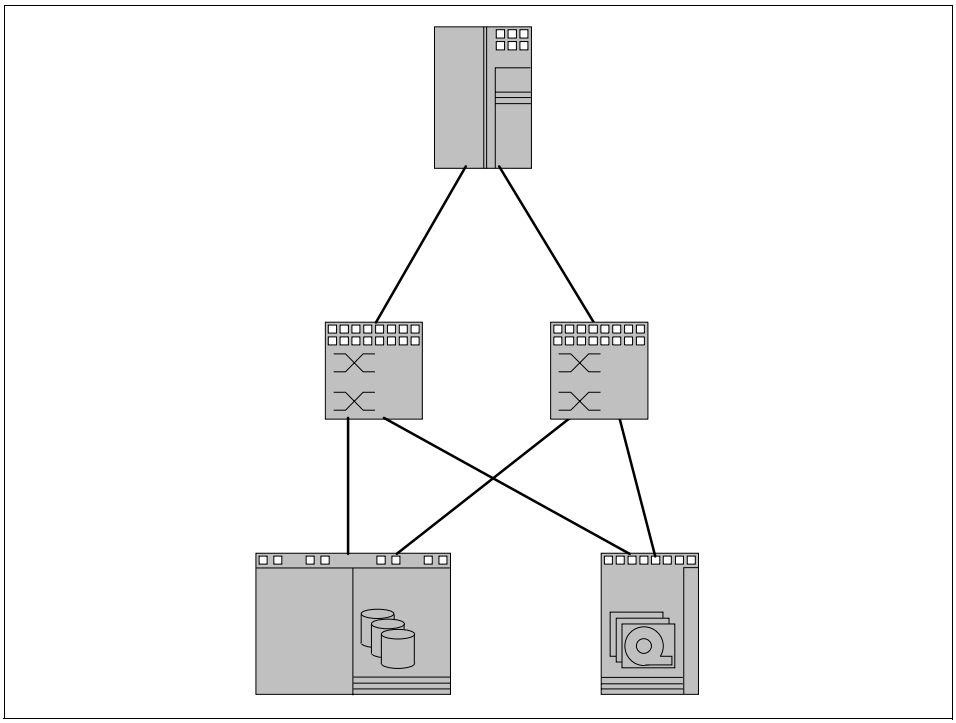


Figure 7-17 Two director class products with FC-AL support solution

This would be the preferred solution if you do not have any price limits. The design has the following attributes:

- Redundant SAN design
- 99.999% available components
- Support for FC-AL devices on all ports
- No single point-of-failure in the SAN
- Single point-of-failure in the fabric components
- Highly scalable solution, allows you to add new ports without any intervention into existing connections

Design with two directors without FC-AL support

We show this design in Figure 7-18.

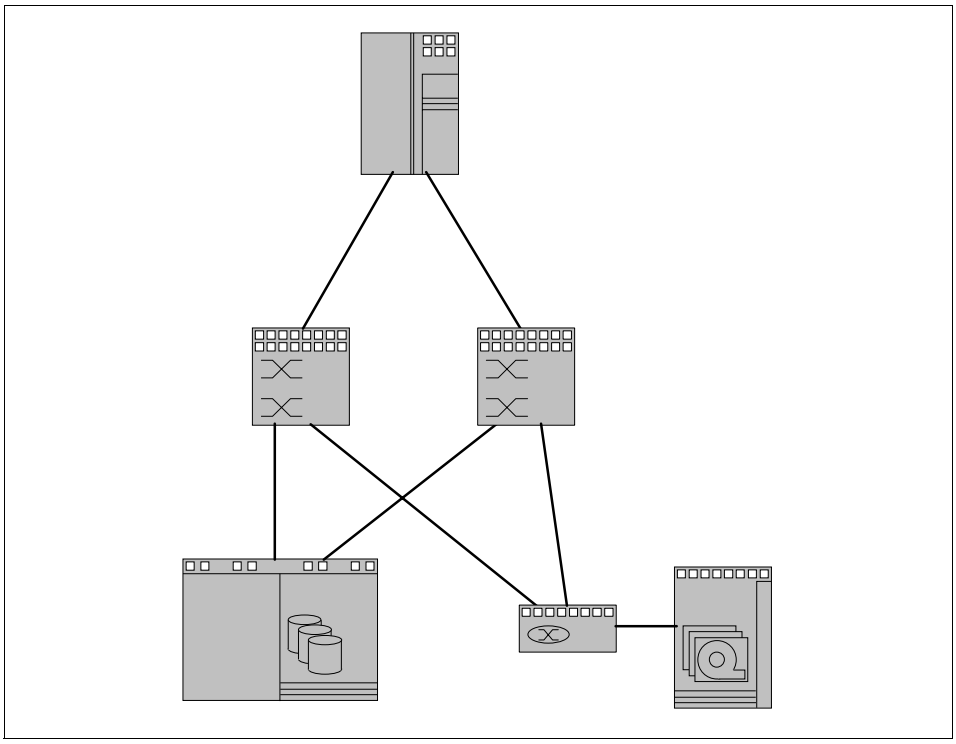


Figure 7-18 Two director class products without FC-AL support

This would be the preferred solution if you do not have any price limits, but the selected vendor of your director class product does not support FC-AL connections, and you want to add more FC-AL devices in the future at no extra cost. The design has the following attributes:

- Redundant SAN design
- 99.999% available components except for edge switch which is a 99.9% available component
- Support for FC-AL devices on the edge switch
- Single point-of-failure on the edge switch. You can overcome this by using two edge switches. But this will take four ports on the director class product, so it will be only reasonable if you would have more than two FC-AL devices in your SAN.

- A highly scalable solution which allows you to add new ports without any intervention into the existing connections. You also have a point for adding new FC-AL devices, because these type of edge switches usually have at least eight ports, without consuming expensive director class ports. Some of the products available on the market today allow only one uplink connection to the director class product, and this means that you are introducing a single point-of-failure into the fabric.

  An example of such a solution is shown in Figure 7-19.

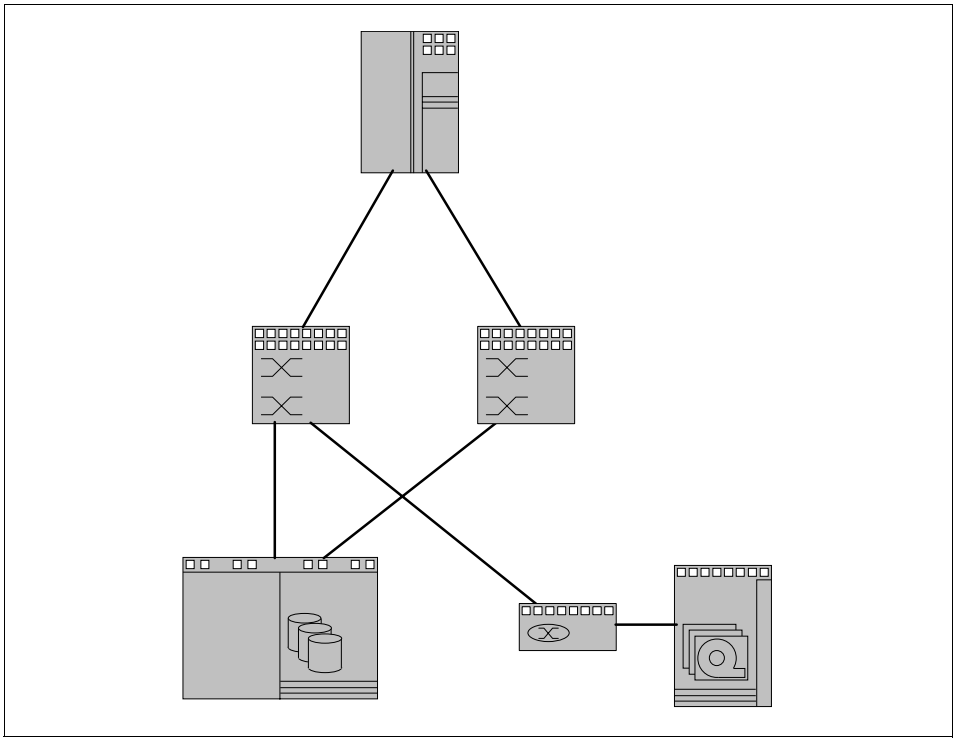


Figure 7-19 Edge switch with only one connection

Design with one director class product without FC-AL support

We show this design in Figure 7-20.

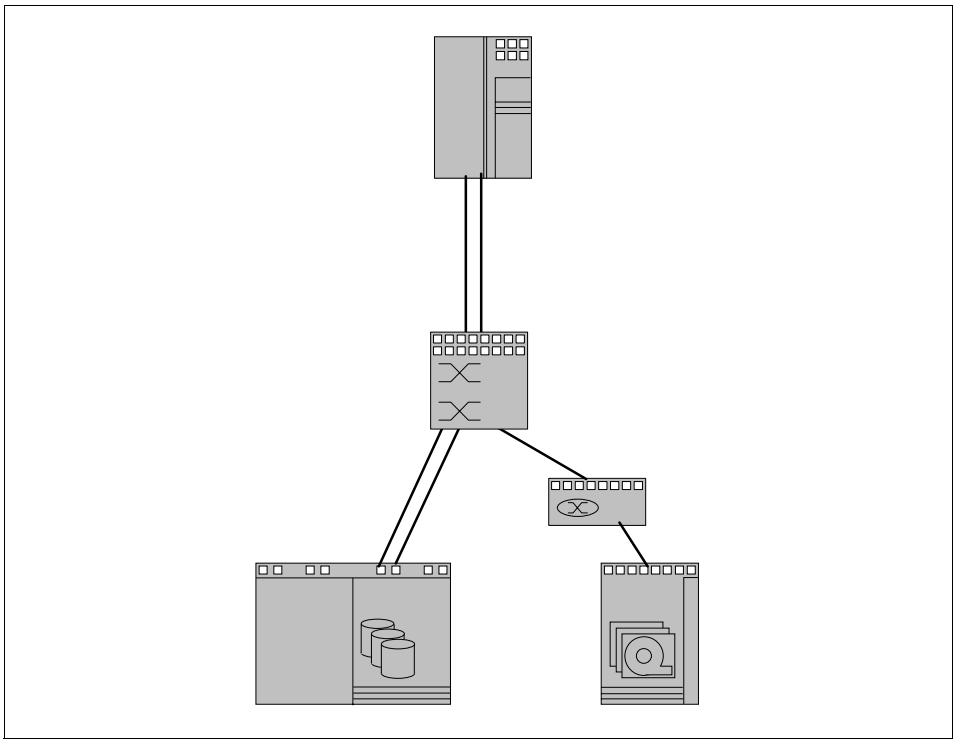


Figure 7-20 One director class product without FC-AL support

This would be the preferred solution if you do not have any price limits and you think that a single backplane is not an issue in your design. With this solution you can also add more FC-AL devices in the future for no extra cost. The design has the following attributes:

- Single point-of-failure in the SAN
- 99.9% and 99.999% available components
- Support for FC-AL devices on edge switch ports
- Single point-of-failure in the fabric components
- Highly scalable solution which allows you to add new ports without any intervention into existing connections. It also allows you to add new FC-AL devices without consuming expensive director class ports.

Design with two switches with FC-AL support

We show this design in Figure 7-21.

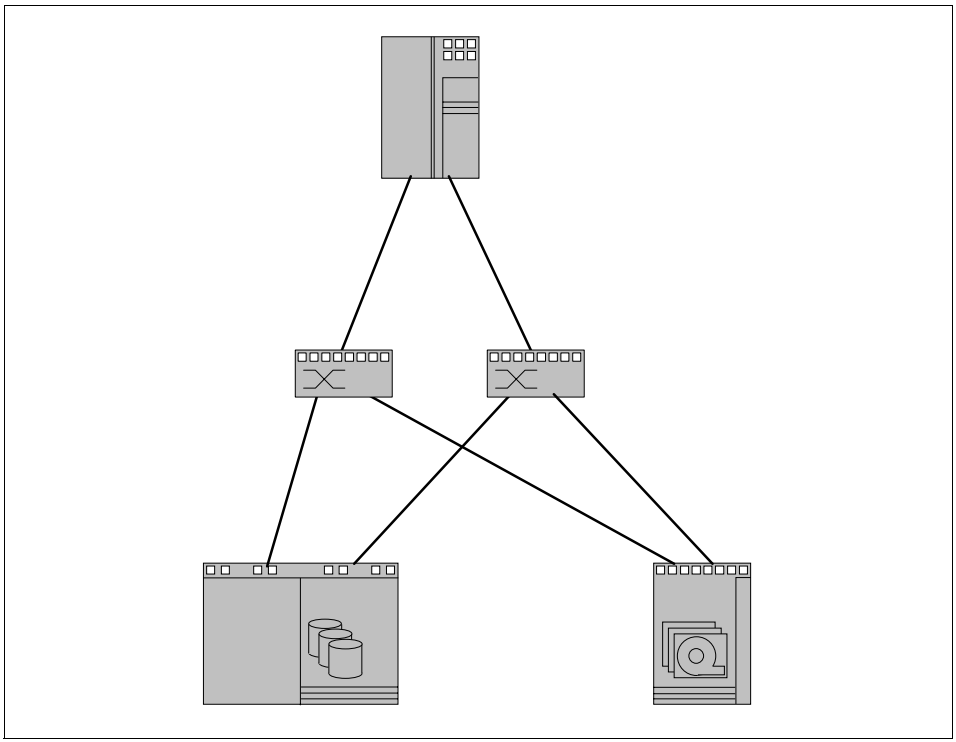


Figure 7-21 Two switch solution with FC-AL support

This would be the preferred solution if you have a price limit. With a two switch solution you are building a redundant SAN. But, because the switch port count is smaller than the director class port count, you can introduce scalability problems in the future. The design has the following attributes:

- No single point-of-failure in the SAN
- 99.9% available components
- Support for FC-AL devices on all ports
- Single point-of-failure in the fabric components
- Not such a highly scalable solution because switches can go up to 32 ports today. In the case of adding additional switches you can come to a situation where you need to re-cable the infrastructure, and this can cause operational delays or degradation if you stay with redundant SAN design

Design with two switches without FC-AL support

We show this design in Figure 7-22.

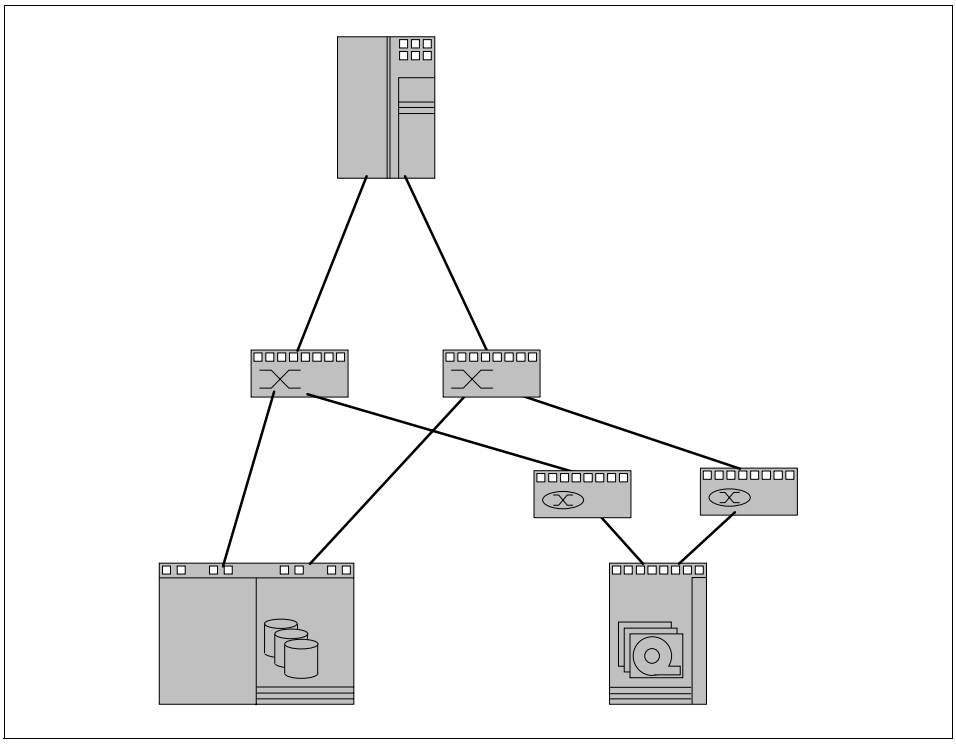


Figure 7-22 Two switch solution without FC-AL support

This would be the preferred solution only if you already have switches which do not support FC-AL devices, and you are building a redundant SAN. But because the switch port count is smaller than the director class port count, you can arrive at scalability problems in the future. The design has the following attributes:

- No single point-of-failure in the SAN
- 99.9% available components
- Support for FC-AL devices on edge switch ports
- Single point-of-failure in the fabric components
- Not a highly scalable solution because switches can go up to 32 ports today. In the case of adding additional switches you can come to a situation where you need to re-cable the infrastructure and this can cause operational delays or degradation if you stay with redundant SAN design
- You are introducing additional levels of complexity with edge switches

Design with single switch with FC-AL support

We show this design in Figure 7-23.

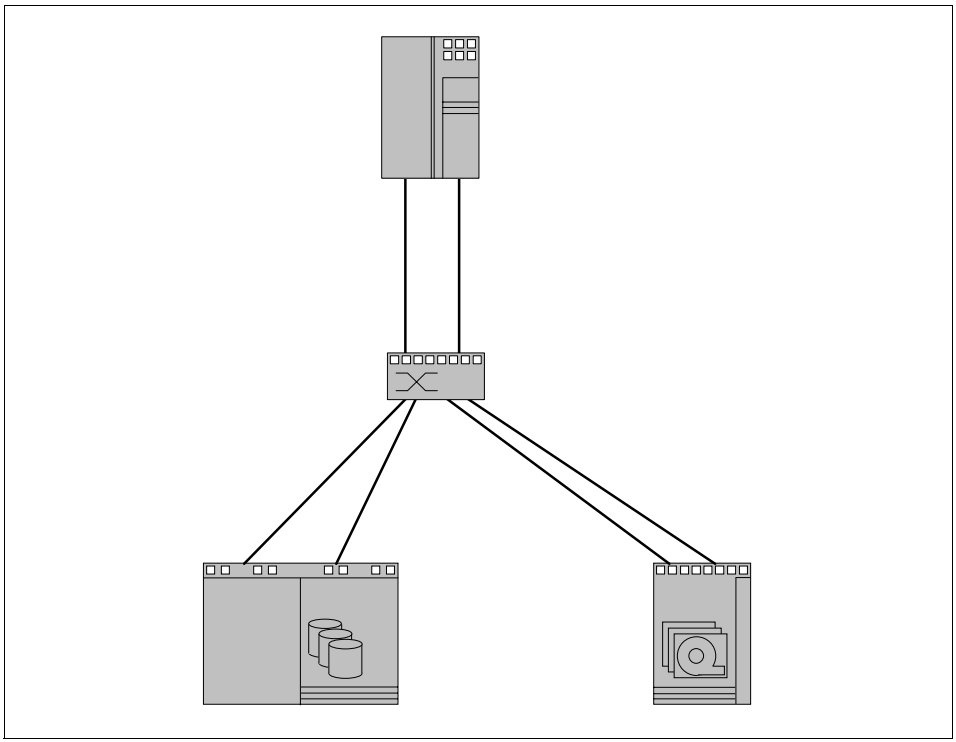


Figure 7-23 Single switch solution with FC-AL support

This type of design should be avoided because you are using a 99.9% available component with a single point-of-failure in the SAN. Also, this is not a very scalable solution. The design has the following attributes:

- Single point-of-failure in the SAN
- 99.9% available components
- Support for FC-AL devices on all ports
- Single point-of-failure in the fabric components
- Not a highly scalable solution because switches can go up to 32 ports today. In the case of adding additional switches you can come to a situation where you need to re-cable the infrastructure and this can cause operational delays or degradation

7.6 General considerations

In the topics that follow we will overview some of those considerations that tend to be ignored when piecing together a solution. Typically, they are viewed as trivial in comparison to some of the major considerations. For example: “Do I choose a switch or a director?” However, a well-designed SAN will have taken these into account.

7.6.1 Ports and ASICs

You can improve performance on some switch products by routing data from storage to host server using ports on the same ASIC. This reduction in switch latency varies between products, and you should refer to the relevant chapter later in the book for specific details.

Of course, all switch or director vendors will claim that the delay in the switch is insignificant compared to the delay caused by the storage devices and servers. Nevertheless, it can still be valuable to understand how ports are grouped within ASICs and use this information to optimize the SAN design.

7.6.2 Class F

When you design SANs with ISLs you should also incorporate some bandwidth requirements on the ISL for Class F traffic. This traffic is minimal when compared to the traffic used for data transfer. So if you size your ISLs to be almost at maximum capacity, for example more than 90MB/s out of 100MB/s, you should start to consider adding additional ISLs because there will be some bandwidth overhead for Class F traffic.

Class F Service is defined in the FC-SW and FC-SW2 standard for use by switches communicating through ISLs. It is a connectionless service with notification of non delivery between E_Ports, used for control, coordination and configuration of the fabric. Class F is similar to Class 2 since it is a connectionless service, the main difference is that Class 2 deals with N_Ports sending data frames, while Class F is used by E_Ports for control and management of the fabric.

7.6.3 Domain IDs

The number of available domain IDs in the fabric should also be a consideration point. Different vendors support different numbers of switches/directors which can be interconnected in the same fabric. You should refer to the relevant product documentation later in this book for specific numbers, as the number of available domain IDs could limit you in your SAN design.

7.6.4 Zoning

All fabrics should implement zoning. Zoning allows you to partition your SAN into logical groupings of devices that can access each other. Using zoning, you can arrange fabric-connected devices into logical groups, or zones, over the physical configuration of the fabric.

Zones can be configured dynamically. They can vary in size depending on the number of fabric connected devices, and devices can belong to more than one zone. Because zone members can access only other members of the same zone, a device not included in a zone is not available to members of that zone. Therefore, you can use zones to:

- Administer security: Use zones to provide controlled access to fabric segments and to establish barriers between operating environments. For example, isolate systems with different uses or protect systems in a heterogeneous environment.
- Customize environments: Use zones to create logical subsets of the fabric to accommodate closed user groups or to create functional areas within the fabric. For example, include selected devices within a zone for the exclusive use of zone members, or create separate test or maintenance areas within the fabric.
- Optimize IT resources: Use zones to consolidate equipment, logically, for IT efficiency, or to facilitate time-sensitive functions. For example, create a temporary zone to back up non-member devices.

Refer to “Zoning” on page 104 for more information on zoning in general.

Refer to the following sections for more product specific details:

- IBM TotalStorage SAN Switch - “Zoning” on page 328
- Cisco - “Zoning” on page 544
- CNT - “Zoning” on page 405
- McDATA - “Zoning” on page 477

7.6.5 Physical infrastructure and distance

The physical infrastructure can play a big role in your design. Because a SAN offers logical storage consolidation, you can also interconnect servers and storage devices which are not located in the same place. In some cases you can simply lay down cables to all resources you want to use in the SAN. But there can

also be a case where you can only have one connection from one area to another. In such a situation it can force you to design a meshed fabric, using either a director class product or switches, grouping resources in different areas and connecting those areas together in one big SAN.

For large scale implementations, fibre trunking and patch panels should be considered as part of a structured cabling solution. This provides for a tidier and more flexible installation.

With respect to distance considerations in the design, you need to pay attention to buffering. Some vendors use dedicated buffer credits for each port in the switch or director class product; others use buffer credits for a pool of ports. This means that you have to be careful when you are selecting the ports for long distance communication. Usually those ports pool buffers in a group of four, giving you the ability to use only one port out of the four of them for long distance (over 10 KM) communication.

We explain buffer credits and how they are used in 3.3, “Buffers” on page 70.

7.7 Interoperability issues in the design

In this section, we will outline some issues relating to interoperability.

Interoperability

Interoperability means how the various SAN components interact with each other. Usually all vendors have their own labs to perform interoperability testing. Before any design, it is recommended that you check with the vendor of your SAN components as to the tests that they performed, so you can input this data into your decision making.

Usually the most important part in this is the interaction with the storage vendor, because at the end of the day, you have to select the components which were certified by them, or you may not have support from them. For example, one storage vendor may certify one level of firmware for an HBA and a server vendor certifies another level.

When planning your SAN design you need to match the minimum requirements for the components used. If this is not possible you can ask the vendor to certify your solution, or go to a SAN Interoperability Lab., for example, the IGS SAN Interoperability Lab. and perform certification there.

Standards

The SAN component vendors, especially switch makers, are trying to comply to the standards which will allow them to operate together in the SAN environment. The current standard which gives the opportunity to have different components in the same fabric is the FC-SW2 standard from http://www.t11.org

This standard defines FSPF and Zoning exchange and ISL communication. Not all vendors may support the whole standard yet, so in designing today's SAN you should be very careful when trying to design a multi-vendor fabric.

The future standards will also bring with them functions which will allow the management information to be exchanged from component to component, thus giving the option to manage different vendors components with tools from one vendor.

Legacy equipment and technology

Another important consideration when you are introducing a SAN design into your current environment is how to integrate legacy equipment. For example, you could have a lot of eminently usable SCSI attached tapes and drives. By usage of routers and bridges, you can bring those devices into the SAN, to be utilized SAN wide.

The old SCSI attached storage devices, for example, can be used for test purposes, because they are probably too slow when compared to today's Fibre Channel attached devices. Most likely you will reuse the tape devices in your SAN design. When introducing bridges or routers into your design keep in mind that with this you are not getting native attached Fibre devices, but you are bridging protocols. Because you are introducing another point of complexity in the design you need to be very careful in selecting the components for this type of work, and again you must check if they comply to the standards, and that they are tested with other SAN equipment you are using.

Heterogeneous support

We have already mentioned that SAN vendors are trying to establish support for the standards which will give them the opportunity to work together in the same SAN fabric. But this is just one view of heterogeneous support. The other view is from the platforms which will participate in the SAN as the users of the resources.

So, when designing the SAN it is important that you check that the SAN components you are using are certified and tested with the platforms you plan to use in the SAN. This also mean that you need to verify which levels of operating systems are supported and can coexist in the same SAN. If you find that some of

the platforms cannot coexist in the same environment, you can still design around the same physical infrastructure for these platforms and then separate the platforms by using some kind of zoning or VSAN mechanism, which will separate the traffic of each platform from each other.

In the future, when frame filtering is a widespread feature, the function of protecting the storage at the storage level will be moved to protection at the switch level. This will give you flexibility in securing your environment.

7.7.1 Certification and support

As we already mentioned in the 7.7, “Interoperability issues in the design” on page 282, we have to take into account the certification status of the components. Usually vendors will provide information about the various components which are supported.

For example, a storage vendor can provide you with the information as to which SAN components, and at which level, are tested with its storage. If you use the components at the level they were tested at by the storage or the server vendor you will get official support.

But this is usually not enough, as you should test all of your hardware designs to ascertain if they do actually work with your applications. If there is no access to this information with respect to this being tested by a vendor (which could mean either the application or hardware vendor) you can utilize SAN Interoperability Labs to perform tests for you. Some of the labs will then also offer to support these tested configurations, even if they are not certified by the server or storage vendor.

7.7.2 OEM/IBM mixes

A lot of SAN component vendors sell their products through various channels under different names. This means that you can basically buy the same component under different brands.

There may be a case when you need to include those products into your design based around the IBM product. Usually, you can reuse those components without any problems as they are not locked or otherwise protected to work only with specific vendor equipment.

The important things to consider are if you can:

- Provide maintenance and support for those products so you can match the portfolio you are introducing

- Synchronize the software levels in the OEM devices with the levels used by IBM products
- Provide licenses for add on features to match the license level of IBM products

7.8 Pilot and test the design

It does not matter if your SAN design tends to be small or large, it is still critical to perform some kind of pilot test or acceptance test, or maybe both. If your solution is one of the standard solutions which is certified by various vendors and you can use the same releases of hardware and software in your environment, then you can be sure that the solution will also work for you without any problems. And you will probably also get support for this solution, so if something does go wrong, the vendor will help you to solve the problem.

If you are designing a solution which has not been implemented before, or there are components of the solution which are not certified with other components, you should perform a pilot test. Of course, you may not buy the same set of equipment for testing as might be planned for the production design, but you can select the most critical components and scale them down for the test.

There are also test sites such as the SAN Interoperability Labs from IGS which offer pilot testing and proof of concept services. They will give you the option to build your environment in the labs and try it out within your SAN design. It is still better to pay for the proof of concept or pilot test, than buy all the equipment and later on find out that something is not working as it should.

8

IBM TotalStorage SAN b-type family

The various models of the IBM TotalStorage SAN Switch provide Fibre Channel connectivity to a large variety of Fibre Channel-attached servers and disk storage, including the IBM TotalStorage Enterprise Storage Server, IBM TotalStorage DS Family (formerly FAStT) Storage Servers, SAN Data Gateways for attachment of IBM Enterprise Tape systems 358x, and tape subsystems with Native Fibre Channel connections.

In this chapter we provide details on these products and describe their interactions.

© Copyright IBM Corp. 2004. All rights reserved.

8.1 Product description

The IBM TotalStorage SAN Fibre Channel Switch interconnects multiple host servers with storage servers and devices to create a SAN. The switch can be used either as a standalone device to build a simple SAN fabric, or it can be interconnected with other switches to build a larger SAN fabric.

The interconnection of IBM and IBM-compatible switches and hubs creates a switched fabric containing several hundreds of Fibre Channel ports. The SAN fabric provides high performance, scalability, and fault tolerance required by the most demanding e-business applications and enterprise storage management applications, such as LAN-free backup, server-less backup, disk, and tape pooling, and data sharing.

IBM offers seven different models of switches which are OEM products from the Brocade SilkWorm family as follows:

- The IBM TotalStorage SAN Fibre Channel Switch Model 2005-B32 is a 32-port model
- The IBM TotalStorage SAN Fibre Channel Switch Model 2005-H08 is an 8-port model
- The IBM TotalStorage SAN Fibre Channel Switch Model 2005-H16 is a 16-port model
- The IBM TotalStorage SAN Fibre Channel Switch Model 2109-F16 is a 16-port model
- The IBM TotalStorage SAN Fibre Channel Switch Model 2109-F32 is a 32-port model
- The IBM TotalStorage SAN Fibre Channel Switch Model 2109-M12 is a (dual) 64-port model
- The IBM TotalStorage SAN Fibre Channel Switch Model 2109-M14 is a 128-port model

In the following sections, we describe the switches in greater detail with respect to their features, including high availability (HA), system components, zoning, Inter-Switch Links (ISLs) and performance.

8.1.1 2005-B32

The IBM TotalStorage SAN32B-2 (2005-B32) provides a 1U rack mount fabric switch, designed to provide Ports on Demand scalability with 16, 24, or 32 ports, with each port capable of throughput speeds of 1, 2, or 4 Gbps, and maximum latency of 2 microseconds. Additionally, each port offers auto sensing and self-negotiation to the highest speed (1, 2, or 4 Gbps) supported by the attached server, storage, or switch, as well as self-configuration as F_ports, FL_ports, or E_ports. The base function of the SAN32B-2 fabric switch includes Fibre

Channel fabric services, Advanced Zoning, WEB TOOLS, and Fabric Watch, as well as high availability design with redundant hot pluggable power supplies and fans, automatic path routing, and non-disruptive firmware upgrades The SAN32B-2 fabric switch comes with the first sixteen ports enabled in the base, and does not include SFPs within the base product. Short wave and/or long wave SFPs must be ordered separately, with a minimum order of 16 SFPs required on initial purchase. Customers may purchase a variety of additional port activation optional features, for a total of 24 ports, or 32 ports enabled. Short wave and/or long wave SFPs must be separately ordered for each of the enabled ports upon activation

The SAN32B-2 fabric switch supports operation with Small form-factor pluggable (SFP) optical transceivers. Operation with 4 Gbps capable short wave SFPs provides distances up to 150 meters. Operation with 2 Gbps capable long wave SFPs provides distances up to 10 kilometers, and up to 80 kilometers with extended distance SFPs.

The SAN32B-2 fabric switch supports compatibility with IBM TotalStorage SAN b-type switch family (IBM 2005, 2109 and 3534).

8.1.2 2005-H08 and 2005-H16

The IBM TotalStorage SAN Switch H08 and IBM TotalStorage SAN Switch H16delivers next generation performance and functionality for Fibre Channel Storage Area Networks (SAN), comprising 1 and 2 Gb/s auto-sensing capability with fully non-blocking performance and designed to be completely interoperable with other members of IBM TotalStorage SAN Switch family.

The base functionality for both IBM TotalStorage SAN Switch H08 and IBM TotalStorage SAN Switch H16 models includes webtools, advanced zoning, and hot code activation standard in Fabric OS 4.2. Just like all the IBM TotalStorage SAN Switch family, all ports are numbered sequentially starting with zero for the left-most port. The switch faceplate includes a silk screen imprint of the port numbers. With the 2 Gb/s switches, the ports are color-coded into quad-groups to indicate which ports can be used in the same ISL trunking group. Both model supports Fibre Channel classes 2, 3 and F and has a latency of less than 2 µs with no contention (assuming the destination port is free).

IBM TotalStorage SAN Switch H08

TheIBM TotalStorage SAN Switch H08 as shown in Figure 8-1 on page 290 is designed for entry level SAN applications, and provides 8 Fibre Channel ports, a single power supply, and fans in 1U rack height. The base IBM TotalStorage SAN Switch H08 functionality supports one E-Port connection to one other IBM SAN switch. The Entry Fabric configuration includes four SW-SFP transceivers. A mixture of SW, LW and extended LW-SFPs features may also be added.

Figure 8-1 IBM TotalStorage SAN Switch H08

IBM TotalStorage SAN Switch H16

The IBM TotalStorage SAN Switch H16 is a 16-port SAN switch which is designed for mid-range SAN applications and provides dual power supplies and fans in 1U rack height. The base IBM TotalStorage SAN Switch H16 functionality supports connections with up to three other IBM SAN switches. The base configuration includes eight SW-SFP transceivers. A mixture of eight SW, LW and extended LW-SFP features may be added.

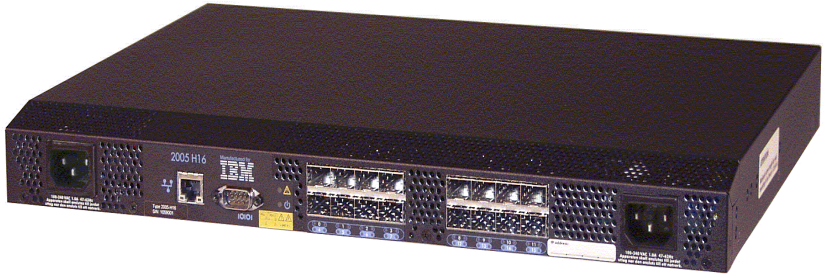


Figure 8-2 IBM TotalStorage SAN Switch H16

Features and functions of model H08 and model H16

The model H08 and H16 provide the following features and functions:

- Provides eight (Model H08) or sixteen (Model H16) Fibre Channel ports with the following characteristics:
  - Automatic negotiation to the highest common speed of all devices connected to port.

  - Compatible port interfaces with small form factor pluggable (SFP) transceivers for both short wavelength (SWL), long wavelength (LWL) and extended long wavelength (ELWL).
  - Universal and self-configuring: capable of becoming an F_Port, FL_Port, or E_Port.
- Supports a domain limit of 2 for Model H08 and a domain limit of 4 for Model H16.
- A single (Model H08) or dual (Model H16) power supply with fixed fans
- A small 1U chassis that can be either rack-mounted or used in a table-top configuration.
- Runs Fabric OS 4.2 and supports the following licensed features: Advance Zoning, ISL, Fabric Watch, Advanced Performance Monitoring, Extended Fabrics, Remote Switch, Web Tools and Secure Fabric OS. It also has a concurrent code activation.
- Full fabric activation feature adds unlimited E_Ports.
- Distance capability feature increases buffer credits.
- One RS-232 serial port, designed to connect to a DTE port.
- One 10/100 MB/s Ethernet port with RJ-45 connector.
- Non-blocking performance architecture and distance support.

For more information, H06 and H16 datasheets can be downloaded at:

http://www.storage.ibm.com/ibmsan/products/2109/library.html#support

8.1.3 2109-F16 and 2109-F32

The F16 and F32 are also built upon the third-generation switch technology that supports link bandwidth of 1 and 2 Gb/s.

The F16 is shown in Figure 8-3 is a 16-port switch.

Figure 8-3 IBM SAN Fibre Channel Switch 2109-F16

The ports of all of the IBM SAN switches are numbered sequentially starting with zero for the left-most port. The switch faceplate includes a silk screen imprint of the port numbers. With the 2 Gb/s switches, the ports are color-coded into quad-groups to indicate which ports can be used in the same ISL trunking group.

The F16 consists of a system board with connectors for supporting up to 16 ports and a Fabric Operating System for building and managing a SAN fabric. The F16 supports Fibre Channel classes 2, 3 and F and has a latency of less than 2 µs with no contention (assuming the destination port is free).

The base model F16 configuration comes with eight SW or LW-SFPs. In addition, a mixture of SW and LW ports can be configured by adding up to sixteen SFP transceivers. High availability is supported by hot-pluggable cooling fans and SFPs, and an additional power supply can be ordered for redundancy. Advanced Zoning and Web Tools is included in the F16 basic configuration.

IBM TotalStorage SAN Switch F16 provides the following features:

- Sixteen Fibre Channel ports, each capable of full-duplex throughput at either 1 or 2 gigabits per second.
- All ports are auto-sensing that self-negotiate to the highest speed supported by the attached server, storage or switch.
- They are all Universal ports that self-configure as F_Ports, FL_Ports, or E_Ports.
- Each Fibre Channel port uses Small Form-Factor Pluggable (SFP) media with options for either shortwave optical connection for distances of up to 300 meters, or longwave optical connection for distances of up to 10 kilometers.
- A 1U package that can be either rack-mounted or used as table-top configuration, with the option of a redundant power supply, providing a highly available switch.
- Hardware zoning controlled at the port level, and at the worldwide name level.
- Cascading support for flexibility in creating scalable fabric topologies.
- Distributed fabric services such as name serving, routing, Advanced Zoning, Fabric Watch and microcode upgrade.
- Web Tools which provides a comprehensive set of management tools that support a web browser interface for easy-to-use operations.

Optional features activated by license includes:

- Extended Fabric, Remote Switch, QuickLoop, Security.
- Performance Bundle, which provides Inter-Switch Link (ISL) and Advance Performance Monitoring.

The F32, as shown in Figure 8-4, is a 32-port switch that shares the same characteristics as the F16.

Figure 8-4 IBM SAN Fibre Channel Switch 2109-F32

The base model F32 configuration comes with 16 SW-SFPs. In addition, a mixture of SW and LW ports can be configured by adding up to sixteen SFP transceivers. High availability is supported by hot-pluggable cooling fans and SFPs. A redundant power supply enabling dual-power and non-disruptive power supply maintenance is included.The Performance Bundle function, Advanced Zoning and Web Tools is also included in the F32 basic configuration.

The IBM TotalStorage SAN Switch 16 provides the following features:

- Thirty-two ports, each ports capable of full-duplex throughput at either 1 or 2 gigabits per second.
- Auto-sensing ports that auto-negotiate to the highest speed available supported by the attached server, storage, or switch.
- Universal porst self-configure as F_Ports, FL_Ports, or E_Ports.
- Each port supports the new Small Form-Factor Pluggable (SFP) media with options for either shortwave optical connection for distances up to 300 meters, or longwave optical connections for distances up to 10 kilometers.
- A 1U package that either can be rack-mounted or used in a table-top configuration with a redundant power supply, providing a highly available switch.
- Hardware zoning controlled at the port level, and at the world wide name.
- Performance Bundle feature, which provides Inter-Switch Link (ISL) Trunking and Advance Performance Monitoring.
- Cascading support for flexibility is creating scalable fabric topologies.
- Distributed fabric services such as name serving, Advanced Zoning, routing.
- Web Tools which provides a comprehensive set of management tools that support a Web browser interface for easy-to-use operation.
- Concurrent code activation, allowing switch firmware upgrades without the need to remove the switch from the fabric.

Optional features activated by License key includes Extended Fabric, Remote Switch, and Security.

For more information, the F16 and F32 data sheets can be downloaded from:

http://www.storage.ibm.com/ibmsan/products/2109/library.html#support

8.1.4 2109-M12

The M12 is a bladed architecture, it consists of two logical 64-port switches in one chassis. Each logical switch has its own:

- Unique domain ID
- Switch World Wide Name (WWN)
- IP address

Both logical switches share:

- Chassis
- Four hot swappable, redundant power supplies:
  - Any two are needed to provide power for a maximum configuration
  - Selective power down of cards (configurable) when only one power supply is working
  - Two AC connectors
- Three hot swappable, redundant fans
  - Any two needed to cool the entire switch
  - One fan can keep the unit running for about one hour
- Dual control processor (CP) operates in active/standby mode

The M12 is shown in Figure 8-5.

.

Figure 8-5 2109-M12

The switch layout is shown in Figure 8-6.

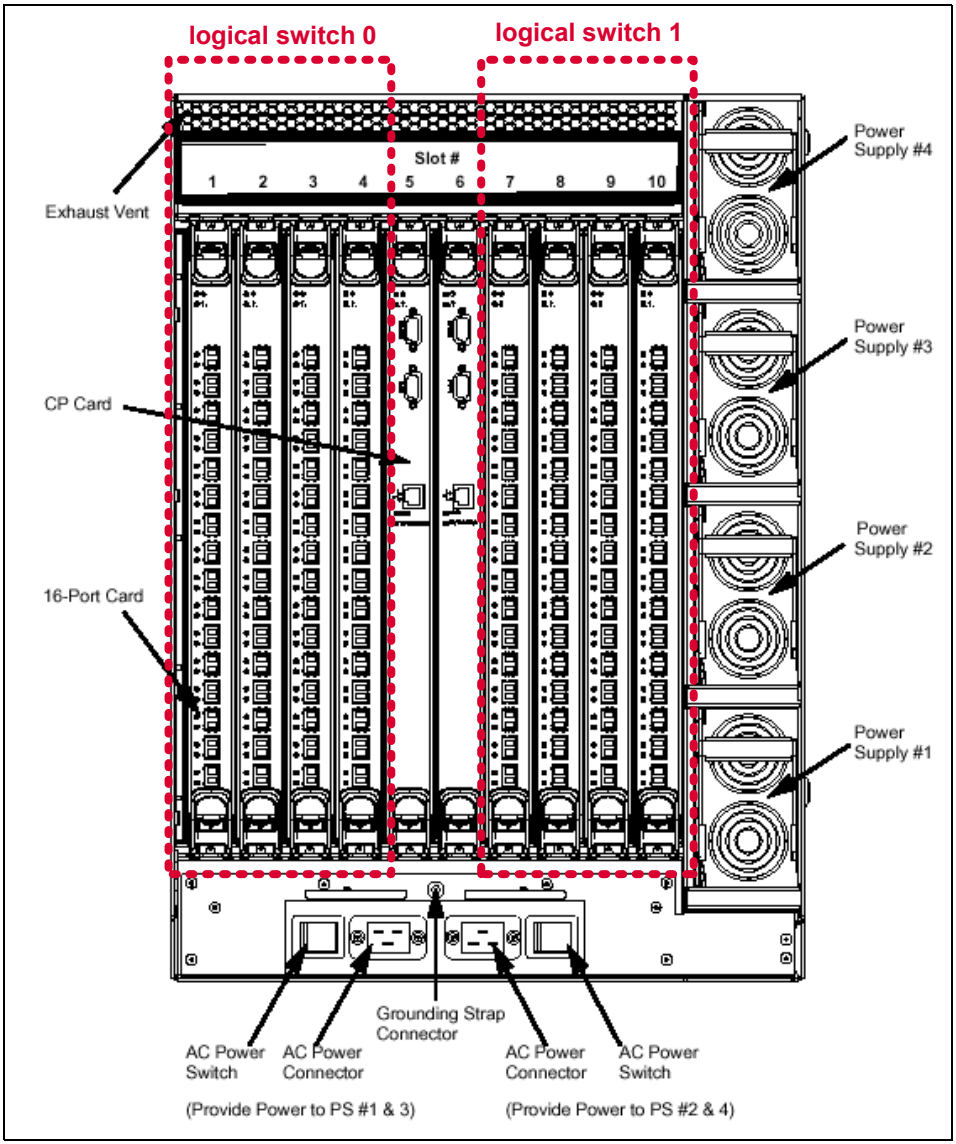


Figure 8-6 Port side of the 2109-M12

Each of the logical switches supports up to four, 16-port Fibre Channel modules (blades) enabling 64 universal (E, F, and FL ports), full duplex, ports. Each port is capable of self-negotiation to the highest speed supported by the attached SAN device. Two CP cards are located in the middle of the chassis, “logical switch 0” occupies blades from slot 1 to 4, ‘logical switch 1’ from 7 to 10.

The chassis holds two redundant hot-swappable CP cards in slots 5 and 6, providing a modem serial port, a terminal serial port, and a 10/100 Mb/s Ethernet port.

Switches and processors are de-coupled, meaning one CP card is active, the other one is in standby mode. The active CP manages both logical switches and only the active CP card controls the M12. If the active CP card fails or is removed, the standby CP card automatically becomes the active CP card. Failover occurs as soon as the active CP card is detected as faulty or is removed.

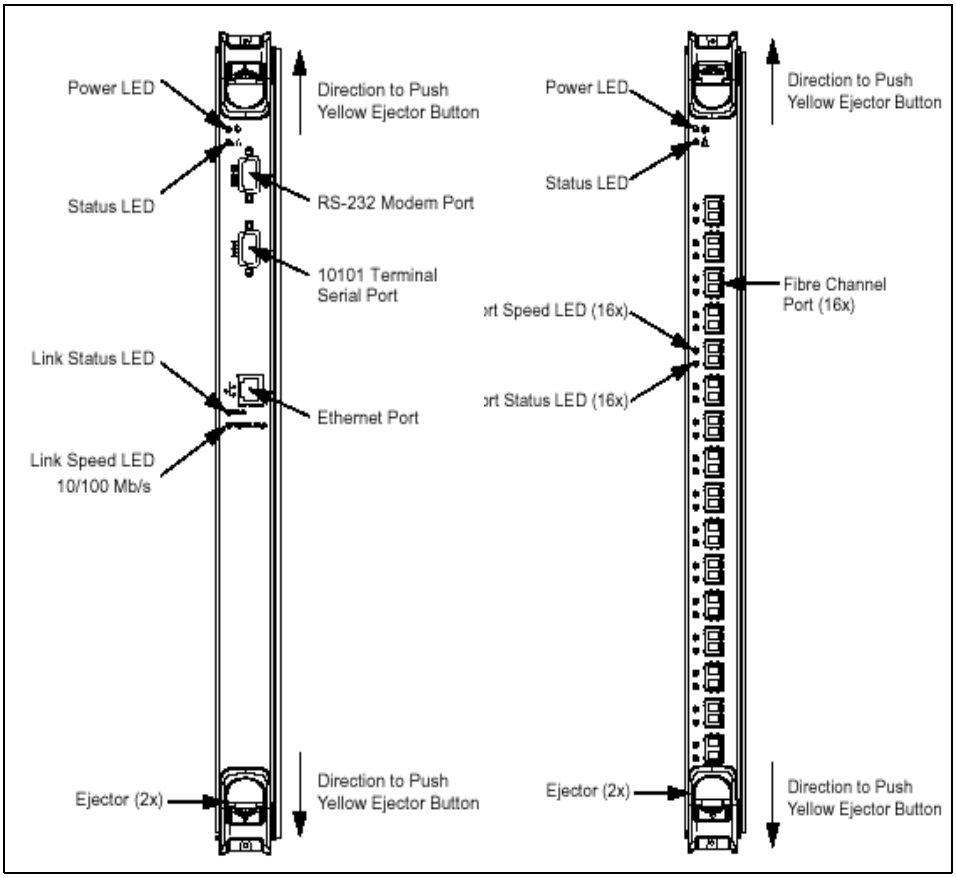


Figure 8-7 CP (left) and Fibre Channel Module (right)

The port blades consist of 16 ports providing SFP transceivers. The same advanced ASIC technology introduced in the Fxx models is used in the M12.

M12 port numbering scheme

Physical port numbering for each port card begins with port 0 at the bottom, and port 15 at the top of the card, as shown in Figure 8-8.

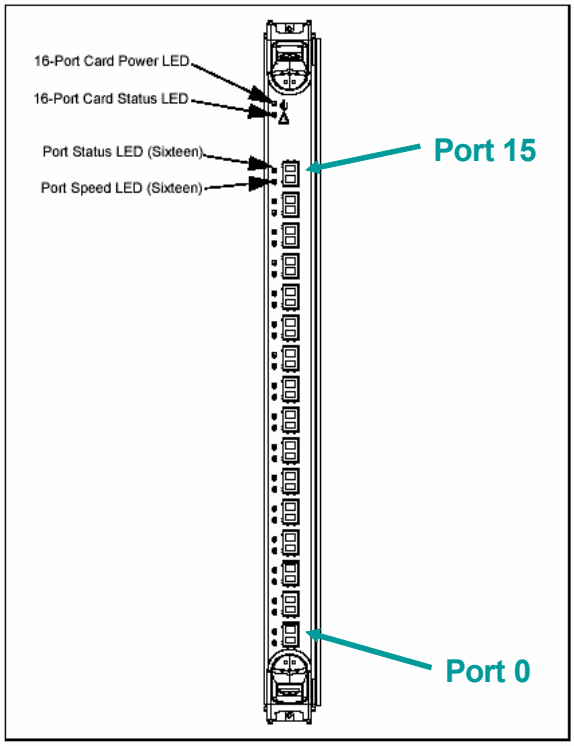


Figure 8-8 Physical port numbering

Because each logical switch can have up to 64 ports, it is necessary to number these ports from a switch perspective, not just at a blade level. This switch port numbering is known as port area (sometimes referred to as the absolute port number) or area, numbering the ports for each logical switch from area 0 through 63.

Using the command line interface (CLI), zoning commands use the port area numbering, while other commands require the slot/port method.

The table in Figure 8-9 shows the *area* number for each physical port location:

- The Physical Slot number refers to the logical switch 0 / logical switch 1 slot position.
- The Logical Slot numbering is only used to help define the FC address.

Logical Slot 0		Logical Slot 1		Logical Slot 2		Logical Slot 3	
Physical Slot 1/7		Physical Slot 2/8		Physical Slot 3/9		Physical Slot 4/10	
Area #	Physical Port	Area #	Physical Port	Area #	Physical Port	Area #	Physical Port
15	15	31	15	47	15	63	15
14	14	30	14	46	14	62	14
13	13	29	13	45	13	61	13
12	12	28	12	44	12	60	12
11	11	27	11	43	11	59	11
10	10	26	10	42	10	58	10
9	9	25	9	41	9	57	9
8	8	24	8	40	8	56	8
7	7	23	7	39	7	55	7
6	6	22	6	38	6	54	6
5	5	21	5	37	5	53	5
4	4	20	4	36	4	52	4
3	3	19	3	35	3	51	3
2	2	18	2	34	2	50	2
1	1	17	1	33	1	49	1
0	0	16	0	32	0	48	0

Figure 8-9 Physical port location to area numbering cross reference

The IBM TotalStorage SAN Switch M12 accommodates two logical 64-port switches in one chassis. In order to form a 128-port fabric, both logical switches have to be connected by one or more external ISLs. For critical applications and high availability, it is not recommended to use the dual switches in a single chassis as redundant fabrics or redundant core, since there is still the potential for a single point of failure.

The base configuration provides an IBM rack fully equipped with power supplies and fans. It includes two CP cards and two FC cards holding 32 SFP either SW or LW, to form a 32-port switch. The basic software license comes with:

- WebTools
- Advanced Zoning
- Fabric Watch
- Performance Bundle:
  - ISL Trunking
  - Global Performance Monitoring

For IBM TotalStorage SAN Switch M12, the chassis-WWN is used to license all the software products. The QuickLoop feature is not supported directly by the M12 switches, but private loop devices may be attached to QuickLoop-capable

switches in this fabric.For more information, the IBM TotalStorage SAN Switch M12 data sheets can be downloaded from:

http://www.storage.ibm.com/ibmsan/products/2109/library.html#support

8.1.5 IBM TotalStorage SAN Switch M14

The director class IBM TotalStorage SAN Switch M14 is a 128-port non-blocking core fabric SAN switch that delivers high performance, scalability, flexibilty, functionality and availability. It can have a 32 to 128 ports in a single domain. The M14 bladed switch architecture expands connectivity in sixteen port increments providing full-duplex link speeds of 1 and 2 Gb/s capable of automatically negotiating to the highest speed supported by the attached server, storage or switch. A mixture of shortwave and longwave ports can be configured by adding optical SFP transceivers.

This director class IBM TotalStorage SAN Switch M14 is designed to provide high-availability with redundant hot pluggable components. It protects Customer's investments, since they are compatible to all existing IBM TotalStorage SAN switches.A fully-configuredIBM TotalStorage SAN Switch M14 shown in Figure 8-10 also consumes approximately 60% less power compared to a fully-configured IBM TotalStorage SAN Switch M12.

Figure 8-10 IBM TotalStorage SAN Switch M14

The IBM TotalStorage SAN Switch M14 enhancesIBM TotalStorage SAN Switch M12 functionalities and provides the following features:

- 128 ports in a single chassis that provides a high-density, rack-ready solution for SAN backbones.
- Enables non-disruptive scalability from 32 to 128 ports since it can support 128 concurrently active 2 Gb/s full-duplex ports in a single domain.
- High-availability platform for mission-critical SAN designed applications.
- Dual, redundant control processors (CP) which enables nondisruptive software upgrades.

- Nonblocking architecture that provides all 128 ports to operate at full 2 Gb/s speed simulataneously.
- Universal ports that self-configure as E_Port, F_Port, or FL_Port.
- Each ports supports Small Form-Factor Pluggable (SFP) optical transceivers with options for either short wavelength (SWL), long wavelength (LWL) and extended longwavelength (ELWL) optical media on a single switch module (the 16-port card).
- Multiprotocol design supports additional blades, such as application platform blades and blades that provide iSCSI and FCIP capabilities.
- Support for 1 and 2 Gb/s autosensing Fibre Channel ports (trunking technology groups up to four ports togethet to create high-performance 8 Gb/s ISL trunks between switches.

Standard License features includes:

- Fabric Watch
- Advance Web Tools
- Advanced Zoning
  - ISL Trunking
- Advanced Performance Monitoring

Optional features activated by License key includes:

- Extended Fabrics
- Remote Switch
- Advanced Security
- Fabric Manager

Hardware components

The IBM TotalStorage SAN Switch M14 has a modular and scalable mechanical construction that allows a wide range of flexibility in switch installation, fabric design, and maintenance. To give you an overview, we wil describe the components that consists M14:

1. **Backplane**
   - Backplane is the heart of M14 which has been designed to let future enhancements and continuous 100% utilization and nonblocking throughput on all ports simultaneously. The design also allows for the hotplugging of the blade assemblies.
2. **Port cards**

- Each port card provides sixteen 1 or 2 Gb/s auto-sensing Fibre Channel ports which is based on the same ASIC technology used in IBM TotalStorage Switches. Up to eight hot-swappable 16-port cards, delivering up to 128 Fibre Channel ports are supported in a single chassis.
- M14 ports uses a numbering scheme on which they are numbered from 0-15, starting from the bottom of the card.

3. CP card

- M14 consists of two central processor (CP) cards installed in slot 5 and slot 6. A single active CP card can control all 128 ports in the chassis while the stanby CP card assumes control of the switch if the active CP fails.
- Each CP card has a PowerPC 440GP 466 MHz microprocessor on the assembly. Each CP card comes with two serial ports(modem port “RS-232” that is used for remote dial-in and a serial port “10101” for local management) and a 10/100 BASE-T ethernet port to connect them to separate networks for greater availability.

4. Power supplies

Looking at Figure 8-11, we can see four power supplies on the right hand side which are split across the two AC inputs. Power supplies 1 and 3 are fed from the left-hand AC input and power supplies 2 and 4 are fed from the right hand AC input. This power distribution configuration on a fully-configured chassis is capable of contineuous operation durimg periods of losing any two power supplies.

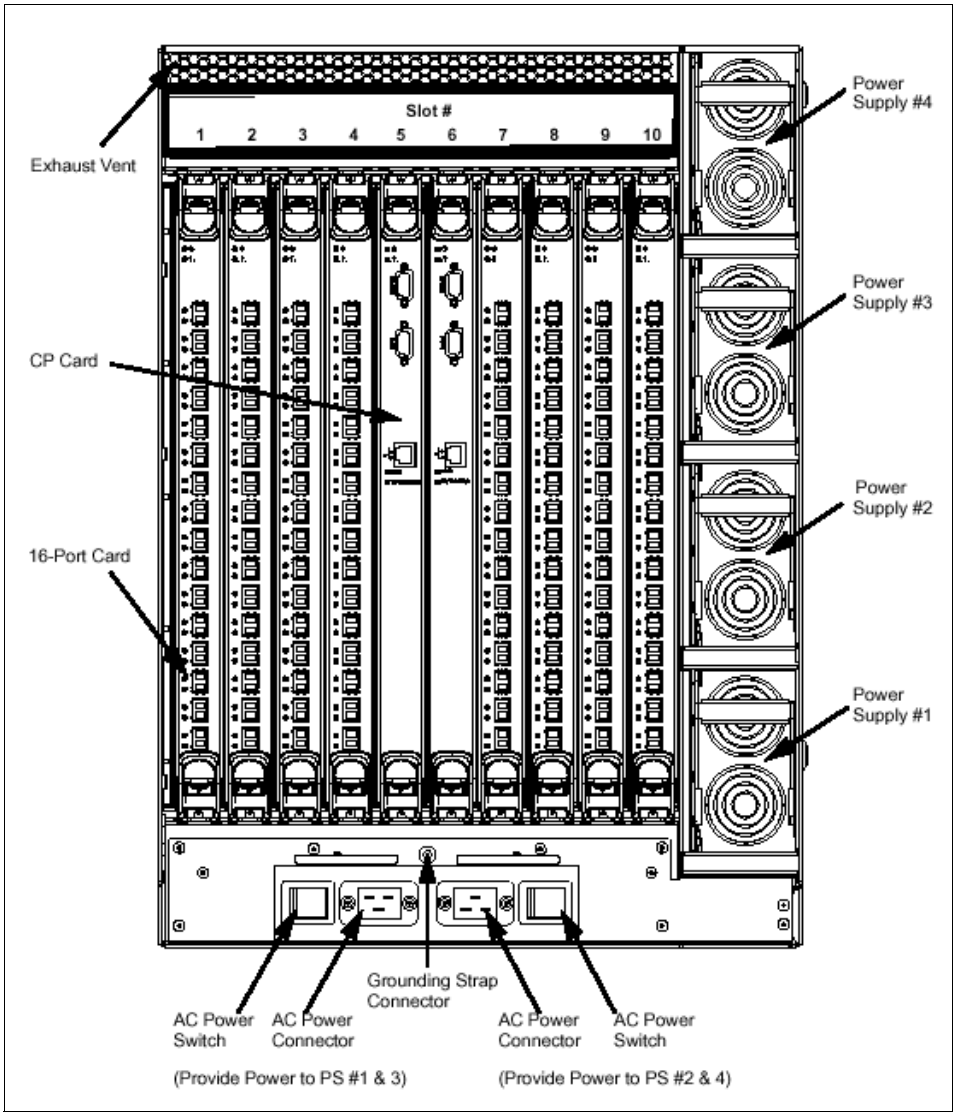


Figure 8-11 Port side of M14

5. Blowers

- The M14 use a redundant system of three hot-swapable blowers. If a blower fails, the remaining two blowers increase their speed to continue adequate cooling.
- Should a second blower fails, the M14 is designed to sequence down the switch blades, so as not to overheat and damaged any

components. This sequence is predefined, but may also be modified to suit your configuration, and can maintain a degraded system configuration during such circumstances.

Numbering scheme

The IBM TotalStorage SAN Switch M14 use a numbering scheme that progresses from left to right and bottom to top in numerical order. The reference location is from the cable side to chassis.

- Blade assemblies are numbered from 1-10 (see Figure 8-11 on page 304, from left to right.
- Ports are numbered from 0-15, bottom to top as seen in Figure 8-12 on page 305

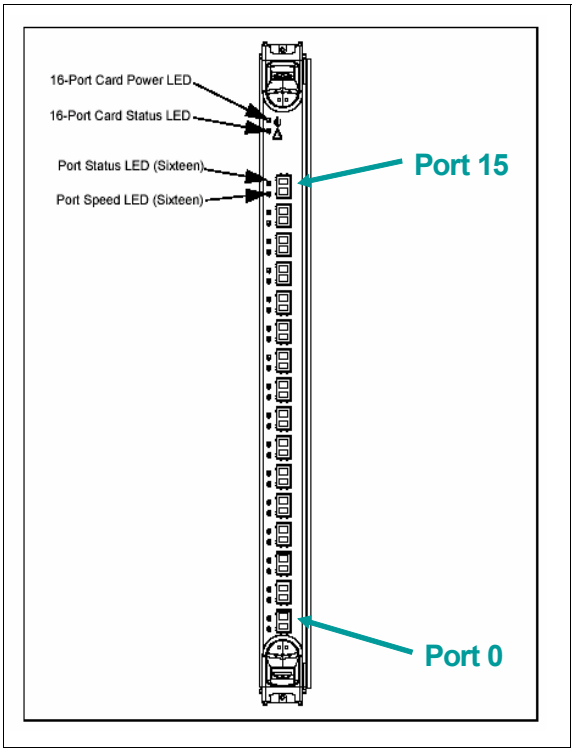


Figure 8-12 Port card

- Power supplies are numbered from 1-4, bottom to top as referred to Figure 8-11 on page 304.
- Blowers are numbered from 1-3, left to right showed in Figure 8-11 on page 304.

8.2 Features

In the following sections we will discuss some features available in IBM TotalStorage SAN Switch product model Mxx, Hxx, and Fxx.

Advanced Web Tools

Advanced Web Tools is an intuitive graphical user interface (GUI) which allows network managers to monitor and manage SAN fabrics consisting of switches using a Java-capable Web browser from standard desktop workstations. By entering the network address of any switch in the fabric, the built-in Web server automatically provides a full view of the switch fabric. From that switch, the administrator can monitor the status and perform administration and configuration actions on any switch in the SAN.

Advanced Performance Monitoring

Advanced Performance Monitoring is essential for optimizing fabric performance, maximizing resource utilization, and measuring end-to-end service levels in large SANs. It helps to reduce total cost of ownership (TCO) and over-provisioning while enabling SAN performance tuning and reporting of service level agreements. Advanced Performance Monitoring enables SAN administrators to monitor transmit and receive traffic from the source device all the way to the destination device. Single applications such as Web serving, databases, or e-mail can be analyzed as complete systems with near-real-time performance information about the data traffic going between the server and the storage devices. This end-to-end visibility into the fabric enables SAN administrators to identify bottlenecks and optimize fabric configuration resources.

Advanced Security

Advanced Security (AS)significantly reduces the security holes left by traditional SAN implementations and greatly improves the ability to minimize SAN-specific vulnerabilities by providing a comprehensive,policy based security system for IBM SAN Switch fabrics. With its flexible design it enables organizations to customize SAN security in order to meet specific policy requirements.

Advanced Zoning

Advanced Zoning segments a fabric into vitual private SANs. It provides data exchange between devices in the same zone and prohibits exchange to any device not in the same zone. Advance zoning of the 2 Gb/s switches compared with the zoning of 1 Gb/s switches enlarges the range of hardware enforcement and so provides the switch with more security access control functions as before, preventing an unauthorized devices from accessing the fabric.

Extended Fabric

The Extended Fabric feature provides extensions within the internal switch buffers. This maintains performance with distances greater than 10 km, and up to 120 km, by maximizing buffering between the selected switch interconnect links. With the Extended Fabric feature, the ISLs are configured with up to 60 buffer credits and optimizes buffers for up to 120 km on 1 Gb/s fiber optic link, and up to 60 km on 2 Gb/s fiber optic link.

Fabric Manager

Fabric Manager provides a Java-based application that can simplify management of a multiple switch fabric. It administers, configures, and maintains fabric switches and SANs with host-based software. Web Tools and Fabric Manager run on the same management server attached to any switch in the fabric. Fabric Manager requires a Windows NT/2K or Solaris 7 server with a Netscape or Internet Explorer Web browser.

Fabric Watch

Fabric Watch enables switches to continuously monitor the health of the fabrics, watching for potential faults based on defined thresholds for fabric elements and events, so making it easy to quickly identify and escalate potential problems. It monitors each element for out-of-boundary values or counters and provides notification to SAN administrators when any exceed the defined boundaries. SAN administrators can configure which elements, such as error, status, and performance counters within a switch, are monitored.

ISL Trunking

ISL Trunking feature allows up to four ISLs between the same pair of switches to be grouped and to act as a single, high speed “pipe” or trunk with a capacity of up to 8 Gb/s. In-order delivery of frames is guaranteed and allows ISLs to be added to or removed from the trunking group seamlessly.

Remote Switch

The Remote Switch feature enables two fabric switches to be connected over an asynchronous transfer mode (ATM) connection. This requires a compatible Fibre Channel to ATM gateway, and can have a distance of up to 10 km between each switch and the respective ATM gateway.

8.3 Advanced Security

As organizations grow their SANs and connect them over longer distances through existing networks, they have an even greater need to effectively manage

SAN security and policy requirements. To help these organizations improve security, Advanced Security (AS) provides a comprehensive security solution for IBM-based SAN fabrics. With its flexible design it enables organizations to customize SAN security in order to meet specific policy requirements. In addition, it works with a security practice which is already deployed in many SAN environments: Advanced Zoning.

The most complete solution for securing SAN infrastructures, AS provides following features to Fabric OS:

- Fabric Configuration Servers (FCS, trusted switches)
- Management Access Controls (MAC)
- Device Connection Controls (DCC)
- Switch Connection Controls (SCC)
- Secure Management Communications

These features will be used in a structured way by defining through the Fabric Management Policy Set (FMPS). It specifies access controls to apply to the fabric management capabilities and the physical connections and components within the fabric. FMPS handles several different types of policies, each with different aspects. The policies provide control over management access to the fabric. Together with the potential points of vulnerability of fabric devices , organizations use FMPS to define their security requirements for a fabric by establishing a set of security domains. These domains typically define different categories of communications that must be protected by the fabric security architecture. These domains include:

Host-to-Switch Domain

In host-to-switch communications, individual device ports are bound to a set of one or more switch ports using Access Control Lists (ACLs). Device ports are specified by WWN and typically represent HBAs. The AS OS DCC feature enables binding by WWN (port) and ACL to secure the host-to-switch connection for both normal operations and management functions.

Administrator-to-Security Management Domain

Because security management impacts the security policy and configuration of the entire SAN fabric, administrator access controls work in conjunction with security management functions. In addition, administrator-level fabric password access provides primary control over security configurations.

Security Management-to-Fabric Domain

AS secures certain elements of the management communications — such as passwords — on some interfaces between the security management function and a switch fabric. The security management function encrypts appropriate data elements (along with a random number) with the switch's public key. The switch

then decrypts the data element with its private key. For more information about public and private keys, see “Encryption” on page 172.

Switch-to-Switch Domain

In secure switch-to-switch communications, the switches enforce security policy. The security management function initializes switches by using digital certificates and ACLs. Prior to establishing any communications, switches exchange these credentials during mutual authentication. This practice ensures that only authenticated and authorized switches can join as members of the SAN fabric or a specific fabric zone. This authentication process prevents an unauthorized switch from attaching to the fabric through an E_Port.

8.3.1 Fabric configuration servers

Fabric Configuration Servers are trusted SAN switches responsible for managing the configuration and security parameters (including zoning) of all other switches in the fabric. Any number of switches within a fabric can be designated as Fabric Configuration Servers as specified by WWN, and the list of designated switches is known fabric-wide.

As part of the security policy configuration process, organizations select a primary Fabric Configuration Server and potential backup servers. Among these, only the primary Fabric Configuration Server can initiate fabric wide management changes, and all initiation requests must be identified to ensure fabric security: a capability that helps eliminate unidentified local management requests initiated from subordinate switches.

8.3.2 Management access controls

Management Access Controls enable organizations to restrict management service access to a specific set of end points: either IP addresses (for SNMP, Telnet, HTTP, or API access), device ports (for in-band methods such as SES or Management Server), or switch WWNs (for serial port and front-panel access). Disabling front-panel access of the older S16 switch prevents unauthorized users from manually changing fabric settings.

IBM TotalStorage SAN switches enable secure IP-based management communications (like SSL) between a switch and Web Tools. Elements of the manager-to-switch-communications process, such as passwords, are encrypted to increase security.

The IBM TotalStorage SAN Switch M14 also provides secure Telnet access through SSH Secure Shell, a network security protocol that helps ensure secure remote login and other network services over insecure networks.

8.3.3 Device connection controls

Device connection controls, also known as WWN Access Control Lists (ACLs) or Port ACLs, enable organizations to bind an individual device port to a set of one or more switch ports. Device ports are specified by WWN and typically represent HBAs. These controls secure the server-to-fabric connection for both normal operations and management functions.

By binding a specific WWN to a specific switch port or set of ports, device connection controls can prevent a port in another physical location from assuming the identity of a real WWN. This capability enables better control over shared switch environments by allowing only a set of predefined WWNs to access particular ports in the fabric.

8.3.4 Switch connection controls

Switch connection controls enable organizations to restrict fabric connections to a designated set of switches, as identified by WWN. When a new switch is connected to a switch that is already part of the fabric, the two switches must be mutually authenticated. As a result, each switch must have a digital certificate and a unique public/private key pair to enable truly authenticated switch-to-switch connectivity.

New switches receive digital certificates at the time of manufacture. However, organizations with existing switches will need to upgrade them with certificate and key information at the installed location.

Switch-to-switch operations are managed in-band, so no IP communication is required. This capability prevents users from arbitrarily adding switches to a fabric. Any new switch must have a valid certificate and also appear in the fabric-authorized switch ACL. Digital certificates ensure that the switch name (which is the WWN) is authentic and has not been modified.

8.3.5 Fibre Channel Authentication Protocol

The Switch Link Authentication Protocol (SLAP/FC-SW-3), establishes a region of trust between switches. For an end-to-end solution to be effective, this region of trust must extend throughout the SAN, which requires the participation of fabric-connected devices, such as HBAs. The joint initiative between Brocade and Emulex establishes Fibre Channel Authentication Protocol (FCAP) as the next-generation implementation of SLAP. Customers gain the assurance that a region of trust extends over the entire domain.

FCAP has been incorporated into its fabric switch architecture and has proposed the specification as a standard to ANSI T11 (as part of FC-SP). FCAP is a Public

Key Infrastructure (PKI)-based cryptographic authentication mechanism for establishing a common region of trust among the various entities (such as switches and HBAs) in a SAN. A central, trusted third party serves as a guarantor to establish this trust. With FCAP, certificate exchange takes place among the switches and edge devices in the fabric to create a region of trust consisting of switches and HBAs.

Because a network is only as secure as its weakest link, all switches in the fabric must support AS in order to achieve the highest level of security fabric-wide.

Advanced Security is covered in more depth in the IBM Redpaper:

- *Advanced Security in an IBM SAN*, REDP3726

Details of how to implement Advanced Security can be found in this edition of the IBM Redbook:

- *Implementing an Open IBM SAN,* SG24-6116-03

8.4 ISL

Although redundancy provides an excellent way to enhance availability, it does not protect against all types of outages. Just as companies have embraced client/server networking to overcome the limitations of the mainframe-centric IT infrastructure, many are taking a similar approach to SANs. A networked SAN is a flexible architecture that can be easily implemented and quickly adapted to changing requirements — extending the availability characteristics of hardware and software components into the SAN fabric itself.

To build a SAN network, switches need to be connected by Inter-Switch Links (ISL). An ISL is created simply by connecting two switches with a fiber optic cable. Both switch ports turn immediately into E_Ports, and the switch automatically discovers connected switches and creates the FSPF routing table used by the entire fabric. No programming of the fabric is necessary, as the FSPF table will be updated as new switches join in.

The network can be scaled from the number of ports needed at the edge, as well as being scaled at the core switch level, to provide higher bandwidth and redundant connectivity. In fact, SAN fabrics can feature multiple levels of availability, including meshed tree topologies of switches, single fabrics with dual connectivity, and dual fabrics with dual connectivity for environments that require the highest levels of availability.

8.4.1 ISLs without trunking

ISLs provide for connection between switches. Any switch in the fabric can have one or more links to another switch in the fabric. At switch start-up, these links are initialized and at fabric login of the Fibre Channel devices, these ISLs are allocated in a round-robin fashion to share the load on the system. The switch guarantees in-order delivery, however, it means that if one Fibre Channel device loads up its dedicated ISL highly and for lengthy periods of time, a second device dedicated to this very ISL may not get all of its data through, as shown in Figure 8-13.

At the same time, a parallel ISL that is dedicated to another Fibre Channel device may be idle.

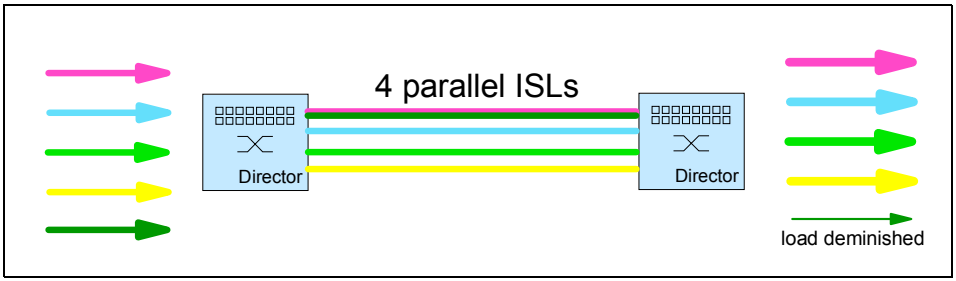


Figure 8-13 Parallel ISLs without trunking

However, there are some features that can be used to increase inter-switch bandwidth:

- Adding an ISL between switches is dynamic and can be done while the switch is active. Adding a new ISL will result in a routing re-computation and new allocation of ISL links between source and destination ports. Similarly, removing a link will result in FSPF routing re-computation across the fabric and possible fabric re-configuration.
- Adding ISLs will cause routing traffic/zoning data to be updated across the fabic via a spanning tree. The total number of ISLs is not so relevant as the amount of configuration changing, as each change will result in a re-calculation of routes in the fabric. When numerous fabric reconfigurations occur (removing or adding links, rebooting a switch, and so on) the load on the switches CPUs will be increased and some fabric events may time-out waiting on CPU response. This occurs only during fabric reconfiguration activities and does affect frame traffic per se, only tasks that require use of the CPU (no CPU intervention is required for normal frame routing, this is all done by switch hardware).

- No more than eight ISLs between any two switches is supported. More than eight ports can be used on a switch for ISL traffic as long as no more than eight go to a single adjacent switch.

Note: A spanning tree connects all switches from the so called principal switch to all subordinate switches. This tree spans in a way such that each switch (or leaf of the tree) is connected to other switches, even if there is more than one ISL between them - that is to say, there are no loops.

8.4.2 ISLs with trunking

It is possible that there are drawbacks in using parallel ISLs as it was implemented with the 1 Gb/s switches. With the 2 Gb/s switches, there is an optional feature called ISL Trunking. Trunking is ideal for optimizing performance and simplifying the management of a multi-switch SAN fabric.

When two, three, or four adjacent ISLs are used to connect two switches, the switches automatically group the ISLs into a single logical ISL, or trunk. The throughput of the resulting trunk is 4, 6, or 8 Gb/s.

ISL trunking is designed to significantly reduce traffic congestion. As shown in Figure 8-14, four ISLs are combined into a single logical ISL with a total bandwidth of 8 Gb/s. It can support any number of devices and Figure 8-14 simplifies the case by showing five exchanges at a time over four ISLs.

To balance the load across all of the ISLs in the trunk, each incoming frame is sent across the first available physical ISL in the trunk. As a result, transient workload peaks for one system or application are much less likely to impact the performance of other devices of the SAN fabric.

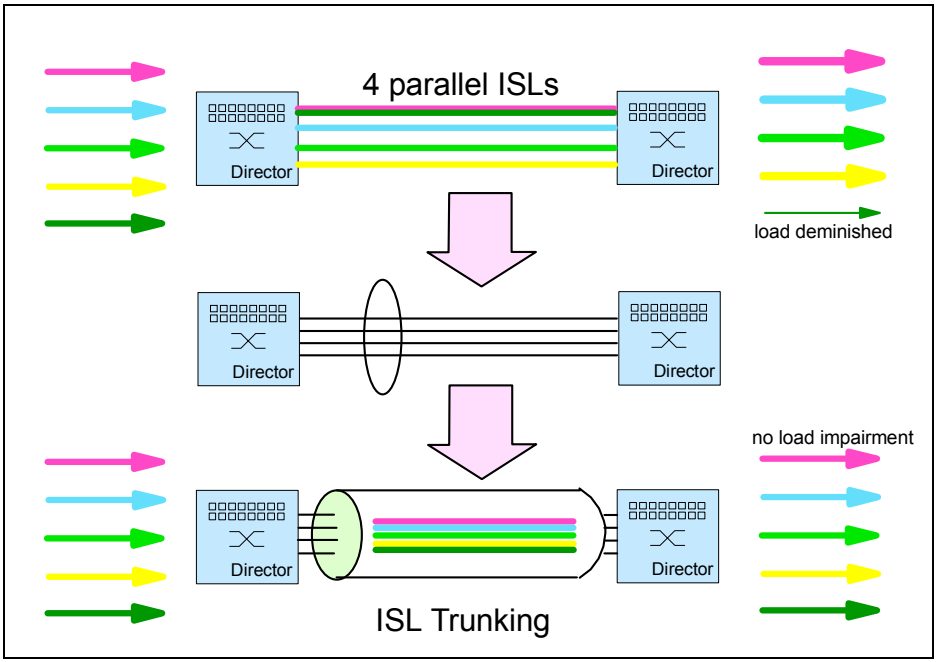


Figure 8-14 2109 ISL trunking

Because the full bandwidth of each physical link is available with ISL trunking, no bandwidth is wasted by inefficient load sharing. As a result, the entire fabric is utilized more efficiently. Moreover, Fabric OS and Management software like Fabric Watch views the group of physical ISLs as a single logical ISL. A failure of a single ISL in a trunk is only a reduction of the available bandwidth and not a failure of the complete route. Therefore, no re-calculation of the routes at that time is needed, and bandwidth automatically increases when the ISL is repaired.

ISL trunking will help simplify fabric design, lower provisioning time, it enhances switch-to-switch performance, simplifies management, and improves the reliability of SANs and in-order delivery is still guaranteed.

Load sharing and load balancing: Non-trunked, parallel ISLs always shared load or traffic in a rough, server-oriented way: next server gets the next available ISL, regardless of the amount of traffic each server is generating. Load balancing, however, is the means to find an effective way to use all of the cumulative bandwidth of the parallel ISLs.

8.4.3 Switch count

The ultimate limitation in fabric design is a maximum of 239 physical switches, whatever switches are used. This limit is imposed by the actual number of domain IDs that can be uniquely established in the Fibre Channel Device ID header on the frames. The practical limit and what has been tested is much fewer switches. Tests are conducted on SAN fabrics of up to 32 switches, with no more than seven hops supported from the source port to the destination port.

The hop count limit is set by the Fabric OS and is used to derive a frame hold time value per switch. The hold time value is the maximum amount of time a frame can be held in a switch before it is dropped (class 3) or F_BSY (class 2) is returned. A frame would be held if its destination port is not available. The hold time is derived from the error detect time-out value and the resource allocation time-out value using a formula as follows:

- **E_D_TOV:** Error detect time-out value. When this time is exceeded and the sending port has not been notified of receipt of data by the receiving port for a transmission, this error condition occurs (2 s default in IBM SAN Switches)
- **R_A_TOV:** Resource allocation time-out value. A fabric resource with a reported error condition that is not cleared will be locked out from reuse for this time. Minimum R_A_TOV computes to two times E_D_TOV (10 s default in IBM SAN Switches)
- **Holdtime:** The Holdtime = (R_A_TOV - E_D_TOV) / (Hop Count +1) / 2 ms (where time value is in milliseconds. For 7 hops, and the default E_D_TOV of 2000 milliseconds, the hold time per switch is 500 ms.

Note: The value of 7 for maximum hops is a Fabric OS parameter used to derive hold time. The actual hops in the fabric are not monitored and restricted to 7. More hops are possible, increasing R_A_TOV from the default will allow for longer switch hold times prior to an error condition; however, the default value for the hops has been chosen as a reasonable limitation in fabrics composed of up to 32 switches. This value has been chosen so there should be more than adequate time to allow for frame traffic to traverse the fabric, unless there is a problem preventing a port from responding.

8.4.4 Distributed fabrics

The data transmission range is up to 500 m for shortwave fiber link and up to 10 kilometers for longwave fiber link. There are also extra long distance GBICs on the market which can drive the optical signal distances of up to about 70 km.

To distribute fabrics over extended distances, IBM offers two new optional features, which we describe in the following sections.

Extended Fabric

This enables fabric interconnectivity over Fibre Channel at distances up to 120 km. In this implementation, ISLs use either DWDM devices, extended LW-GBICs or dark fiber repeater-connections to transfer data. The Extended Fabric feature optimizes switch buffering to ensure the highest possible performance over ISLs. With the Extended Fabric feature the ISLs are configured with up to 60 buffer credits and optimizes buffers for up to 120 km on 1 Gb/s fiber optic link, and up to 60 km on 2 Gb/s fiber optic link.

In a fabric consisting of 2109-Mxx and -Fxx switches, the long distance ISL that connects both locations, must be installed between edge port switches of the same model. An Extended Fabric does not work if the long distance ISL is installed between non-matching edge port switches, for example between an M12 and an F16.

The enhanced switch buffers help ensure that data transfer can occur at near-full bandwidth to efficiently utilize the connection over the extended links. To enable the Extended Fabric feature, the license must be installed on each switch in the fabric, the long distance Extended Fabric configuration has to be set only at the edge port connector switch. This switch automatically manages the rest of the switches in the extended fabric.

A high level view of an extended fabric is shown in Figure 8-15.

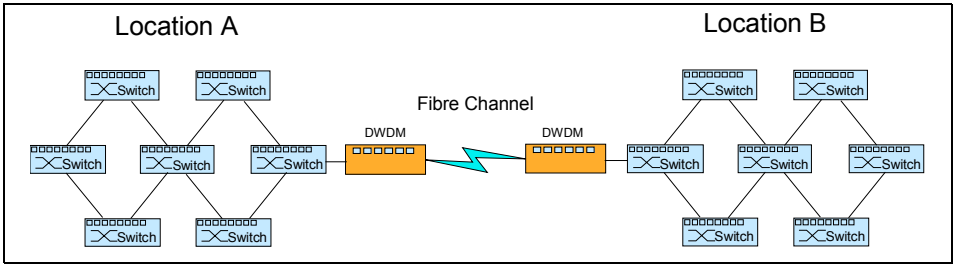


Figure 8-15 Extended Fabrics feature using dark fiber and DWDM

Remote Switch

This feature enables two switches to interconnect over a WAN by *gateways* (or network-bridges). The gateway supports both Fibre Channel Physical Interface as well as a secondary interface such as ATM. It accepts Fibre Channel frames from one side of a Remote Switch fabric, tunnels them across the network, and passes them to the other side of the Remote Switch fabric. This implementation is ideal for environments where dark fiber is not available or when the distance between the two sites exceeds 100 km. To enable the Remote Switch feature, the Remote Switch license must be installed on both switches connecting to the gateway, and the configuration has to be changed on this switch pair.

Both of these optional features are enabled through software capabilities in the switch. A SAN implemented via the Extended Fabric or Remote Switch feature provides all the facilities currently available in locally connected SANs such as these:

- **Single, distributed fabric services such as the name server and zoning:** Each device attached to the SAN appears as a local device, simplifying deployment and administration.
- **Comprehensive management environment:** All management traffic flows through internal SAN connections (IP over Fibre Channel) to allow the fabric to be managed from a single administrator console using Web Tools switch management software.

An example of a remote switch fabric is shown in Figure 8-16.

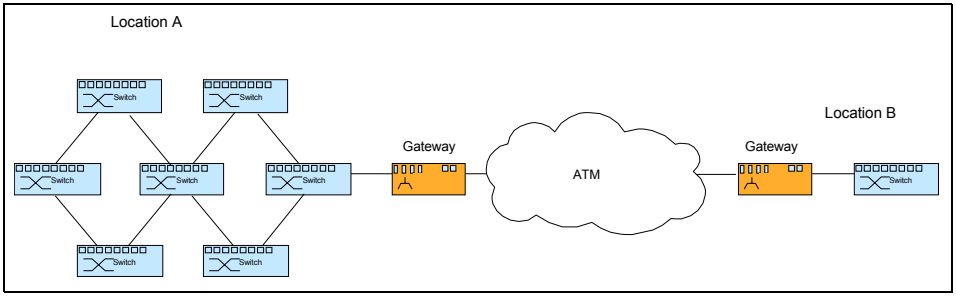


Figure 8-16 Remote Switch feature using ATM

8.5 FICON

FICON switching is supported by IBM TotalStorage SAN Switch M12 and IBM TotalStorage SAN Switch M14 which includes:

FICON servers

Native FICON is automatically supported on Fabric OS 4.2 for IBM TotalStorage SAN Switch M14 and Fabric OS 4.1 and higher for IBM TotalStorage SAN Switch M12. This native FICON is available in one Director per fabric.

Intermixed FICON and FCP

FICON intermix allows customers to run together both FICON and FCP through a shared director-class IBM TotalStorage SAN Switch.

Cascaded FICON support

FICON support of cascaded directors means that a Native FICON (FC) channel or a FICON CTC can connect a server to a device or other server via two same vendor directors in between. This support is for a two switch, single hop configuration only.

To enable Cascaded FICON support function, you need a FICON license and Secure Fabric OS. Cascaded FICON support is available in two Directors per fabric.

8.6 Fabric management

The switch can be managed using several remote and local access methods. Telnet, SNMP, and Web Tools require an IP network connection to the switch; either out-of-band via switch Ethernet port, or in-band via the Fibre Channel. The switch must be configured with an IP address fitting into the environment's IP addressing scheme.

In Table 8-1 we show a comparison of the access methods.

Table 8-1 Comparison of management access method

Management method	Description	Local	In-band (Fibre Channel)	Out-band (Ethernet)
Serial Port	CLI locally from serial port on the switch	Yes	No	No
Telnet	CLI remotely via Telnet	No	Yes	Yes
SNMP	Manage remotely using the simple network management protocol (SNMP)	No	Yes	Yes
Management Server	Manage with the management server.	No	Yes	No
SES	Manage through SCSI-3 enclosure services	No	Yes	No
Web Tools	Manage remotely through graphical user interface	No	Yes	Yes

8.6.1 Web Tools

Web Tools (also called the IBM TotalStorage StorWatch Switch Specialist) is an intuitive graphical user interface (GUI) which allows network managers to monitor and manage SAN fabrics consisting of switches using a Java-capable Web browser from standard desktop workstations. By entering the network address of any switch in the fabric, the built-in Web server automatically provides a full view of the switch fabric. From that switch, the administrator can monitor the status and perform administration and configuration actions on any switch in the SAN.

Web Tools can manage the switches in the fabric either using in-band Fibre Channel connections or out-of-band Ethernet connections.

To increase SAN management security, Web Tools can operate over a secure browser using the Secure Sockets Layer (SSL) protocol. This protocol provides data encryption, server authentication, message integrity, and optional client authentication for TCP/IP connections. Because SSL is built into all major browsers and Web servers, installing a digital certificate activates the SSL capabilities.

Central status monitoring

Web Tools enables management of any switch in the fabric from a single access point. Using a Web browser, Web Tools is quickly accessed by simply entering the name or IP address of any switch. The Web Tools menu then appears in the Web browser's window, where information about all switches can be retrieved.

The Web Tools menu includes the following views:

- **SAN Fabric View:** Displays all switches in the fabric on a single screen. This graphical display shows all switches currently configured in the fabric and provides a launch point for monitoring and administrating any switch in the SAN. It scales well to large fabrics via a Summary View, which can show twice as many switches as the default detail view.
- **Fabric Event View:** Displays events collected across the entire fabric from the built-in messaging system on each switch, or more detailed and managed information provided by Fabric Watch, an optional feature. Fabric events may be sorted by key fields such as date-time, switch source, or severity level.
- **Fabric Topology View:** Summarizes the physical configuration of the fabric from the perspective of the local domain (the domain of the switch entered as a URL in the Web browser).
- **Name Server View:** Displays information about all hosts and storage devices that are currently registered in the fabric. The Name Server Table is automatically updated when new hosts or devices join the fabric.

- **Zone Admin View:** Administrative privileges are required to access this view. The SAN administrator will manage the switch configuration by menu selection, including a check for possible zoning conflicts before applying the changes to the switch.
- **Detail & Summary View:** Select this to view either the Summary or Detail version of the Fabric View. The Summary version shows abbreviated switch panels. The default view is Detail.
- **Refresh View:** Select this to update the Fabric View to display the latest changes. The Refresh button icon will flash when there have been changes. The Refresh button is only available on switches running Fabric OS 4.x. and higher.

Switch access

From the fabric view, SAN administrators click on any switch icon to establish communication with individual switches for in-depth monitoring or to access configuration options. Individual switch views include:

- The Switch View is an active point-and-click map of the selected switch. Each port icon displays current port status. A click on a port takes the user to the Port Detail View. The states of power supply, fan, and temperature health are updated dynamically. Tool Icons in the Switch View permit direct access to the Event View, the Administrative View, the Performance View, the Fabric Watch Configuration Page (if licensed), the Administrative View, and the Switch Beaconing function.
- The Event View provides a sortable view of all events reported by the switch.
- The Performance View graphically portrays real-time through-put information for each port and the switch as a whole.
- The Telnet icon provides an interface to Telnet functions to perform special switch functions and diagnostics via a Command Line Interface.

Central zoning administrative control

For multi-switch fabric configurations that include the zoning feature, Web Tools enables users to update zoning functions through a graphical user interface.

Fabric OS instantly distributes zoning configuration changes to all switches in the fabric. In addition, users of QuickLoop may use WebTools to configure QuickLoop and integrate QuickLoop with zoning.

Administration and configuration

With Web Tools, SAN administrators can configure and administer individual ports or switches. Web Tools provides an extensive set of features, which enable

SAN administrators to quickly and easily perform the major administrative functions of the switch, such as these:

- Configuring individual switches' IP addresses, switch name, and SNMP settings
- Rebooting a switch from a remote location
- Upgrading switch firmware and controlling switch boot options
- Maintaining administrative user logins and passwords
- Controlling individual ports
- Managing license keys
- Updating multiple switches with similar configurations

8.6.2 Advanced Performance Monitoring

Advanced Performance Monitoring is essential for optimizing fabric performance, maximizing resource utilization, and measuring end-to-end service levels in large SANs. It helps to reduce total cost of ownership (TCO) and over-provisioning while enabling SAN performance tuning and reporting of service level agreements. Advanced Performance Monitoring enables SAN administrators to monitor transmit and receive traffic from the source device all the way to the destination device. Single applications such as Web serving, databases, or e-mail can be analyzed as complete systems with near-real-time performance information about the data traffic going between the server and the storage devices. This end-to-end visibility into the fabric enables SAN administrators to identify bottlenecks and optimize fabric configuration resources.

Advanced Performance Monitoring supports loop, and switched fabric topologies.

Here are some examples of what can be monitored using Telnet:

- **AL_PA monitoring:** Provides information regarding the number of CRC errors in Fibre Channel frames in a loop configuration. It collects CRC error counts for each AL_PA attached to a specific port.
- **End-to-End monitoring:** Provides information regarding performance between a source and a destination on a fabric or a loop. Up to eight device pairs can be specified per port. For each of the pairs, the following information is available:
  - CRC error count on the frames for that device pair
  - Fibre Channel words that have been transmitted through the port for them
  - Fibre Channel words that have been received by the port for them
- **Filter-based monitoring:** Provides information about a filter's hit count. All user-defined filters are matched for all FC frames being transmitted from a port. A filter consists of an offset (byte offset into the FC frame that is) and up

to four values. A filter will match, if all the values specified are found in the FC frame at the specified offset.

You can also administer and monitor performance using Web Tools, if a Web Tools license is also installed. The enhanced Advanced Performance Monitoring features in Web Tools provide:

- Pre-defined reports for AL_PA, end-to-end, and filter-based Performance Monitoring
- User-definable reports
- Performance canvas for application level or fabric level views
- Configuration editor (save, copy, edit, and remove multiple configurations)
- Persistent graphs across reboots (saves parameter data across reboots)
- Print capabilities

Advanced Performance Monitoring makes powerful underlying capabilities simple and easy to use. An enhanced graphical user interface launched from Web Tools gives administrators at-a-glance information needed to anticipate and resolve problems. Administrators can display up to eight performance graphs on a single user-defined management canvas.

Different canvasses can address different users, scenarios, or host applications. Saved canvas configurations enable administrators to change views quickly and easily. Because there is no need to identify a single management console, administrators can access and run the tool from any switch using the Web Tools browser at any location. Moreover, setting up end-to-end monitoring is straight-forward, even for large SAN configurations. To further improve productivity, administrators can use powerful sort, search, and selection features to identify source-to-destination device pairs, dragging and dropping them from the topology tree.

A rich set of predefined graphs are provided for the most common tasks. In addition, administrators can customize predefined performance graphs on virtually any parameter, and add them to canvas configurations. They can also set up and generate printouts or reports in minutes by using previously saved or customized layouts, along with drag-and-drop screens.

Advanced Performance Monitoring can be implemented and used on any IBM 2 Gb/s switch. The Performance Monitoring features can be used as long as the data path of the target flows through a switch that has Frame Filtering capabilities. Existing switches do not need to be replaced or modified.

8.6.3 Fabric Watch

Fabric Watch health enables switches to continuously monitor the health of the fabrics, watching for potential faults based on defined thresholds for fabric

elements and events, so making it easy to quickly identify and escalate potential problems. It monitors each element for out-of-boundary values or counters and provides notification to SAN administrators when any exceed the defined boundaries. SAN administrators can configure which elements, such as error, status, and performance counters within a switch, are monitored.

Accessing Fabric Watch

Fabric Watch runs on Fabric OS since version 2.2, and can be accessed through either Web Tools, Telnet interface, SNMP-based enterprise manager, or by modifying and uploading the Fabric Watch configuration file to the switch. It is designed for rapid deployment: simply enabling Fabric Watch (it comes with pre-configured profiles) permits immediate fabric monitoring. It is also designed for rapid custom configuration.

Range monitoring

With Fabric Watch, each switch continuously monitors error and performance counters against a set of defined ranges. This and other information specific to each monitored element is made available by Fabric Watch for viewing and, in some cases, modification.

Terminology: The set of information about each element is called a *threshold*, and the upper and lower limits of the defined ranges are called *boundaries*.

Fabric Watch monitors the following elements:

- Fabric events:
  - Topology reconfigurations
  - Zone changes)
- Switch environment:
  - Fans
  - Power supplies
  - Temperature)
- Ports:
  - State changes
  - Errors
  - Performance
  - Status of smart GBICs (Finisar SMART GBICs FTR-8519-3)

Fabric Watch lets SAN administrators define how often to measure each switch and fabric element and specify notification thresholds. Whenever fabric elements

exceed these thresholds it is considered as an event. Fabric Watch automatically provides an event notification, they come in two flavours:

- **Continuous Alarm:** Provides a warning message whenever a threshold is breached, and it continues to send alerts until the condition is corrected. For example, if a switch exceeds its temperature threshold, Fabric Watch activates an alarm at every measurement interval until the temperature returns to an acceptable level.
- **Triggered Alarm:** Generates one warning when a threshold condition is reached and a second alarm when the threshold condition is cleared. Triggered alarms are frequently used for performance thresholds. For example, a single notice might indicate that port utilization exceeds 80 percent. Another notice would appear when port utilization drops 80 percent.

These alarms (one or more) generated for the relevant threshold can have three consequences:

- **SNMP trap:** Following an event, Fabric Watch can transmit critical event data as an SNMP trap. Support for SNMP makes Fabric Watch readily compatible with both network and enterprise management solutions.
- **Entry in the switch event log:** Following an event, Fabric Watch can add an entry to an individual switch's internal event log, which stores up to 256 error messages.
- **Locking of the port log to preserve the relevant information:** Following an event, Fabric Watch can add an entry to an individual switch's internal port log and freeze the log to ensure detailed information is available.

Integration with existing management tools

SAN administrators can easily integrate Fabric Watch with existing enterprise systems management tools. Fabric Watch is designed for seamless interoperability with:

- **SNMP-based Enterprise Managers:** The Fabric Watch Management Information Base (MIB) lets system administrators configure fabric elements, receive SNMP traps generated by fabric events, and obtain the status of fabric elements through SNMP-based Enterprise Managers.
- **Web Tools:** By running Fabric Watch with Web Tools, SAN administrators can configure Fabric Watch and query fabric events from this graphical user interface.
- **Syslog daemon:** Through its integration with the UNIX operating system's standard interface for system logging and events, Fabric Watch will send SAN events into a central network log device.

8.6.4 Fabric Manager

Fabric Manager is a highly scalable, Java-based application that manages multiple switches and fabrics in real time. In particular, Fabric Manager provides the essential functions for efficiently configuring, monitoring, dynamically provisioning, and managing fabrics on a daily basis.

Through its single-point SAN management platform, Fabric Manager enables the global integration and execution of processes across multiple fabrics. Moreover, Fabric Manager helps to lower the cost of SAN ownership by intuitively facilitating a variety of SAN management tasks. As a result, Fabric Manager provides a flexible and powerful tool optimized to provide organizations with rapid access to critical SAN information across multiple fabrics.

Fabric Manager increases the efficiency levels of SAN administrators who are responsible for managing multiple SANs. With the unique ability to provide real-time information and streamline SAN management tasks, Fabric Manager provides the following capabilities:

- **A single-console global SAN management platform**: Fabric Manager has the intelligence to manage multiple switch elements spanning up to eight fabrics. It dynamically collects (in real time) all SAN fabric elements and portrays them within the single console, allowing intuitive iconic and explorer tree operations
- **Enhanced SAN visibility:** Fabric Manager can globally capture and present reliable status for all SAN objects. Status is projected through-out the entire SAN management environment. This context-sensitive feature enables SAN administrators to dynamically discover and control the status of all components.
- **An intuitive and functional object management platform:** Fabric Manager's visual display works efficiently with multiple SANs, Fabric Manager is a powerful, secure, and highly scalable management platform for configuring and administrating multiple SANs medium to large, and up to 200 switches. Fabric Manager provides the object status of critical fabric elements, such as ISL Trunking and fabric events, capturing this information in real time across multiple fabrics and fabric security levels.

Fabric Manager provides unique and intuitive methods for managing SANs, including these:

- **User-controlled SAN object grouping:** Fabric Manager enables fabric switches to be placed into any logical, user-defined groups, which are then dynamically propagated throughout Fabric Manager. The groups can be utilized at any time to simplify global management tasks, reducing execution time and ultimately lowering SAN management costs.

- **Global password control:** Fabric Manager enables the management of a user-definable set of SAN fabric switch passwords. SAN administrators can utilize these secure and encrypted objects across all secure features within the platform and across logical groups.
- **Advanced license key management:** Fabric Manager can manage license keys across all SAN fabrics under its control. License management is fully integrated with security, group, and password control.
- **Profiling, backup, and cloning:** Fabric Manager enables organizations to capture a profile of a switch within any fabric, back-up the snapshot to a safe place, and compare the backup to the original fabric device. Cloning facilitates the distribution of profiles to switches within the fabric.
- **Highly flexible firmware download:** This feature is dynamically configurable and scalable across logical groups, password controls, multiple fabrics, and SAN infrastructures with multiple security levels. When utilized with sequenced reboot, Fabric Manager provides a fully configurable environment for controlling the Fabric OS download process
- **Tight integration:** Fabric Manager is tightly integrated with all components of the IBM SAN management family and can in some cases extend those products' capabilities (such as Web Tools and Fabric Watch). As a result, Fabric Manager reduces the time and costs of managing SANs.

In-band and out-of-band: Web Tools uses in-band discovery mechanisms (through the SAN network itself) to discover devices. The in-band discovery mechanisms use SCSI inquiry commands. Many simple disk drives are discovered using in-band discovery. Another type involves out-of-band discovery mechanisms using Simple Network Management Protocol (SNMP) capabilities through TCP/IP and then correlating the results. Hosts and storage subsystems usually have out-of-band management capabilities. The out-of-band discovery gathers device and topology information.

Note: Switches can be accessed simultaneously from different connections. If this happens, changes from one connection may not be updated to the other, and some may be lost. Make sure, when connecting with simultaneous multiple connections, that you do not overwrite the work of another connection.

8.6.5 SCSI Enclosure Services

SCSI Enclosure Services (SES) allows an SES-enabled host connected to a fabric switch to manage all switches in the SAN. This is done remotely in-band using a Fibre Channel link. Therefore, SES serves as the access management method of choice for SCSI-based legacy environments where no Fibre Channel IP driver is available. The SES implementation complies with the SCSI-3 protocol standards used for implementing SES.

- Any SCSI-enabled host connected to the fabric can manage any switch.
- There is no single point of failure in the network.
- The SES capability automatically scales without needing additional resources as the fabric enlarges.

Managing a SAN using SES

To manage a SAN using SES, a host must have a Fibre Channel link to a switch in the fabric. The host must support FCP (Fibre Channel Protocol for SCSI-3) and recognize the FCP target at the Management Server well-known address (FFFFFAh). The host needs to perform the normal N_Port login procedure with the Management Server. It may then initiate an appropriate SES request.

Switch identification in SES

A switch is identified at the FCP level by its Logical Unit Number (LUN). To get a list of LUNs (switches) in the network, the FCP host sends a command to LUN 0 of the target at the Management Server well-known address. Thereafter, the host specifies a particular LUN during a management SES request.

Based on the management information obtained from SES, the SES host may perform a configuration, performance, and/or enclosure function on a switch. For instance, it may enable or disable a switch port, take the temperature sensor readings of a switch, or monitor the performance or error counters of a switch port.

SES helps to maintain a highly available environment for databases and business-critical information in distributed storage environments that are exclusively SCSI-based.

SES switch management

SES is an in-band mechanism for managing a switch within a fabric or other enclosures. SES commands are used to manage and sense the operational status of the power supplies, cooling devices, displays, indicators, individual drives, and other non-SCSI elements installed in a switch (enclosure). The command set uses the SCSI SEND DIAGNOSTIC and RECEIVE DIAGNOSTIC RESULTS commands to obtain and set configuration information for the switch.

Initiator communication

SES allows a SCSI entity (or initiator) to communicate with a switch through a standard FCP connection into the fabric. SES does not require supporting another protocol or additional network links such as Ethernet.

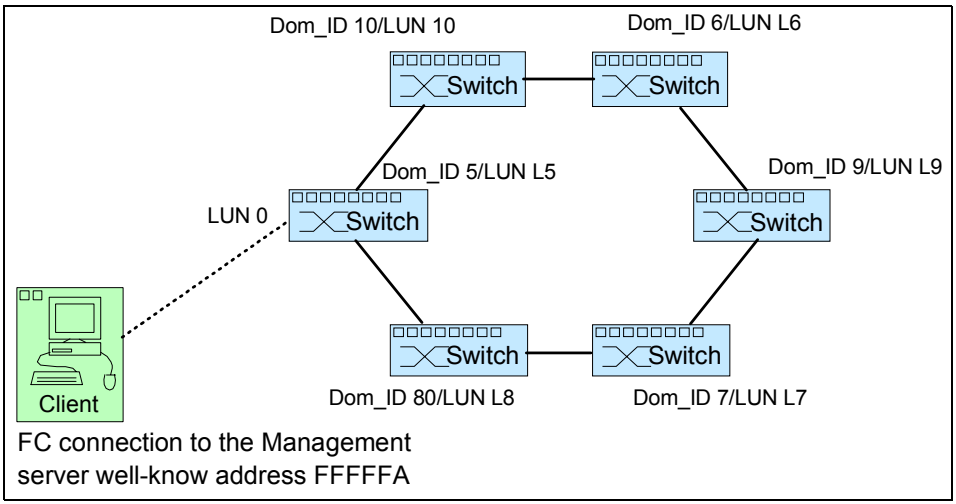


Figure 8-17 SES management

The switch's domain_ID is used as the LUN address to identify each switch including the switch used for access using SES.

Note that the connection to the fabric is through the switch labeled LUN L5 and which is also called LUN 0. The connection to the well-known management address (x'FFFFFA') is always labeled LUN 0 (value in hexadecimal is 00000000 00000000) no matter which switch is used.

Additionally, there can also be a LUN L0 with a hex value of 01000000 00000000. The figure also shows that the left most switch is assigned both LUN L5 and LUN 0. LUN L5 because the switch's domain_ID is L5, and LUN 0 because the client is physically connected to the switch.

8.7 Zoning

Zoning allows you to partition your SAN into logical groups of devices that can access each other. Using zoning, SAN administrators can automatically or dynamically arrange fabric-connected devices into logical groups (zones) across the physical configuration of the fabric. Although zone members can access only

other members in their zones, individual devices can be members of more than one zone.

This approach enables the secure sharing of storage resources, a primary benefit of storage networks. The number of devices that can participate in a zone and the number of zones that can be created are virtually unlimited. SAN administrators can specify zones at a port-level, at server- or storage-level or at department-level. Likewise, zones can vary in size and shape, depending on the number of devices included and the location of the devices. Multiple zones can be included in saved configurations, providing easy control over the enabling or disabling of configurations and avoiding manual changes to specific zones.

Because zone members can access only other members of the same zone, a device not included in a zone is unavailable to members of that zone. Therefore, you can use zones as follows:

- Administer security: Use zones to provide controlled access to fabric segments and to establish barriers between operating environments. For example, isolate systems with different uses or protect systems in a heterogeneous environment.
- Customize environments: Use zones to create logical subsets of the fabric to accommodate closed user groups or to create functional areas within the fabric. For example, include selected devices within a zone for the exclusive use of zone members, or create separate test or maintenance areas within the fabric.
- Optimize IT resources: Use zones to consolidate equipment, logically, for IT efficiency, or to facilitate time-sensitive functions. For example, create a temporary zone to back up non-member devices.

Figure 8-18 shows four zones which allow traffic between two Fibre Channel devices each:

- iSeries server and ESS (Zone A)
- UNIX server and ESS (Zone B)
- zSeries server and ESS (Zone C)
- Windows server and ESS (Zone D)

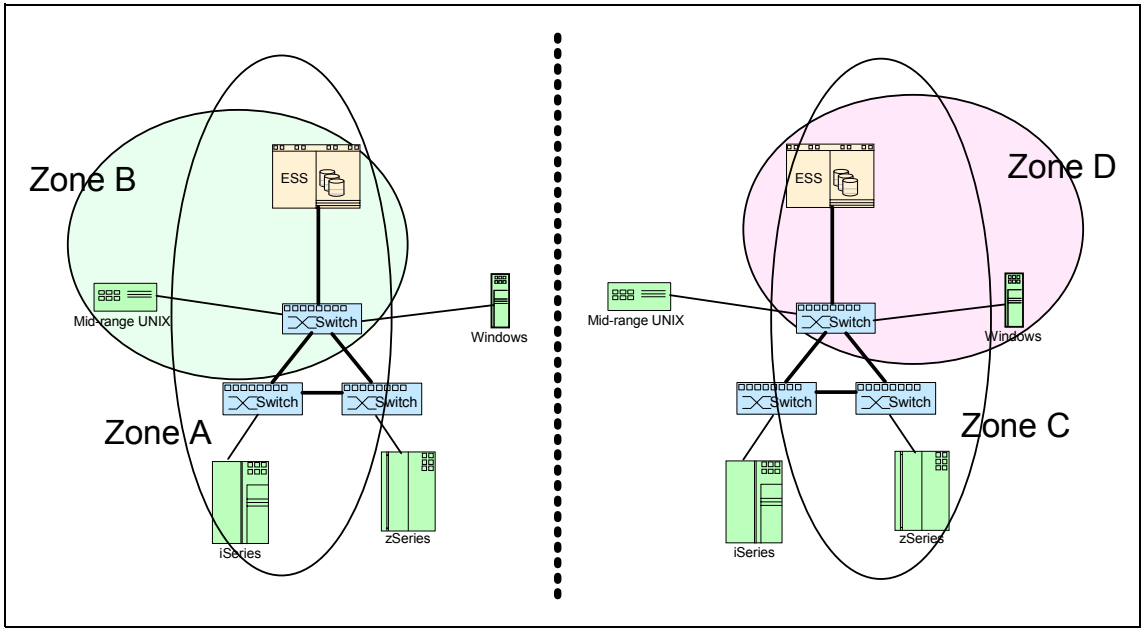


Figure 8-18 IBM SAN Switch zoning

Without zoning, failing devices that are no longer following the defined rules of fabric behavior might attempt to interact with other devices in the fabric. This type of event would be similar to an Ethernet device causing broadcast storms or collisions on the whole network instead of being restricted to one single segment or switch port. With zoning, these failing devices cannot affect devices outside of their zone.

8.7.1 Preparing to use zoning

Before you start using zoning, you should consider the naming conventions that you will be applying to zone related components. In the long run, adherence to a well documented and thought intensive naming convention will make life easier for all concerned.

Before implementing zoning, remember that the zoning process itself has the following advantages:

- Zoning can be administered from any switch in the fabric. Any changes configured to one switch automatically replicate to all switches in the fabric; if a new switch is added to an existing fabric, all zone characteristics are automatically applied to the new switch. Because each switch stores zoning information, zoning ensures a high level of reliability and redundancy.

- Zones can be configured dynamically. Configuring new zones does not interrupt traffic on unaffected ports or devices. Zones do not affect data traffic across Inter-Switch Links (ISLs) in cascaded switch configurations.
- Zoning uses policy-based administration. Because zoning uses policy-based administration, separating zone specification from zone enforcement, you can manage multiple zone configurations and easily enable a specific configuration when it is required. A fabric can store any number of zone configurations; however, only one configuration is active at a time. But, because the configurations are pre-determined and stored, a new configuration can be easily enabled.
- Zoning can be configured and administered using the command line interface (CLI) or Web Tools.

8.7.2 Increasing availability

The easiest way to increase system availability is to prevent failures from ever occurring, typically by monitoring fabric activity and performing corrective actions prior to an actual failure. By leveraging advanced SAN features such as zoning and predictive management, companies can deploy a much more reliable and resilient SAN environment. To help prevent localized failures from impacting the entire fabric, specific parts of SANs can be isolated through the use of zoning, in which defined zones limit access between devices within the SAN fabric. SAN administrators can specify different availability criteria at the connection, node, and network level to address the potential impact of certain types of outages.

For instance, several minor outages in one environment might be much less destructive than a single large outage in another environment — even if the total amount of downtime is the same. The use of zoning helps limit the types of interactions between devices that might cause failures, and thus prevents outages. Especially as companies build larger SANs with heterogeneous operating systems and storage systems, zoning is an effective way to prevent failures.

8.7.3 Advanced zone terminology

A zone generally is a group of fabric-connected devices arranged into a specified group. Any device connected to a fabric can be included in one or more zones. Devices within a zone gain awareness of other devices within the same zone by the RSCN protocol (see 3.1.8, “Adding new devices” on page 66); they are not aware of devices outside of their zone. By these means, zoning provides data exchange between devices in the same zone and prohibits exchange to any device not in the same zone.

Advanced zoning of the 2 Gb/s switches compared with the zoning of 1 Gb/s switches enlarges the range of hardware enforcement and so provides the switch with more security access control functions as before, preventing unauthorized devices from accessing the fabric.

Attention: As zoning functions have developed, some items within the zone terminology have changed slightly with the availability of the 2 Gb/s switches.

Zone members

A zone member must be specified either by:

- Switch port (domain, port)
- World Wide Node Name (WWNN)
- World Wide Port Name (WWPN)
- Alias
- AL_PA in QuickLoop configurations

Zone members are grouped into a zone. To participate in a zone, the members must belong to the appropriate Access Control List (ACL) maintained in the switch hardware. Any number of ports in the fabric can be configured to the zone, so the number of members of a zone is unlimited.

Zones can overlap; a device can belong to more than one zone and a fabric most likely will consist of multiple zones. A zone configuration can include both hard and soft zones. There can be any number of zone configurations resident on a switch; however, only one configuration can be active at a time. Because the number of zones allowable is limited only by memory usage, the maximum number is virtually limitless.

Alias: Aliases exist to make life easier for administration. They are defined by [domain,port] or WWN and provide the opportunity to assign a nickname to a port or a device such as: `Server_Adrian_HBA0` instead of having to deal with a WWN such as `20:00:00:00:c9:2b:db:f0`.

The option to overlap zones is essential for secure storage resource sharing. The ability to share storage resources through logical volume allocation requires that multiple servers share a physical path into the storage. Overlapping zones enable each server (or groups of servers) to reside in a zone with the storage connection while another hardware zone can also see the storage connection. The two distinct zones and all their devices cannot communicate with each other, only with the shared storage system.

Figure 8-19 shows three servers separated by zone A,B and C, able to exchange data with the ESS.

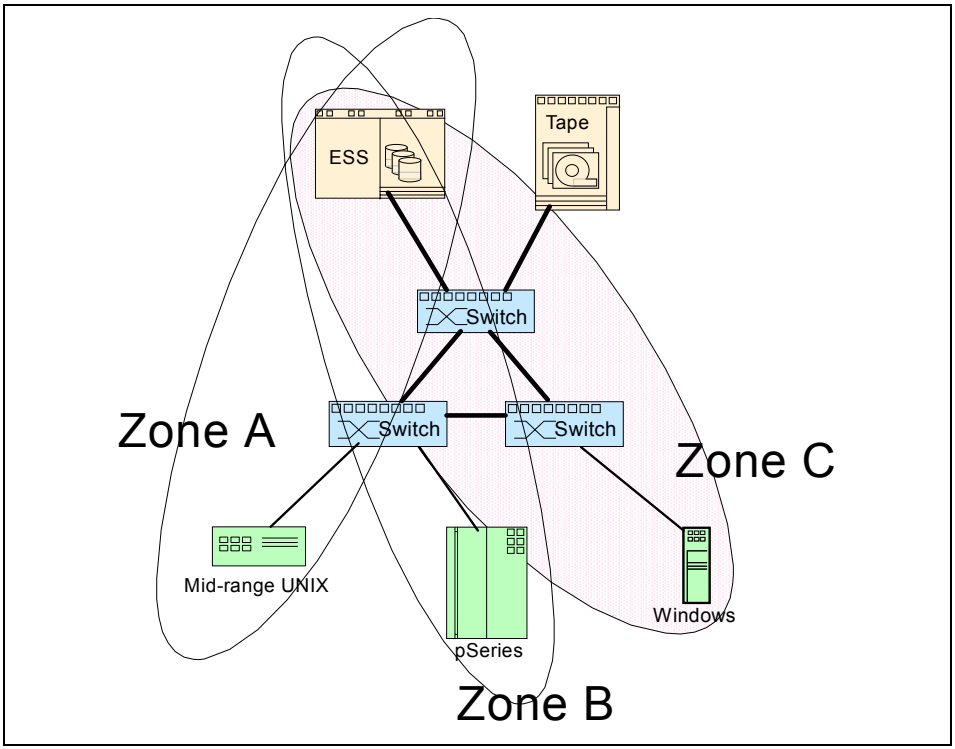


Figure 8-19 Overlapping zones

8.7.4 Zoning types

With advanced zoning, there are two different kinds of enforcement and four different kinds of zoning, as we explain in the following topics.

Hardware enforcement

This is achieved by the following types of zoning:

- **Hard port zoning:** In this case, all members are specified by [domain, port].
- **WWN zoning:** In this case, all members are specified by WW(P)N.

All zone members are specified exclusively either by [domain, port] or by WWN (that includes also WWPN). Hardware enforced zones mean that each frame is checked by hardware before it is delivered to a zone member, and discarded if there is a zone mismatch. Overlapping zones (zone members appearing in multiple zones) are permitted, and hardware enforcement will continue as long as all of the overlapping zones have either all WWNs or [domain,port] entries.

- **Broadcast zoning:** The broadcast zone is a special case; it controls the delivery of broadcast packets within a fabric. Used in conjunction with IP-capable HBAs, a broadcast zone restricts IP broadcast traffic to those elements included in that zone. Only one broadcast zone can exist within a fabric. Broadcast zones are independent of any other zones. A broadcast transfer will be delivered to all ports defined in the broadcast zone even though a port is protected by a hard zone.

Switch port in the M12: Area is considered the absolute port number and ranges from 0 through 63 for each logical switch and defines a specific port for zoning commands. For the ease of description, we still note switch ports [domain, port], although it should be noted as [domain, area] within the M12 with respect to Fabric OS v4.x. For more details on how the port numbering scheme within the M12 works.

Software enforcement

This type is also called name server enforcement:

- Soft port zoning: Here, all members are part of port and WWN zoning. Each port that is part of a port zone and a WWN zone is referred to as a “soft port”. That means that it will now follow name server enforcement rules; however, it is still complemented by hardware-assisted authentication. This means that any access of a FC device to the “soft port” is still checked by hardware and refused when the device is not in the same zone.

Hardware-assisted authentication: As fabric login exchanges continue to be enforced by the ASIC, any attempt by a misbehaving, unauthorized device (PLOGI / ADISC / PDISC / ACC) would get aborted before completion and no SCSI transaction could ever be performed, thereby guaranteeing data access control.

With the 1 Gb/s switches, a hardware zone had to be defined by [domain, port] only — all other configurations were soft zones. Now in mixed fabrics, consisting of both types of 1 Gb/s and 2 Gb/s switches, when there is at least one “soft port” in the configuration, each type of switch stays with its unique zoning method. In 1 Gb/s switches, they use soft zoning; in 2 Gb/s switches, they use soft zoning, hardware-assisted.

8.7.5 Zone configuration

Zones are grouped in a configuration. A zone configuration can carry both hardware and software enforced zones of virtually any amount. Switches can store any number of zone configurations in their memory; however only one configuration can be active at a time. The number of zones allowable is limited only by memory usage.

In Table 8-2 we show a comparison of the different zone types that can be in a zone configuration.

Table 8-2 Different zone types

Feature	Hard zone	Soft zone	Broadcast zone
Naming Convention	Zone names must begin with a letter; may be composed of any number of letters, digits and the underscore character "_". Zone names are case sensitive. Spaces are not allowed within the name.		Special name "broadcast"
Name Server Requests	All devices in the same zones (hard or soft) as the requesting elements		NA
Hardware Enforced Data Transfers	Yes	Yes	Yes
Registered State Change Notification (RSCN)	State changes on any devices within the same zones.		NA
Eligible Devices	All members specified either by [domain, port] or WW(P)N	One member specified by [domain, port] and WW(P)N	Fabric Port Numbers or World Wide Names
Maximum Number of Zones	Limited by total available memory		1
Maximum Number of Zone Members	Limited by total available memory		
Fabric Wide Distribution	Yes	Yes	Yes
Aliases	Yes	Yes	Yes
Overlap	An element can be a member of an unlimited number of zones in any combination of hard and soft zones and be a member of the broadcast zone.		

Managing multiple zones

This is a policy-based administration which allows the user to manage multiple zone specifications and rapidly enable a specific configuration when required.

This provides flexibility in rapidly making SAN configuration changes with minimal impact and risk.

Multiple zone configurations

A fabric can store multiple zone configurations with any one configuration being active at a time. This capability can be used in many ways. For example, a policy can be defined to provide access to the tape library to Windows hosts during the day for continuous backup, but migrate access to UNIX hosts at end of day.

Policy based management

As an example, imagine you have a storage subsystem that under normal circumstances is shared among multiple hosts. Your disaster policy provides that this storage subsystem can be used exclusively by a single host to recover critical data. Using policy-based zoning administration, both zoning configurations are configured and stored in the fabric. In the event of disaster, the SAN administrator would simply enable the pre-configured zoning configuration — a few mouse clicks — and the fabric would automatically enforce your pre-determined policy.

8.7.6 Zoning administration

Zoning administration can be managed either using the Command Line Interface to any switch in the fabric or by using Web Tools. Configuring zones consists of four steps:

- **Create aliases:** The aliases allow you to rapidly give familiar name or group multiple devices into a name. For example, you can create an alias called “NT_Hosts” to define all NT hosts in the fabric.
- **Define zones:** You can create a zone and add members to it. Members can consist of Switch Port Names, WWNs, or aliases. Changes to the zone layout does not take effect until a zone configuration has been defined and enabled.
- **Define a zone configuration:** You can create a zone configuration and add zones to it. This step identifies the zones that should be enabled whenever this configuration is enabled. Changes to the zone configuration will not take effect until that zone is enabled.
- **Enable the zone configuration:** Select the zone configuration to be activated. For hard zones, this action downloads the zone configuration to the switch ASICs and begins the enforcement. For either hard or soft Zones, a State Change Notification (RSCN) is issued to signal hosts to re-query the name server for a list of available devices.

Zoning is a fabric-wide resource administered from any switch in the fabric and automatically distributes itself to every switch in the fabric, and simplifies

administration. Zone definitions and zones configurations are persistent and remain in effect across reboots and power cycles until deleted or changed.

A new switch added to the fabric automatically inherits the zoning configuration information and immediately begins enforcement. The fabric provides maximum redundancy and reliability since each switch stores the zoning information locally and can distribute it to any switch added to the fabric.

8.7.7 QuickLoop

The QuickLoop feature combines arbitrated loop and fabric support for private loop servers. Private-loop storage devices like disk and tape can be connected to and used on each switch without the QuickLoop feature. Address translation for targets (storage devices) is basically implemented in each 2109 OS.

Looplets

A QuickLoop consists of multiple private arbitrated looplets (a set of devices connected to a single port) that are connected by a fabric. All devices in a QuickLoop share a single AL_PA space and behave as if they are in one loop. This allows private devices to communicate with other devices over the fabric, provided they are in the same QuickLoop.

QuickLoop topology has the following characteristics:

- A QuickLoop can include up to two switches and support up to 126 devices.
- Each individual switch can only be included in one QuickLoop.
- A QuickLoop can include either all or a subset of the ports on an individual switch.
- Multiple QuickLoops can exist in a fabric of multiple switches.
- QuickLoop enabled switches can exist in the same fabric as non-QuickLoop enabled switches.
- A device attached to a QuickLoop can communicate with all other devices attached to the same QuickLoop.
- A private device in a QuickLoop can communicate with devices in the same QuickLoop only. Existing PLDA capable host drivers need no modification to perform I/O operations with storage devices.
- Public devices that are arbitrated loop capable are treated as private devices when connected to QuickLoop ports (their fabric login, or “FLOGI,” is rejected).
- QuickLoop supports the use of legacy loop devices, allowing them to be attached to a fabric and operate as if in a Private Loop Direct Attach (PLDA) environment.

- QuickLoop functionality can be enabled or disabled for either the entire switch or for individual ports. When QuickLoop is disabled on an individual port, that port returns to fabric mode.
- Each looplet in a QuickLoop has its own unshared bandwidth and can support transfer rates up to 100 MB/s.
- Multiple devices can communicate simultaneously and at full bandwidth within multiple looplets located in the same QuickLoop.
- If a looplet error is detected, QuickLoop automatically takes the looplet out of service. If the error condition is cleared, the looplet is automatically reinstated.

Private loop migration

QuickLoop provides a potential migration path from deploying a single private loop to deploying a fabric-based SAN. Initially, QuickLoop enabled switches can be used to replace hubs when the SAN is first deployed and only has private devices attached. Then, as the SAN grows, fabric switches can be added without any detrimental effect to the QuickLoop enabled switches.

Address translation

QuickLoop address translation is transparent and requires no actions on the part of the user. It is achieved through hardware translative mode (also known as phantom mode), in which a device not physically located in a looplet is made addressable by a unique AL_PA in that looplet. There are two hardware translative modes available to a QuickLoop enabled switch:

- **Standard translative mode:** Allows public hosts to communicate with private target devices across the fabric.
- **QuickLoop mode:** Allows private hosts to communicate with private target devices across the fabric.

The switch automatically determines and sets the appropriate mode.

QuickLoop and zoning

QuickLoop can be used in conjunction with zoning. Using the products together provides the following additional features:

- AL_PAs from multiple QuickLoops can be used to add members to a zone. This is due to the Zoning ability to name QuickLoops and therefore identify the QuickLoop to which the AL_PA belongs.
- Additional control over access to QuickLoop devices is possible. Fabric devices in a zoned fabric can only access the QuickLoop (and fabric) devices that are in the same zone.

- Zones can be created within QuickLoops. Zoning can be used to partition QuickLoops. This creates “QuickLoop zones” (as opposed to fabric zones), which support identification by either physical port number or AL_PA.

8.8 Switch interoperability

IBM TotalStorage SAN Switch model Fxx, Hxx, and Mxx are OEM products from Brocade are identical to Brocade’s SilkWorm switches:

- IBM TotalStorage SAN Switch B32 — Brocade SilkWorm 4100
- IBM TotalStorage SAN Switch H08 — Brocade SilkWorm 3250
- IBM TotalStorage SAN Switch H16 — Brocade SilkWorm 3850
- IBM TotalStorage SAN Switch F16 — Brocade SilkWorm 3800
- IBM TotalStorage SAN Switch F32 — Brocade SilkWorm 3900
- IBM TotalStorage SAN Switch M12 — Brocade SilkWorm 12000
- IBM TotalStorage SAN Switch M14 — Brocade SilkWorm 24000

In order to refer to the latest compatibility information, we advise you to refer to the Web site:

http://www.storage.ibm.com/ibmsan/products/2109/library.html#support

8.9 Feature codes

IBM TotalStorage SAN Switch 2005-B32

- **FC2210** 2 Gbps SW SFP Transceiver
- **FC2220** 2 Gbps LW SFP Transceiver
- **FC2235** 2 Gbps 35 Km Extended Distance LW SFP Transceiver
- **FC2280** 2 Gbps 80 Km Extended Distance LW SFP Transceiver
- **FC2410** 4 Gbps SW SFP Transceiver
- **FC2214** SW SFP Transceiver - 4 Pack
- **FC2224** LW SFP Transceiver
- **FC5601** Fiber Cable LC/LC 1m multimode
- **FC5605** Fiber Cable LC/LC 5m multimode
- **FC5625** Fiber Cable LC/LC 25m multimode
- **FC5701** LC/LC Coupling Cable 1 m
- **FC5702** LC/SC Coupling Cable 2 m
- **FC5720** 9 Micron Fibre Cable 31m, singlemode SC/LC
- **FC5721** 9 Micron Fibre Cable 31m, singlemode LC/LC
- **FC5722** 9 Micron Fibre Cable 2m, singlemode SC/LC
- **FC7510** B32 Ports 16-23 Activation
- **FC7511** B32 Ports 24-31 Activation
- **FC7512** B32 Ports 16-31 Plant Activation
- **FC7513** B32 Ports 16-31 Field Activation
- **FC7551** B32 Fabric Manager V4 Max Domains
- **FC7552** B32 Remote Switch Activation
- **FC7553** B32 Extended Fabric Activation
- **FC7554** B32 Advanced Security Activation
- **FC7555** B32 Performance Bundle
- **FC7556** B32 Performance Monitor
- **FC7557** B32 Enhanced Trunking

IBM TotalStorage SAN Switch 2005-H08

- **FC2210:** Shortwave SFP Transceiver
- **FC2214:** Shortwave SFP Transceiver Four Pack

- **FC2220:** Longwave SFP Transceiver
- **FC2224:** Longwave SFP Transceiver Four Pack
- **FC2235:** Extended longwave SFP transceiver, 35 Km
- **FC2280:** Extended longwave SFP transceiver, 80 Km
- **FC7701:** Fabric Manager V4, maximum domains
- **FC7702:** H08 Remote Switch activation
- **FC7703:** H08 Extended Fabric activation
- **FC7720:** H08 Full Fabric activation
- **FC7721:** H08 Advance security activation

IBM TotalStorage SAN Switch 2005-H16

- **FC2210:** Shortwave SFP Transceiver
- **FC2214:** Shortwave SFP Transceiver Four Pack
- **FC2220:** Longwave SFP Transceiver
- **FC2224:** Longwave SFP Transceiver Four Pack
- **FC2235:** Extended longwave SFP transceiver, 35 Km
- **FC2280:** Extended longwave SFP transceiver, 80 Km
- **FC7701:** Fabric Manager V4, maximum domains
- **FC7802:** H16 Remote Switch activation
- **FC7803:** H16 Extended Fabric activation
- **FC7820:** H16 Full Fabric activation
- **FC7821:** H16 Advance Security activation

IBM Total Storage SAN Switch 2109-F16

- **FC2210:** Shortwave SFP Transceiver
- **FC2214:** Shortwave SFP Transceiver Four Pack
- **FC2220:** Longwave SFP Transceiver
- **FC2224:** Longwave SFP Transceiver Four Pack
- **FC2235:** Extended longwave SFP transceiver, 35 Km
- **FC2280:** Extended longwave SFP transceiver, 80 Km
- **FC6203:** Additional power supply
- **FC7203:** Fabric Manager V4
- **FC7204:** Fabric Manager V4.1 Upgrade

- **FC7302:** Remote Switch activation
- **FC7303:** Extended Fabric activation
- **FC7421:** Performance Bundle
- **FC7423:** F16 Advance Security activation
- **FC9205:** Non-rack mount install

IBM SAN TotalStorage Switch 2109-F32

- **FC2210:** Shortwave SFP Transceiver
- **FC2214:** Shortwave SFP Transceiver Four Pack
- **FC2220:** Longwave SFP Transceiver
- **FC2224:** Longwave SFP Transceiver Four Pack
- **FC2335:** 35 Km extended distance longwave SFP transceiver
- **FC2380:** 80 Km extended distance longwave SFP transceiver
- **FC7203:** Fabric Manager V4
- **FC7204:** Fabric Manager V4.1 Upgrade
- **FC7423:** F32 Advance Security activation
- **FC7502:** Remote Switch activation
- **FC7503:** Extended Fabric activation
- **FC9205:** Non-rack mount install

IBM SAN TotalStorage SAN Switch 2109-M12

- **FC2310:** M12 Shortwave SFP Transceiver
- **FC2314:** Shortwave SFP transceiver four pack
- **FC2320:** M12 Longwave SFP transceiver
- **FC2324:** Longwave SFP transceiver four pack
- **FC2335:** 35 Km extended distance longwave SFP transceiver
- **FC2380:** 80 Km extended distance longwave SFP transceiver
- **FC3126:** M14 Control Process Upgrade
- **FC3216:** M12 16-Port Switch Blade
- **FC3226:** M14 16-Port Switch Blade
- **FC7250:** Fabric Manager 4.1, 10 domains
- **FC7251:** Fabric Manager 4.1, Max domains
- **FC7252:** Fabric Manager 4.1 Upgrade to Max domains

- **FC7253:** Fabric Manager 3 Upgrade to 4.1 Max domains
- **FC7603:** Extended Fabric activation
- **FC7623:** M12 Advanced Security activation

IBM TotalStorage SAN Switch 2109-M14

- **FC2310:** M12 Shortwave SFP Transceiver
- **FC2314:** Shortwave SFP transceiver four pack
- **FC2320:** M12 Longwave SFP transceiver
- **FC2324:** Longwave SFP transceiver four pack
- **FC2335:** 35 Km extended distance longwave SFP transceiver
- **FC2380:** 80 Km extended distance longwave SFP transceiver
- **FC3226:** 16-Port M14 Switch Blade
- **FC7250:** Fabric Manager 4.1, ten domains
- **FC7252:** Fabric Manager 4.1, upgrade ten to maximum domains
- **FC7602:** Remote Switch activation
- **FC7603:** Extended Fabric activation
- **FC7623:** M14 Advance Security activation

9

IBM TotalStorage SAN n-type family and replication appliance

The IBM TotalStorage SAN n-type family of enterprise director and switch products expand the IBM TotalStorage portfolio. The IBM TotalStorage n-type family is designed to support ultra-scalability and enterprise level availability and performance, flexible management system architecture, FCP and FICON, operation, FICON Cascading, and FICON CUP support, as well as compatibility with the CNT FC/9000 and UltraNet Multi-service Director (UMD) (IBM 2042) product lines.

Figure 9-1 on page 346 demonstrates the new storage networking vision of the CNT Ultranet Product family. The Ultranet Replication Appliance offers the storage and extension services currently, and they will be ported to the SAN256N plug-in module in the future.

© Copyright IBM Corp. 2004. All rights reserved.

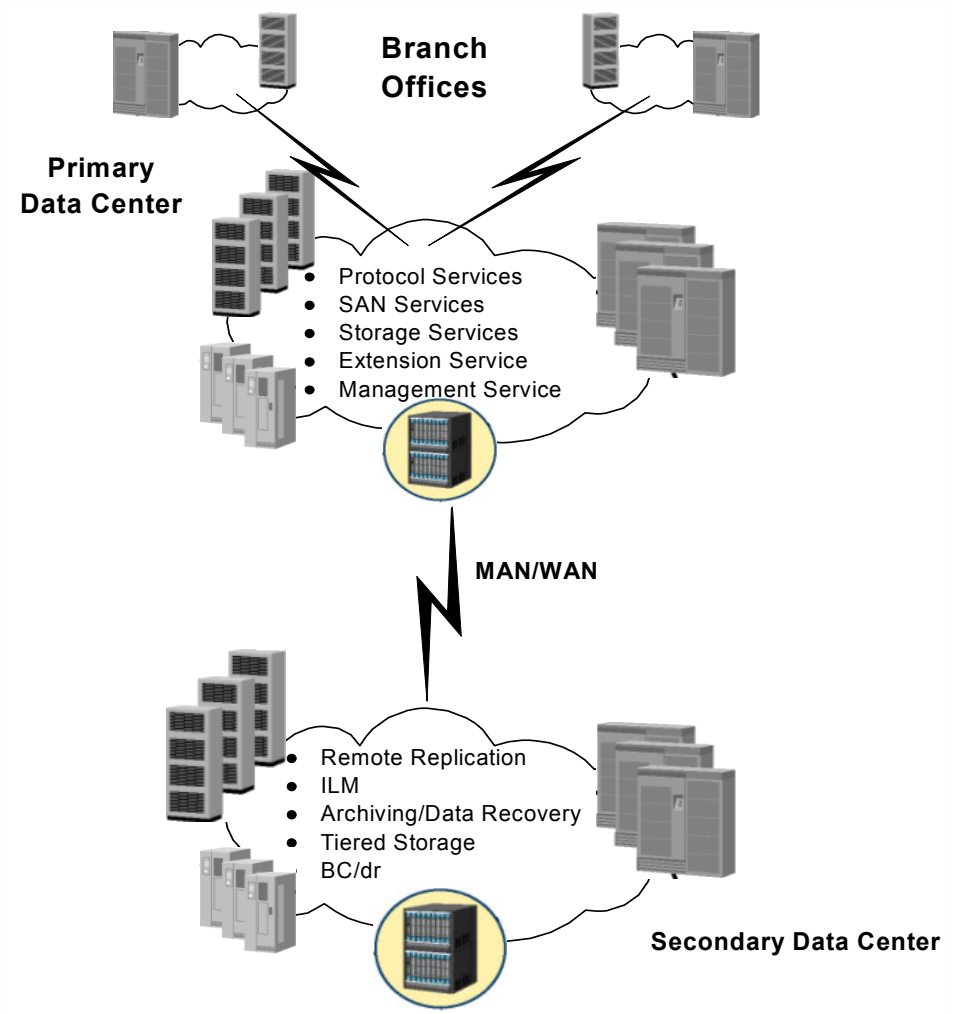


Figure 9-1 CNT Ultranet Multi-services Director Heterogeneous Extended Fabric

Figure 9-2 on page 347 shows an example of the FC/9000 director in a local fabric. The FC/9000 supported FCAL links only with the older 1 Gbps I/O module. In this example, the FCAL public and private loop devices are connected to the older 1 Gbps module. For newer 2 Gbps I/O modules or the SAN256N director, the FCAL devices would be connected to a switch used as an edge switch via an ISL to the FC/9000 or SAN256N.

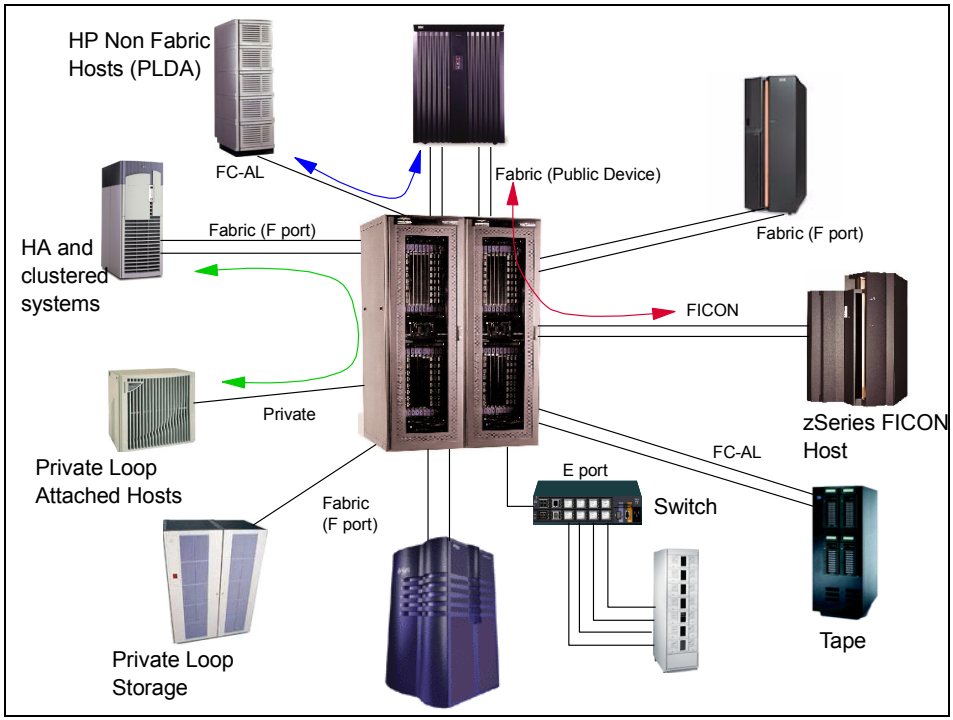


Figure 9-2 CNT FC/9000 director based heterogeneous storage networking

9.1 CNT UltraNet family product description

In these sections we describe the CNT UltraNet family of products.

9.1.1 CNT Ultranet Edge Storage Router

The CNT Ultranet Edge Storage Router enables any-to-any connectivity of Fibre Channel or FICON storage area network islands over an IP, ATM, or SONET infrastructure. The IP Storage Architecture of the router enables intelligent data movement across the IP network with predictable throughput, efficient utilization of bandwidth, and data integrity with no distance limitation.

The CNT Ultranet Edge Storage Router allows geographically distributed SAN locations to replicate data concurrently, ensuring consistency of information. The unique compression capabilities can compress the data transferred anywhere from 2:1 to 20:1 to reduce bandwidth costs.

Figure 9-3 on page 348 demonstrates the usage of the Ultranet Edge Storage Router to expand the Storage Area Network across a Wide Area Network.

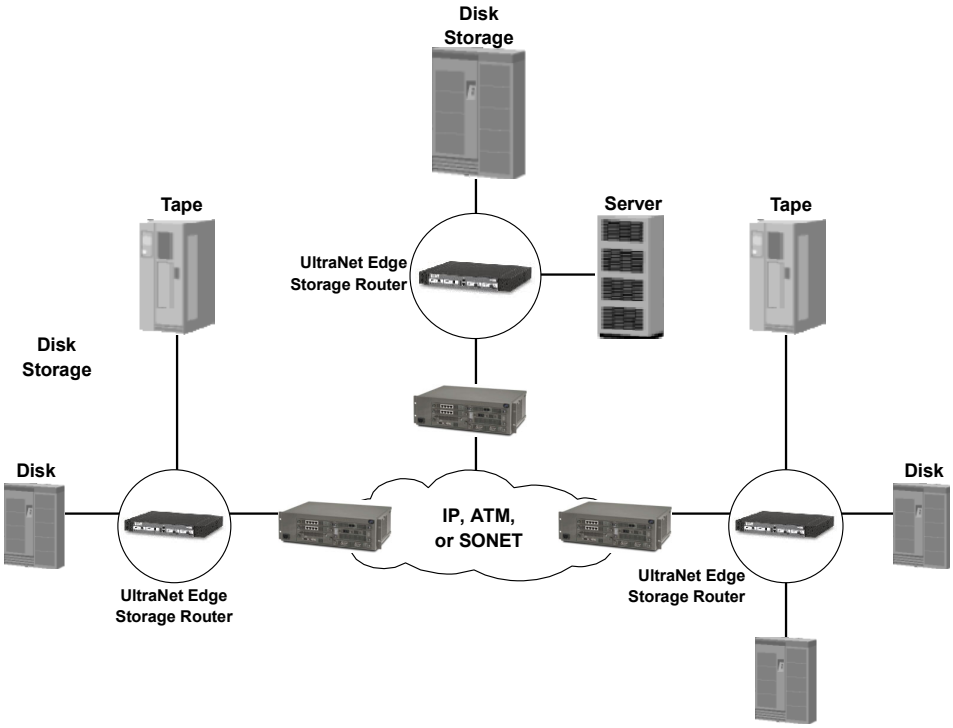


Figure 9-3 Ultranet Edge Storage Router extending the fabric across WAN

The router can be used as a tape controller in the primary site, eliminating latency issues with remote backup. The Router emulates a buffered device controller, allowing backup in multiple locations.

The router utilizes a store and forward architecture with FC and IP payload matching for maximum efficiency. Data integrity is assured by the use of end-to-end hardware CRC. The router offers traffic rerouting to alternate paths in case of line failure.

The router includes support for various IP technologies such as 100BaseT and Gigabit Ethernet, and ATM/SONET technologies including OC-3/12 ATM and packet over SONET/SDH. Maximum performance for standard IP packet sizes is assured for ease of interoperability.

Figure 9-4 on page 349 shows the CNT Ultranet Edge Storage Router.

Figure 9-4 CNT Ultranet Edge Storage Router

9.2 CNT Ultranet Replication Appliance

The CNT Ultranet Replication Appliance (URA) consists of hardware, software, and services that provide flexible SAN data replication for any Fibre Channel storage environment over any distance. This version of the CNT Ultranet Storage Appliance is specifically tuned for IBM SAN Volume Controller support.

The URA replication server is a dedicated storage networking appliance which attaches between application servers and the IBM SAN Volume Controller. It is a dual-processor server running a special LINUX kernel. There are numerous features and solution packages available to customize the distance replication for the customer.

The URA supplies the following functions for the SAN environment:

- Asynchronous mirroring
- Snapshot Replication with Delta
- Heterogeneous Storage Support for Mirroring

Figure 9-5 on page 350 shows a picture of the CNT Ultranet Replication Appliance.

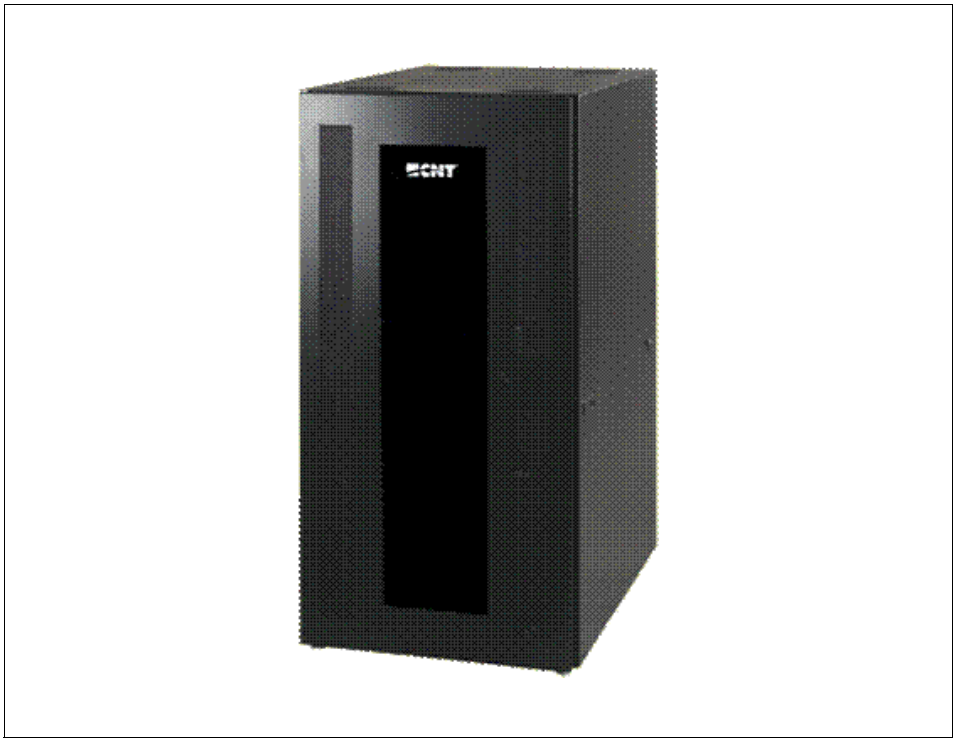


Figure 9-5 CNT Ultranet Replication Appliance

The CNT Ultranet Replication Appliance provides data replication services. It is inserted in the data path (inband) between the open system hosts and the storage, or the open system hosts and the SAN Volume Controller . The SVC vdisks (LUNs normally presented to the open system hosts) are presented to the URA. The URA then presents a view of the LUNs to the open system hosts.

The URA also needs a disk buffer allocated on the back end storage. Size is determined by distance and bandwidth requirements.The URA is then able to both locally replicate and/or remotely replicate the intercepted LUNs across the wide area network to a geographically separate location at extreme distance. The remote location could contain a full URA capable of bi-directional replication, or just an Ultranet Edge Storage Router which allows unilateral replication. Both local and remote storage can consist of SAN Volume Controller managed storage pools, or heterogeneous DASD. Therefore the local storage could be high performance ESS storage, while the remote site less costly lower performance storage.

There are two main components of the CNT Ultranet Replication Appliance:

Asynchronous Replication Model

The application software supplies a policy-driven replication process which can switch between continuous replication and delta replication based on the write speed. This application runs in the network-based appliance.

Local and remote solutions can be created using the following functions:

- Long distance asynchronous mirroring with minimal exposure to data loss.
- Acknowledgement of data writes occur immediately to the host application while the data is buffered for sending to the remote site. Data is buffered, then sent to the local and remote sites in the order received.
- Snapshot copy which can use application-aware database agents to ensure the transactional integrity of the data copies
- The URA can track multiple point-in-time images of any data volume and roll back to a previous version of the data. The user can define snap intervals called TimeMarks associated with a time stamp that is tracked by URA. If its necessary to recreate or restore the data volume for any reason, it can be restored instantly based on any of the existing TimeMarks.
- Replication of incremental changes
- A copy of the data volume can be maintained in either a local or remote site. The customer can establish policies based on time or data change to trigger the replication, which uses snapshot to send the delta changes to the replica volume. Point-in-Time replication can be done locally through the local fabric or remotely over the Ultranet Edge Storage Router.

Network Services Model

This supplies wide area networking technology with hardware-based data compression, data management and flow control, and network error detection and recovery including CRCs (cyclic redundancy checks). This feature can integrate the current IP, ATM, or SONET infrastructure with the SAN islands, to extend replication capability to unlimited distances. 2043-NS1 provides support for OC-3 ATM/POS (packet over SONET) and OC-12 ATM/POS Services. The Appliance provides distance-independent application performance.

Figure 9-6 on page 352 shows the Ultranet Replication Appliance with the front cover open. The picture shows the following hardware components:

- Ultranet replication server
- Ultranet Management Services Server
- Ultranet Edge Storage Router
- Brocade Fibre Channel Switch (optional)
- Netgear Ethernet Switch
- Rack-mount LCD Monitor and Keyboard Drawer

- Keyboard Video Mouse (KVM) Switch

Figure 9-6 Ultranet Replication Appliance with front cover open

CNT URA cabinet

The URA cabinet feature is a 19-inch rack with a 16-port Ethernet switch for communication between the building blocks, a LCD Monitor/ Keyboard Drawer, a Keyboard Video Mouse Switch, and a power distribution system. Cabinet space is sufficient to integrate two 2U replication servers, two 1U Edge WAN routers, a 1U network manager server, a 1U backup UPS power supply and two 1U IBM SAN Switches.

- Dimensions: 129.5 cm H x 61.0 cm W x 91.4 cm D
- (51 in x 24 in 36 in)
- Weight: 135 - 270 kg (300 - 600 lb.)

CNT URA solution packages

IBM and CNT offer the following preselected disaster tolerant configurations to ease the pain of configuring a solution for each customer:

- Dark recovery site package:

This solution provides global disk vaulting for customers needing to replicate IBM SAN Volume Controller volumes from the primary site to the recovery site, over any major WAN transport. Recovery requires restoring primary site volumes.

- Primary site with dual URA replication servers, WAN router and network management server
- Recovery site with WAN router
- Warm recovery site package:

This solution provides global disk mirroring from the primary site to the recovery site over any major WAN transport, for customers needing to rapidly switch primary site application server processing to the recovery site.

- Primary site with dual URA replication servers, WAN router and network management server
- Recovery site with one URA replication server and WAN router
- Hot recovery site package:

This solution provides bidirectional global disk mirroring between the primary site and the recovery site over any major WAN transport, for customers needing to rapidly switch primary site application server processing to the recovery site and to switch recovery site application server processing to the primary site.

- Primary site with dual URA replication servers, WAN router and network management server
- Recovery site with dual URA replication servers and WAN router

9.3 IBM TotalStorage SAN256N director

The IBM TotalStorage SAN256N director (IBM 2045-N16) is designed to simplify consolidation of Fibre Channel and mainframe FICON storage area networks by utilizing logical domain capability. The SAN256Nfamily provides a resilient storage infrastructure with the capability to extend across multiple data centers and network the entire global storage enterprise. The SAN256N directors offer better high availability than current directors by including active-active load sharing and fail-over capabilities, completely hot swappable components, advanced diagnostics and proactive reporting capabilities.

Figure 9-7 on page 354 shows a heterogeneous fabric consisting of two Ultranet Multi-service Directors with their associated servers, disks, and tape storage, with the fabric extended across a wide area network for mirroring or replication purposes.

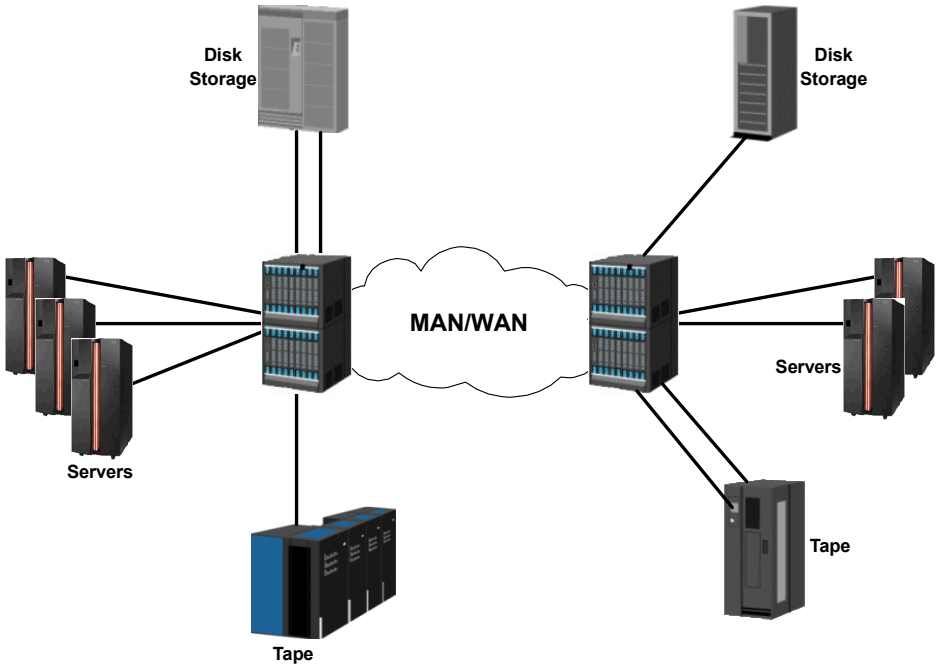


Figure 9-7 CNT Ultranet Multi-services Director Heterogeneous Extended Fabric

The base system provides a redundant bladed switch design for high availability, and 32 ports that can be upgraded in 16-port increments to 256 ports. The SAN256N director is a fourth generation design utilizing an expandable switching

core with 2 Gbps non-blocking performance on all ports. The new UltraScale architecture is extendible for multiprotocol and greater performance. Current protocols include FICON(CUP), Fibre Channel (FCP), and iSCSI. The basic design provides 5 Tbps of system throughput for up to 512 non-blocking ports at 4 Gbps data rates. The design is protocol and speed agnostic, non-blocking and low latency, and delivers predictable and consistent performance across every port regardless of load.

Figure 9-8 on page 355 shows the Model N16 of the CNT Ultranet Multi-service Director.

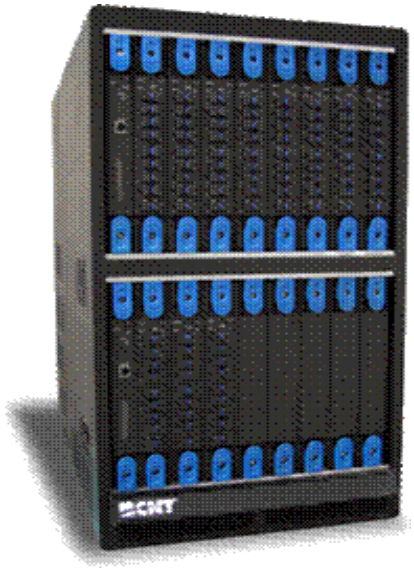


Figure 9-8 CNT Ultranet Multi-service Director Model N16

The SAN256N provides 256 ports in 18U height with a new cabinet design 2042-C80 which can contain one or two SAN256Ns. This provides 512 ports in a single footprint. This is a 4:1 density improvement compared to the FC/9000. Alternately, the SAN256N can be installed in the 2042-C40 FC/9000 cabinet to utilize extra space or upgrade an existing cabinet (one or two SAN256Ns in one 2042-C40). The SAN256N director is interoperable with the FC/9000 director.

The SAN256N I/O module contains 16 ports at 2 Gbps. There are both shortwave and longwave 2 Gbps SFP transceivers available in 4 packs. There are also extended longwave 35Km and 80Km SFP transceivers available to provide disaster tolerant metro mirroring.

The SAN256N physical director may be partitioned via the blades into a maximum of four logical domains (logical switches). A single domain may include a maximum of 256 ports. A logical domain can consist of two or more I/O modules. For example, a fully populated 256-port SAN256N could be partitioned into two 64-port logical switches and one 128-port logical switch. Partitioning also provides the ability to connect the logical domains together with virtual interswitch links (ISLs). Virtual ISLs require no director ports or physical cables, and have no ISL latency.

Logical domain partitioning requires the SAN services bundle. This feature allows the SAN to be partitioned into isolated logical switches if desired so that the customer can separate different protocols like FICON and FCP instead of utilizing physically separated switches. In addition, each logical partition can be managed independently, and firmware upgrades of its components can be done independently with respect to other partitions. At GA of the SAN256N, logical domains could be created only on adjacent blades. Logical domains will be extendible on any blade in the very next release.

New management software integrates a variety of services as needed by the enterprise, including protocols, security, WAN and MAN extension, storage services like virtualization, and management services. Four levels of Quality of Service (QoS) allows the user to set policies for prioritization of traffic between ports giving the users more control over data flows.

The inVSN Storage Network Manager will support the new SAN256N director and seamlessly integrate all CNT products on the SAN/WAN for an end-to-end view of the application across the global infrastructure. The inVSN software will collect and analyze network information to aid in problem identification and isolation.

The SAN256N roadmap includes 4 and 10 Gbps I/O modules, multiprotocol, and advanced intelligence features. Near term features include an N02 model for multiprotocol requirements such as 1 Gbps Ethernet, iSCSI, and FCIP. Other new features will include Flow Group for FICON cascading performance, global replication for the IBM SAN Volume Controller, and new management services for global infrastructure management.

9.3.1 Ultranet Multi-service Director Components

The SAN256N contains the following components documented in the *UMD Installation & Maintenance Guide,* 3016837-HIG Rev X1. The components that plug into the front of the SAN256N chassis are called both modules and blades in the IBM and CNT literature. The terms are interchangeable.

TFIO 16-Port I/O Module

The SAN256N Fibre Channel I/O module (TFIO) is a sixteen-port input/output line card that plugs into a slot in the front of the chassis. The SAN256N chassis can support up to sixteen TFIO blades, giving a total port count of 256 ports per chassis. Each TFIO module has sixteen (16) Fibre Channel ports to accept qualified SFP Shortwave (850nM) and Longwave (1310-1550nM) modules. The SFP modules can operate at either 1.0625 Gbps or 2.125 Gbps. All ports are self-configuring and have 100 buffer credits per port providing support for a distance of up to 80 km distance with extended distance long-wave SFPs.

Figure 9-9 on page 357 shows the SAN256N TFIO 16 port I/O Module.

Figure 9-9 CNT SAN256N TFIO Module

The TFIO module front panel contains indicators, switches, and mirror ports. The front panel contains LED status indicators showing the current state of the module, and for each of the individual I/O ports. The I/O port LEDs will be solid or blinking green for 2 Gb/s links, and solid or blinking yellow for 1 Gb/s fibre speed. Red indicates an error. The board fault indicators show the power status, heartbeat, fault status, and whether online to the fabric. There are also two mirror ports available, one for viewing the transmit stream of the port to be diagnosed called Mirror Transmit (MTx), and one for viewing the receive stream called Mirror Receive (MRx). Both ports transmit data only which allows connectivity to a Fibre Channel analyzer. Finally, there is a reset switch on the TFIO module that will force a hard reset, which is not recommended.

SAN256N TCM Control Module

The SAN256N Control Module (TCM) provides overall control, monitoring, and external communications for the SAN256N system. The module acts as a proxy for all external communications destined for other modules in the system, utilizing a 100Mb/s ethernet connection on the front panel. The TCM provides control and management of communications and the system clock for the TCM, the TFIO and the TSW modules. The SAN256N requires a redundant SAN256N Control Module (TCM) in the system acting as a standby.The TCM monitors and reports the general health of the following components in the SAN256N system:

- TFIO Modules
- TSW Modules
- TCM Modules
- Power Supply Assemblies
- Fans
- Upper and Side Assemblies.

Figure 9-10 on page 358 shows the SAN256N Control Module (TCM). Note the RJ-45 100 Mbps Ethernet adapter for connecting to the Enterprise Manager PC. There are also LED indicators and an LCD display. The Board fault indicators show the power status, heartbeat, fault status, and whether online to the fabric. Three additional LEDs indicate system power and fan status.

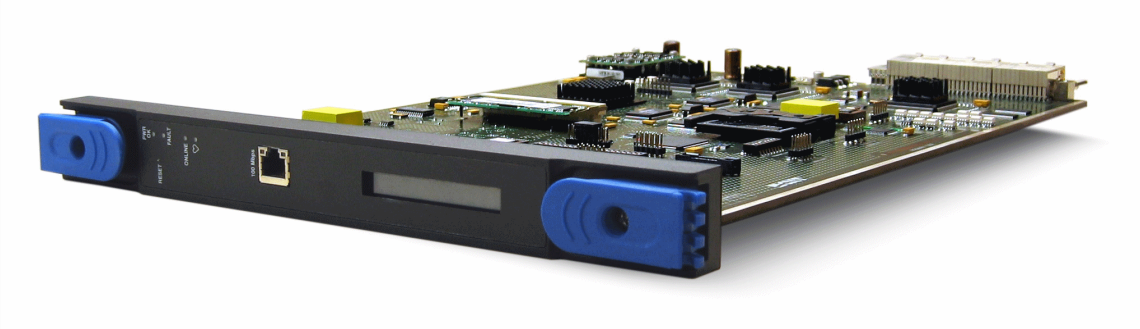


Figure 9-10 CNT SAN256N Control Module (TCM)

The 8-character alphanumeric LCD display on the TCM front panel can display system information, such as:

- Switch Name
- IP Address
- Status of the Field Replaceable Units (FRUs):
  - TFIO Modules
  - TSW Modules
  - TCM Modules
  - Power Supply Assemblies

– Fans, Upper and Side Assemblies

SAN256N TSW Switch Module

The SAN256N Switch Modules (TSW) provides the physical and logical connectivity between the TFIO modules installed in the SAN256N chassis. The TSW modules provide non-blocking any-to-any switching capabilities between the switch fabric ports. The SAN256N needs only two TSW modules to provide the full bandwidth required to support 256 ports at 4Gbps, but is shipped with three modules for full availability. The three TSW modules share the fabric load in an active-active configuration. No failover is needed is needed should a TSM module fail. A fully populated SAN256N with four TSW modules will supply 256 ports non-blocking connectivity with full availability at 10 Gbps.

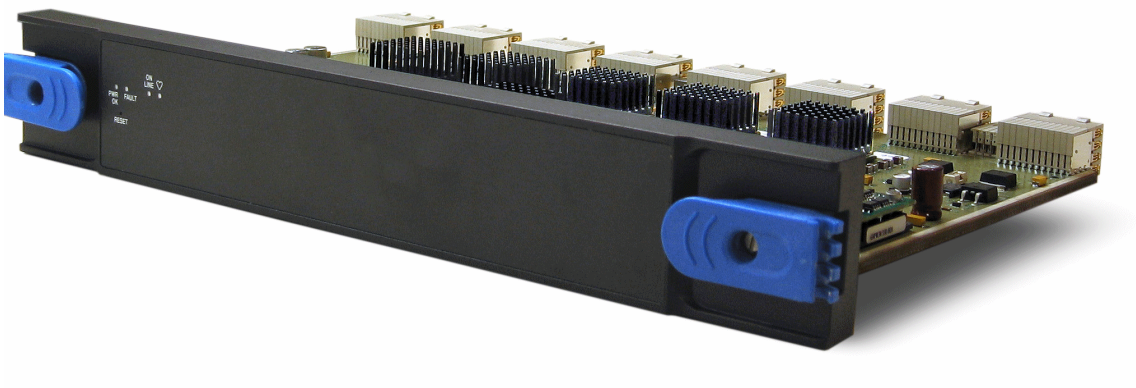


Figure 9-11 CNT SAN256N TSW Switch Module

SAN256N Power Supply

Figure 9-12, “CNT SAN256N Power Supply Assembly” on page 360 shows an example of one of the dual power supply assemblies located in the rear of the chassis. Each Ultranet Multi-service Director is shipped with two power supplies for full redundancy (DC load sharing) and availability. Each power supply is fully integrated, removable, and hot-swappable. Integrated fans within the power supply cool the assembly separately from the system fans.

AC input power is supplied to each power supply via a separate line cor.d. Each supply also has a DC Power On/Off switch for itself. The power supply assembly provides visual display, via two front panel LED indicators, indicating that VAC input voltage is within specification and that the DC voltage output is within specification.

Figure 9-12 CNT SAN256N Power Supply Assembly

Ultranet Multi-service Director fan assemblies.

The Ultranet Multi-service Director contains multiple fan assemblies for full redundancy and availability. Figure 9-13 on page 361 shows the front and rear views of the upper fan assembly.

Three upper fan module assemblies, located in the upper rear of the chassis, provide the cooling capability for the SAN256N and airflow around the I/O modules or other installed modules in the sixteen slots. The fans cool by pushing air out of the chassis. The upper portion of the chassis is cooled by a push-pull airflow drawn from the lower air inlets, located at the front and sides of the chassis, and exhausted at the top of the chassis rear.

Any of the three fan modules can be replaced with the same fan module. In the event of a single fan failure, the remaining fans increase speed to a higher operating rpm to compensate for the single fan failure. A Fan Status indicator lights green to indicate proper operation of the fan module, red to indicate an under-speed condition, and orange to indicate a marginal fan under-speed fault. The LED remains off when indicating a circuit fault. Each upper fan module consists of two fans mounted vertically in a pull-out module.

Figure 9-13 CNT SAN256N Upper fan assembly

Figure 9-14 on page 362 shows the side fan assembly. Two fan module assemblies are located in the right-hand side, middle rear of the chassis). They are used to pull air into the SAN256N chassis to aid in cooling the chassis components (in particular, the switch modules). The middle portion of the chassis is cooled by a push-pull airflow drawn from the middle side air inlets, located at the rear portion of the chassis, and exhausted at the right-hand side of the rear chassis.

Any of the two fan modules can be replaced with the same type fan module. If there is a single fan failure, the remaining fans increase speed to a higher operating rpm to compensate for the outage. A Fan Status indicator lights green to indicate proper operation of the fan module, red to indicate an under-speed condition, and orange to indicate a marginal fan under-speed fault. The LED remains off when indicating a circuit fault.

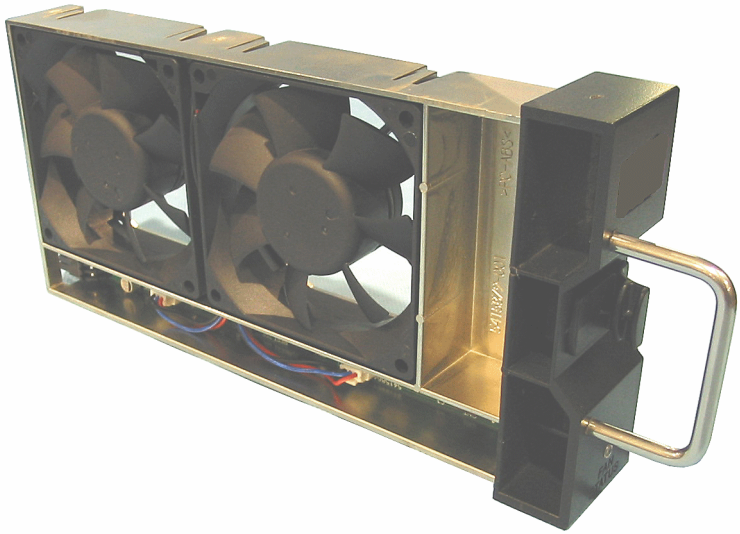


Figure 9-14 CNT SAN256N side fan assembly.

Each side fan module consists of two fans mounted horizontally in a pull-out module.

SAN256N air baffles

There are two types of air baffles for filling empty slots. Air baffles are installed in slots not containing I/O or switch modules to maintain the proper air flow required in the SAN256N chassis. The I/O blade air baffles have the same physical dimensions as an actual I/O module, only with a blank front panel, The switch module air baffles have the same physical dimensions as an actual switch module and a blank front panel. The baffles are installed into the chassis in the same manner as I/O blades and switch modules.

Figure 9-15 on page 363 shows an example of the air baffle.

Figure 9-15 CNT SAN256N Air Baffle

9.3.2 Ultranet Multi-service director supported attachments

The CNT SAN256N director provides the following types of attachments:

- FICON
- FCP
- FCIP
- iSCSI
- ETHERNET
- SONET
- FICON Cascading
- ISL

In addition to the usual Fibre Channel and FICON attachments, the CNT SAN256N will have integrated MAN/WAN interfaces for Ethernet (IP) and SONET networking. This simplifies infrastructure management. The integrated MAN/WAN interfaces will be available in a blade design to plug into one of the sixteen front slots.

9.3.3 SAN256N Multiple Services

The CNT Ultranet Multi-service Director offers the following expanded services not usually found in a SAN director:

Management

The SAN256N director offers a full-featured in-band element manager which produces statistics, diagnostics and logs for ease of fabric management across multiple domains and fabrics. The manager is an updated version of the inVSN Enterprise Manager Server.

The SAN256N director offers third party management support by supporting the following specifications:

- GS-3-ANSI standard for switch services
- SMI-S-SNIA management specification
- FC-SWAPI-ANSI standard épée support for third-party management

Examples of third party vendors currently supported are:

- IBM OS390 System Automation
- Tivoli Storage Network Manager
- Also solutions from - EMC, Hitachi, HP, Veritas, Creekpath, AppIQ
- inVSN Storage Network Manager – end-to-end visualization and pro-active monitoring of your enterprise storage network

Security

- FC-SP-ANSI standard for Fibre Channel security protocols that define authorization, authentication, data integrity and confidentiality
- Fabric level hard zoning
- Hardware assisted port and switch binding
- Logical partitioning

Storage

- iSCSI gateway to Fibre Channel storage
- Support for storage virtualization, replication and management

Extension Services

- SAN extension across local, metro, or wide area networks with data compression and performance acceleration techniques for disk and tape

SAN256N high availability

Figure 9-16 on page 365 illustrates SAN256N's Ultrascale architecture designed for continuous availability. The software architecture is distributed across the redundant control, port, and switching modules. Each module's firmware can be upgraded non-disruptively. There are multiple redundant switch paths. There is active-active failover between switch modules, control cards, power supplies, and fans.

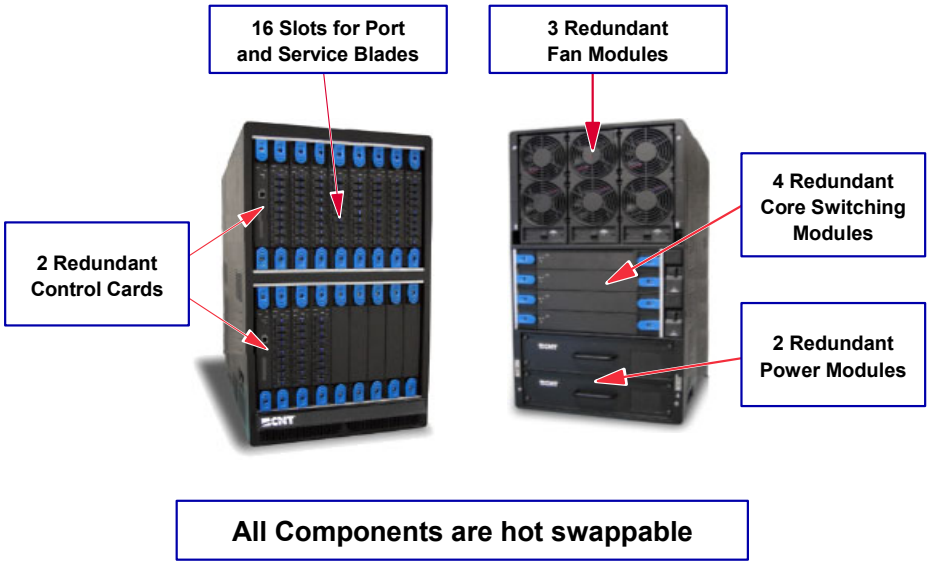


Figure 9-16 SAN256N design for high availability

9.4 FC/9000 Director product description

The FC/9000 director design is based upon server requirements for a large number of ports in a single director (256 ports in a single footprint) as well as the ability to upgrade to larger configurations in the future. The director is also designed to provide the high levels of availability, performance, and integrity required by today's most demanding data centers.

The director features N+1 redundancy at all critical points of design, automatic internal failover, extensive hot-swapping, non-disruptive firmware updates, and automatic fault detection and isolation. In addition, call-home and pager capabilities can automatically alert support and maintenance personnel to accelerate problem resolution.

The CNT director is available in three models providing 64, 128, or 256 ports. These ports can be a mix of either FICON or Fibre Channel (with 1 Gb/s and 2 Gb/s ports being supported) and they are expandable in 8-port increments. The ports are hot-pluggable and allow non-disruptive upgrades.

The director includes redundant power supplies and N+1 cooling fans together with redundant active components that support automatic failover. All field-replaceable units support hot-swapping, and concurrent firmware updates

are also supported. Call home and pager capabilities can be configured to automatically alert support and maintenance support teams.

A High-Availability Option provides redundant internal pathing, redundant control modules and power modules for all active electronic components.

The director provides a high performance, non-blocking architecture that provides 200 MB/s throughput to all ports in the director.

Previous models of the CNT director which are 1 Gb/s and 128-port capable are field upgradeable to the newer 256-port/2 Gb/s ready director model.

In the following sections we present an overview of the 2 Gb/s 64 port, 128 port and 256 port CNT directors.

9.4.1 FC/9000-64 Fibre Channel Director

The FC/9000-64 Fibre Channel Director (IBM 2042-001) forms the basic building block for this legacy product range. The base system comes with 24 ports and can be upgraded in 8-port increments to a maximum of 64 ports within the chassis. The CNT director supports both GBIC and SFP compliant ports.

A High-Availability Option (f/c 5020) exists which will provide a configuration with redundant power supply assemblies, redundant system modules and redundant control modules. These modules are discussed in more detail in “FC/9000 components” on page 372.

By ordering the Model 001 Interconnect Kit (f/c 4129), two 64-port CNT directors can be upgraded to form the functional equivalent of the 128-port CNT director.

The 2042-001 will need to be installed inside the 2042-C40 Cabinet. This model will provide an CNTFC/9000 40U Cabinet which can hold up to two 2042-001 Directors.

These are the physical dimensions of the IBM 2042-001 director:

- Width: 43.18 cm (17.00 in)
- Depth: 73.03 cm (22.00 in)
- Height: 84.46 cm (33.25 in)
- Weight: 81.60 kg (180.00 lbs)

Figure 9-17 shows a picture of the 64 port director.

Figure 9-17 CNT FC/9000-64 Fibre Channel Director

9.4.2 FC/9000-128 Fibre Channel Director

The IBM 2042-128 is the FC/9000-128 Port Director which includes a fully integrated FC/9000 40U Cabinet containing the director. The director is configured with a minimum 48 ports and a maximum of 128 ports with additional ports being added in 8-port increments.

A High-Availability Option (f/c 5022) exists which will provide a configuration with redundant power supply assemblies, redundant system modules, and redundant control modules. These modules are discussed in more detail in “FC/9000 components” on page 372.

The inVSN Management Server (f/c 7600) with up to 16 Management Clients can also be ordered with the IBM 2042-128 and is external to the cabinet. This is the management tool that is used to configure and administer the CNT director.

By ordering the Model 128 Interconnect Kit (f/c 4256), two 128-port CNT directors with adjacent FC/9000 Cabinets can be upgraded to form the functional equivalent of the 256-port director.

These are the physical dimensions of the IBM 2042-128 director (with 40U cabinet):

- Width: 61.00 cm (24.00 in)
- Depth: 99.10 cm (39.00 in)
- Height: 199.40 cm (78.50 in)
- Weight: 272.20 kg (600.00 lbs)

Figure 9-18 shows a picture of the 128 port director.

Figure 9-18 CNT FC/9000-128 Fibre Channel Director

9.4.3 FC/9000-256 Fibre Channel Director

The IBM 2042-256 is the FC/9000-256 Port Director which includes dual fully integrated FC/9000 40U Cabinets containing the director. The director is configured with a minimum 48 ports and a maximum of 256 ports with additional ports being added in 8-port increments.

A High-Availability Option (f/c 5224) exists which will provide a configuration with redundant power supply assemblies, redundant system modules and redundant control modules. These modules are discussed in more detail below in “FC/9000 components” on page 372.

The inVSN Enterprise Management Server (f/c 7600) with up to 16 Management Clients can also be ordered with the IBM 2042-256 and is external to the cabinet. This is the management tool that is used to configure and administer the CNT director.

These are the physical dimensions of the IBM 2042-256 director (with two 40U cabinets):

- Width: 122.00 cm (48.00 in)
- Depth: 99.10 cm (39.00 in)
- Height: 199.4 cm (78.05 in)
- Weight: 544.30 kg (1200.00 lbs)

Figure 9-19 shows a picture of the 256-port CNT FC/9000 director.

Figure 9-19 CNT FC/9000-256 Fibre Channel Director

9.4.4 FC/9000 director supported attachments

The CNT director provides the following types of attachments:

- FICON
- FCP
- FC-IP
- Private and Public Arbitrated Loop (including Public-Private translation)
- Cascaded directors
- Interoperability port for cascading of fabrics

9.4.5 FC/9000 supported port types

The CNT director supports a comprehensive range of port types to allow for a vast range of connection options. The supported port types include:

- F_Port (fabric)
- FL_Port (public loop)
- E_Port (for multiple CNT directors)
- TL_Port (private to public bridging)
- Fibre Channel connectivity (FICON)

FCAL and TL ports are only configurable on 1 Gb/s FIO boards on FC/9000 directors.

Previous versions of the FC/9000 director also supported a T_Port. This provided a proprietary method of connecting multiple directors. CNT now supports the E_Port standard.

All ports, with the exception of the TL_Port, are self discovering. CNT directors automatically sense the attributes of individual end-nodes, configuring themselves in any combination of fabric or switch-to-switch ports as needed. Manual adjustments are not necessary at the time of installation or as the fabric evolves, because the director adapts to change dynamically — while the SAN is up and running.

9.4.6 Availability

The CNT director provides high availability with components supporting automatic failover, automatic fault detection, and isolation, in addition to call home and pager support to enable rapid problem resolution. It also provides:

- Non-disruptive software and firmware upgrades
- Passive backplane
- Hot swapping for all field replaceable unit (FRU) components

The High-Availability Option (f/c 5020, 5022, or 5224, depending on the model of the director) provides:

- Redundant internal pathing
- Redundant power modules

- Redundant control

The High-Availability Option is highly recommended.

9.4.7 Scalable capacity CNT director high availability option

The High-Availability Option provides redundancy for all electronic and power modules — helping to enable continuous data access and high performance in the event of a single component failure. Together, these capabilities are designed to help provide uninterrupted full-bandwidth service without the loss of data access during periods of failure, repair, maintenance, and capacity upgrades.

Multiple directors can provide a scalable enterprise SAN backbone that supports consolidated storage management applications such as disk sharing, tape pooling, and enterprise-wide data sharing, as shown in Figure 9-2 on page 347.

To meet the demands of a growing enterprise, the CNT director provides one of the most flexible capacity solutions in the market today. With an entry level of 24 ports the Fibre Channel switch can scale, in increments of 8 ports, up to 256 ports. This is illustrated in Figure 9-20.

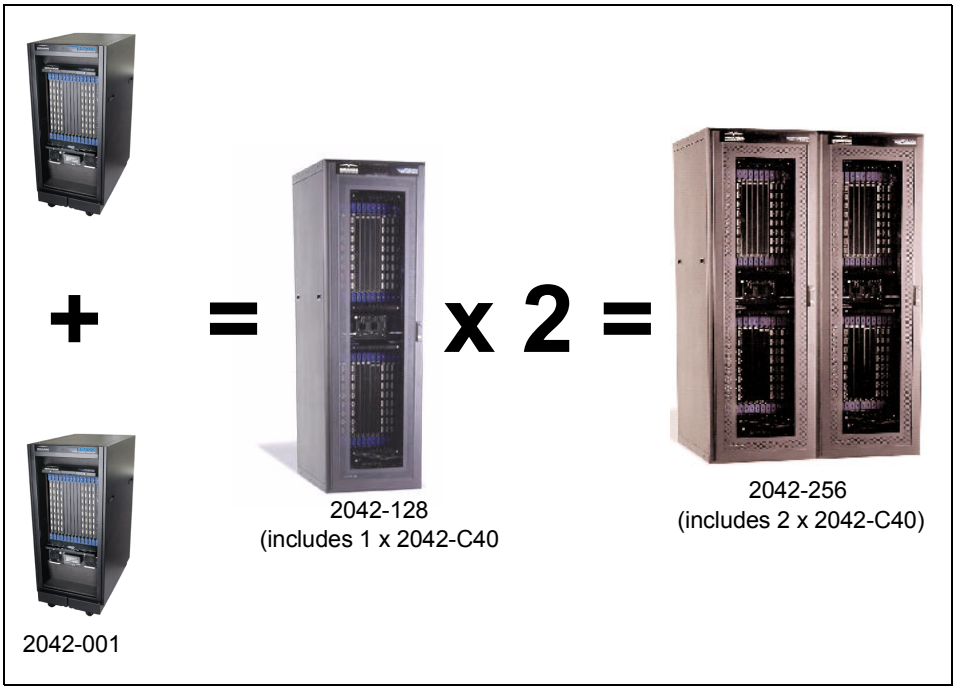


Figure 9-20 FC/9000 director scalability

The expandability of the FC/9000 director is possible due to the Extensible Core Architecture (XCA). The XCA design provides a combination of architecture, hardware and embedded operating system designed from the start to scale to 256 ports. This design provides full any-to-any non-blocking performance between all ports with no degradation, no extra hops due to ISLs and an industry-leading latency of between 0.6 to 2.7 microseconds.

inVSN Enterprise Manager software provides the interface into one or multiple Fibre Channel switches and can support up to 16 inVSN clients for remote management.

9.5 FC/9000 components

In the topics that follow, we describe the main components of the CNT director, along with showing the physical location of the components.

The FC/9000 director uses a blade-based design which can be expanded or upgraded by the addition of new blades or modules. Extra control blades (or modules) can be added as part of the High-Availability Option.

The current version of the FC/9000 director are 2 Gb/s based and capable of 256 ports.

The older models are 1 Gb/s based and 128-port capable. These older models can be upgraded to the newer models by replacing the control modules.

9.5.1 FC/9000 support the following components:

The 2 Gb/s and 256 port capable directors support the following modules:

- **FC/9000 eXtended Fibre I/O Module - 2 Gb (XFIO2):** Used to add additional SFP ports to the director
- **FC/9000 Base I/O module (FIO):** Used to add additional GBIC ports to the director
- **FC/9000 Switching (FWI) module:** Provides the cross-connection architecture of the director
- **FC/9000 Control Module Enhanced (FCME):** Provides the management and communication interface to the director

Filler cards are also used to ensure the correct airflow within the chassis. These are listed below:

- **XFIO2/FIO Slot Filler Module:** Filler module for unused (empty) XFIO2/FIO slots to ensure proper air flow

- **Non-High Availability Filler Module:** Filler module for unused (empty) HA slots to ensure proper air flow
- **Non-High Availability Filler Module/128 port:** Filler module for unused (empty) HA slots to ensure proper air flow for the Model 128

FC/9000 eXtended Fibre I/O Module: 2 Gb/s (XFIO2)

The CNT director uses XFIO2 cards (f/c 5210) to provide the physical connection between the CNT director and the external devices being connected. Each XFIO2 module has 8 SFP ports that support CNT-certified SFP transceivers including multi-mode fiber optics and single-mode fiber optics. This provides Fibre Channel and FICON connectivity to the attached devices.

Each port will operate at 2 Gb/s and will auto-negotiate to either 1 Gb/s or 2 Gb/s. All ports are self-configuring and have 128 buffer credits per port providing support for a distance of up to 80 km distance using extended distance longwave SFPs.

The XFIO2 card also has 2 mirror ports, L-A MTX for transmit and L-A MRX for receive. These are for diagnostic purposes.

The CNT director supports a minimum of three and a maximum of eight XFIO2 modules. When one or more of the XFIO2 blades is not required, an XFIO2 filler module (f/c 5920) must be installed to ensure the correct airflow within the chassis.

The XFIO2 card does not support arbitrated loop connections. To connect to FC_AL devices the FIO card (1 Gb/s and GBIC based card) is required. Refer to “FC/9000 Fibre I/O Module: 1 Gb/s (FIO)” on page 380.

Figure 9-21 shows the various port types available across the director XFIO2 module.

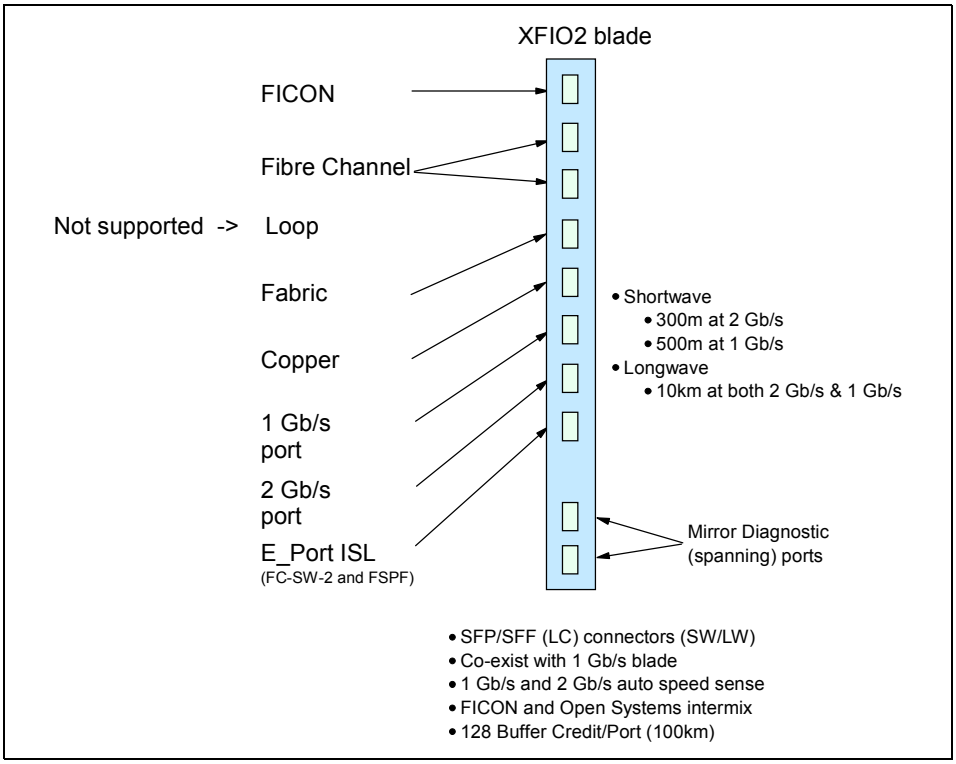


Figure 9-21 CNT director XFIO2 module

The XFIO2 module has two redundant backplane paths through to the redundant FWI module. The FWI is described in “The XFIO2 cards can be installed into chassis slots 0-3 and 9-12.” on page 375. The base XFIO2 module logic manages the synchronized switch over to the spare module. Each backplane base XFIO2 port has a redundant I/O which is routed to a cross point chip, which has two input ports and two output ports that are routed to a spare FWI module. The corresponding chip is then configured to bypass a failed FWI module and switches the connectivity through to the spare FWI module.

Shown in Figure 9-22 is a picture of the XFIO2 module.

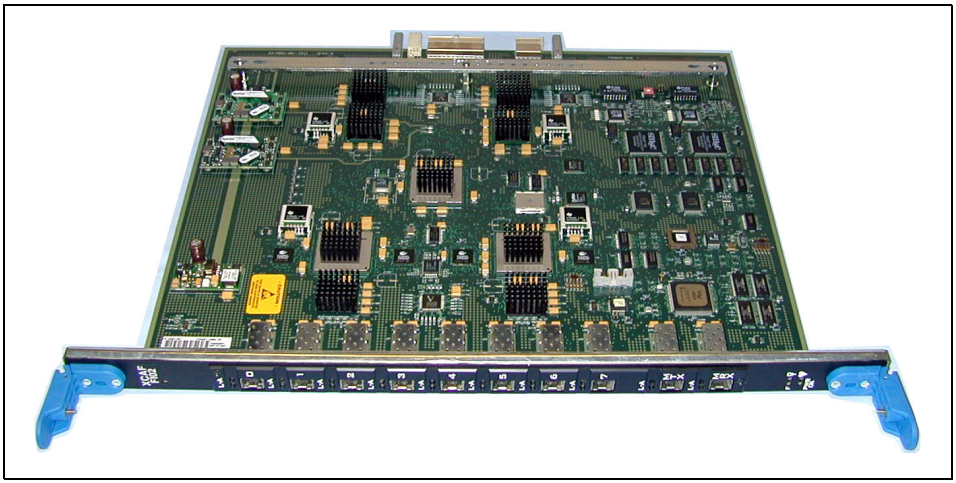


Figure 9-22 8-port XFIO2 module

The XFIO2 cards can be installed into chassis slots 0-3 and 9-12.

The FWI module provides the physical and logical connectivity between the XFIO2 and FIO modules which are installed in the chassis. The FWI module provides the middle or cross-connect section of the director architecture.

The FWI modules are hot-swappable.

A picture of the FWI module is shown in Figure 9-23.

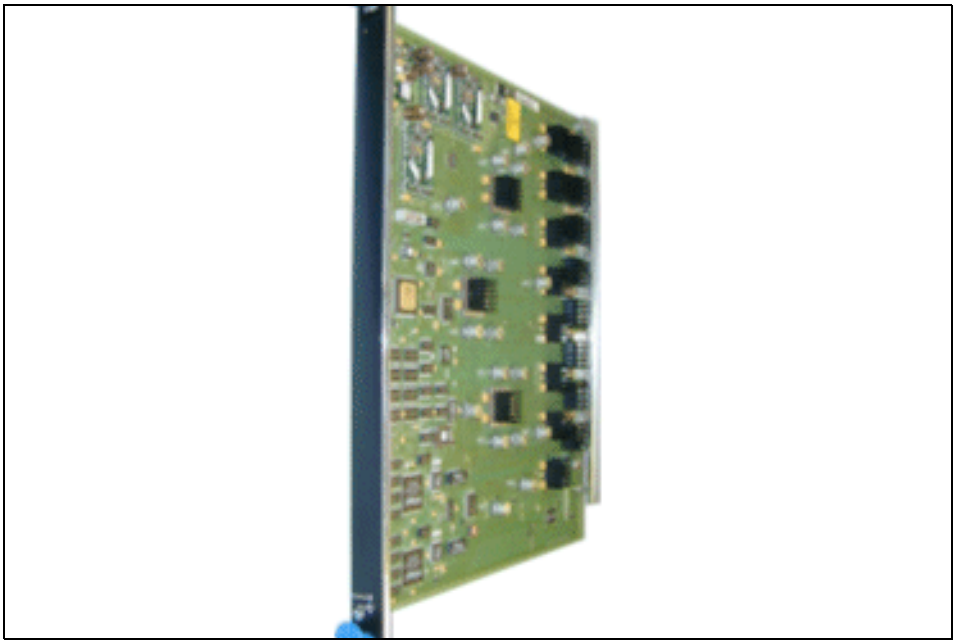


Figure 9-23 FWI module

When the High-Availability Option (f/c 5020, 5022 or 5224 depending on the model of the director) is specified a hot-spare FWI is provided. The redundant FWI will act as the hot-spare for all FWIs within the director. FWI modules are installed in chassis slots 5 and 7 in Port Block 1, 2, 3 and 4 with the redundant FWI being installed in chassis slot 6 in Port Block 1.

FC/9000 Control Module Enhanced (FCME)

The FCME provides the common command and control interface for the CNT director. The FCME monitors the node and acts as a proxy for all external communications going to other modules in the system and all communications from the other modules going to the inVSN Enterprise Manager.

The High-Availability Option (f/c 5020, 5022 or 5224) provides a redundant FCME module that acts as the hot standby. The FCME is a critical module within the system.

A picture of the FCME is shown in Figure 9-24.

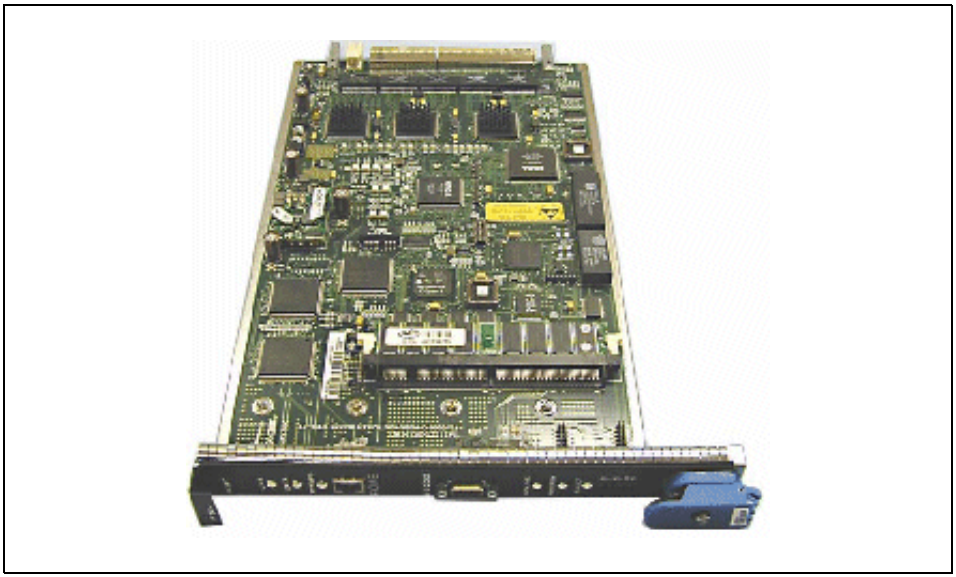


Figure 9-24 FCME card

The FCME modules are installed in chassis slots 13A and 13B.

Power supply assembly

There are two hot-swappable power supply assemblies that provide the CNT director with full redundancy. The second power supply is provided when the High-Availability Option (f/c 5020, 5022 or 5224 depending on the model of the director) is specified.

Each power supply has an LED display indicating that the power supply and DC voltage are functioning correctly.

The electrical specifications of the power supply assembly are as follows:

- Input voltage is 220 V AC nominal; V AC input range is from 200 V AC to 240 V AC, +/- 10%.
- Input frequency is 50/60 Hz nominal; frequency range is 47 Hz to 63 Hz.
- Output is 48 V DC nominal, +/- 5%.

Each power supply is removable and hot-pluggable and contains its own integrated fan which is separate to the chassis fan assembly.

Figure 9-25 shows one of the two power supply assemblies for the director.

Figure 9-25 Director - power supply assembly

Fan assembly

There are four fan modules which either push or draw air to achieve the CNT director cooling. All fans can be replaced independently. In the event of a single fan failure the remaining fans will automatically adjust their speed to compensate for the failing component.

Each fan module is powered by 48 V DC and provides up to 300 cfm per fan assembly. The fan modules are interchangeable.

Figure 9-26 shows one of the four fan assemblies for the director.

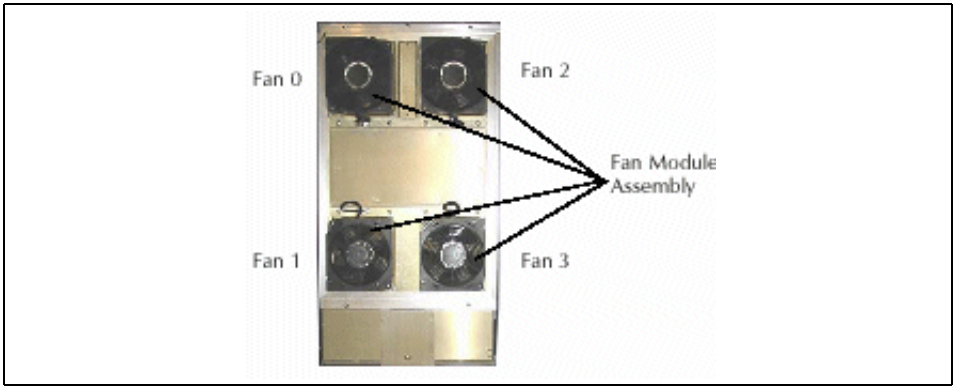


Figure 9-26 CNT director - fan assembly

Backplane module

The 2 Gb/s based CNT director has a passive backplane that provides the connectivity for all system modules. This backplane has the capability to be extended from a 64 port model to a 256-port model. When upgrading a 64-port director to the 256-port director, the backplanes are connected via a special interconnect cable which is supplied as part of the upgrade kit.

The backplane module is a passive component and therefore has no moving components. If physically damaged, it is not a FRU, so the entire CNT director would need to be replaced. The physical replacement of a chassis takes 1 to 2 hours, depending on the number of ports.

Crossbar and field programmable gate arrays

The 2 Gb/s based CNT directors have moved away from ASIC's and now use a crossbar to handle the switching requirements and field programmable gate arrays (FPGA) to manage the protocols for the director.

9.5.2 CNT director: 1 Gb/s and 128 port capable

The 1 Gb/s ready and 128 port capable CNT directors support the following modules:

- **FC/9000 Base I/O module (FIO):** Used to add additional GBIC ports to the director
- **FC/9000 Switching module (FSW):** Provides the cross-connection architecture of the director
- **FC/9000 Control Module (FCM):** Provides the management interface to the director

Filler cards are also used to ensure the correct airflow within the chassis:

- **FIO Slot Filler Module:** Filler module for the unused (empty) FIO slots to ensure proper air flow
- **Non-High Availability Filler Module:** Filler module for the unused (empty) HA slots to ensure proper air flow
- **Non-High Availability Filler Module/128 port:** Filler module for the unused (empty) HA slots to ensure proper air flow for the Model 128

The blade-based architecture of the CNT director is illustrated in Figure 9-27.

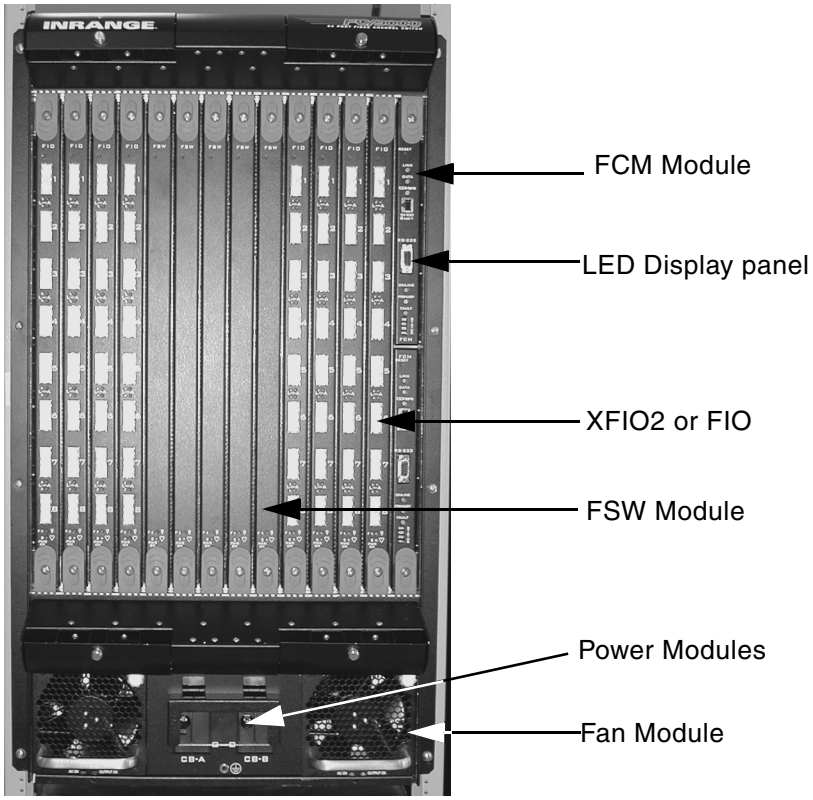


Figure 9-27 CNT FC/9000 director - module locations

FC/9000 Fibre I/O Module: 1 Gb/s (FIO)

The CNT director used the older FIO cards (f/c 5010) to provide the 1Gb/s physical connection between the CNT director and the external devices being connected. Each FIO module has 8 ports that terminate at CNT certified GBIC compliant devices including copper, multi-mode fiber optics and single-mode fiber optics. This provides Fibre Channel and FICON connectivity to the attached devices.

The CNT director has a maximum port count of 64, and currently the minimum IBM configuration requires 3 FIO blades for a total of 24 ports. When one or more of the FIO blades is not required, an FIO filler module (f/c 5910) must be installed.

Shortwave (multi-mode), color coded beige or black exposed surface GBICs, and longwave (single-mode) color coded blue exposed surface GBICs are supported. Each GBIC consists of a transmitter and receiver optical subassembly. Both the shortwave and longwave discrete laser diodes are classified as Class 3B laser products. Supported interface converters (for example FICON and GBICs) can be installed in the FIOs in any combination.

All ports are self-configuring and have up to 64 buffer credits set allowing, if required, all 64 ports to be used for longwave transmission.

The FIO module has two redundant backplane paths through the backplane to the redundant Fibre Channel Switch Module (FSW). The FSW is described in “Fibre Channel Switch Module” on page 382.

The base FIO module logic manages the synchronized switch-over to the spare module. Each backplane base FIO port has a redundant I/O, routed to a cross point chip, which has two input ports and two output ports that are routed to a spare FSW module. The corresponding chip is then configured to bypass a failed FSW module and switches the connectivity through the spare FSW module.

There are three types of memory used on the FIO modules:

- Non-volatile, read-only, used to store hardware configuration, Boot code and Maintenance interface code.
- Non-volatile, block re-writable memory, used to store firmware operation code or user configurable port settings
- Volatile high speed memory, which contains a full copy of firmware, operation code, operating parameters, and data packets for routing or special handling.

In Figure 9-28 we show the FIO module.

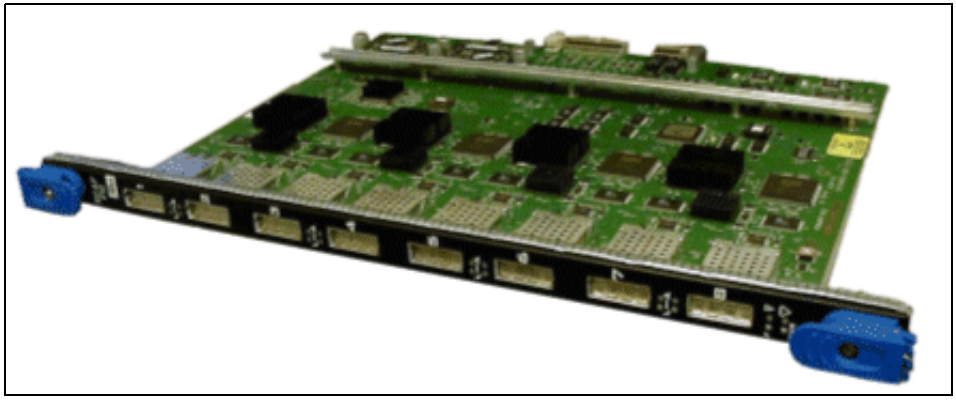


Figure 9-28 8 port FIO module

The FIO cards can be installed in chassis slots 0-3 and 9-12.

Shown in Figure 9-29 are the port types available on a FIO module. The FIO module supports 1 Gb/s and is GBIC-based.

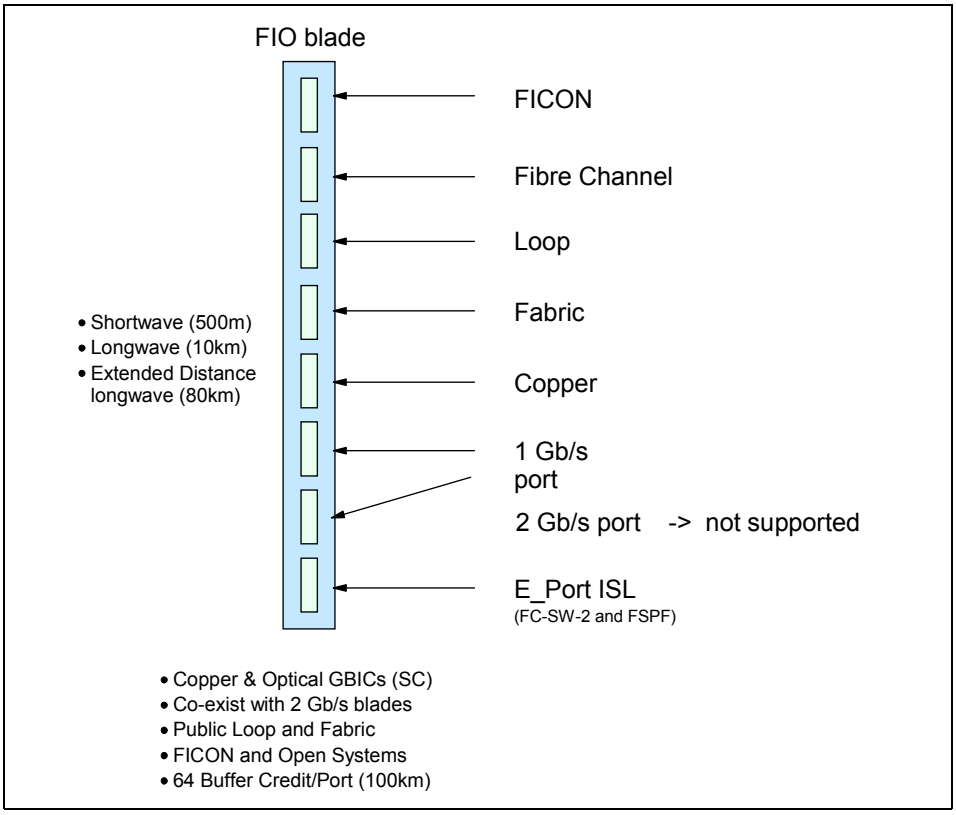


Figure 9-29 CNT director FIO module

Fibre Channel Switch Module

The Fibre Channel Switch Module (FSW) provides the middle or cross connection architecture of the director. There are up to four active and one hot spare FSWs that provide the physical and logical links between the FIO modules. The hot spare FSW is only provided when the High-Availability Option (f/c 5020, 5022 or 5224) is selected. The hot-spare FSW will act as a spare for both chassis for the 128-port model.

FSW modules are hot-swappable modules and are interchangeable within the director.

Shown in Figure 9-30 is a picture of the FSW module.

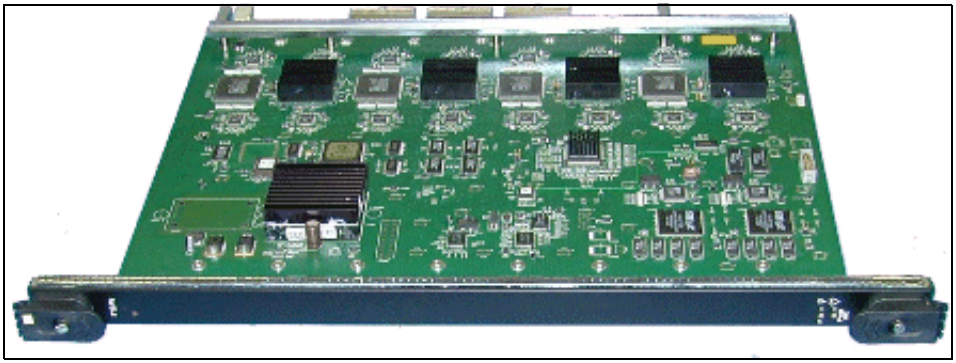


Figure 9-30 FSW module

The major differences between an FIO blade and FSW module are that FIOs have external ports and have serializer/deserializer components, while the FSW modules do not have either of these functions.

Located at the bottom of each FIO blade and FSW module, there are four LED indicators that display the status of the component:

- Over temperature
- Heart beat
- PWR OK
- F LEDS

For detailed information of the color and LED sequences, refer to the *inVSN FC/9000 Fibre Channel Director - Installation Manual*, 9115100-01 and the *inVSN FC/9000 Fibre Channel Director - Maintenance Manual,* 9114572-307-00.

The FSW modules are installed into chassis slots 4 though 8 with one FSW acting as the hot-spare.

Note: The FSW modules are replaced by the FWI modules for all new machines. Two FWI modules provide the same function as four FSW cards.

Fibre Channel Control Module

The Fibre Channel Control Module (FCM) provides the command and control interface for the system. It enables the control management software, inVSN, to configure, modify, and test the CNT director.

The High-Availability Option (f/c 5020, 5022, or 5224, depending on the model of the director) provides a redundant FCM module that acts as the hot standby. The FCM is a critical module within the system.

In addition to the three types of memory in the FSW modules, the FCM module also has Non-Volatile Random Access Memory (NVRAM), and this is used to store persistent system configuration and status information. Error log information is also stored in the NVRAM.

For Customer Engineer access to the diagnostic log and maintenance panels, there is an RS232 port.

For high availability, it is possible to have a redundant hot standby FCM module. The FCM hot spare is only provided when the High-Availability Option (f/c 5020) is selected. The FCM modules are hot swappable units.

Shown in Figure 9-31 is a picture of the FCM.

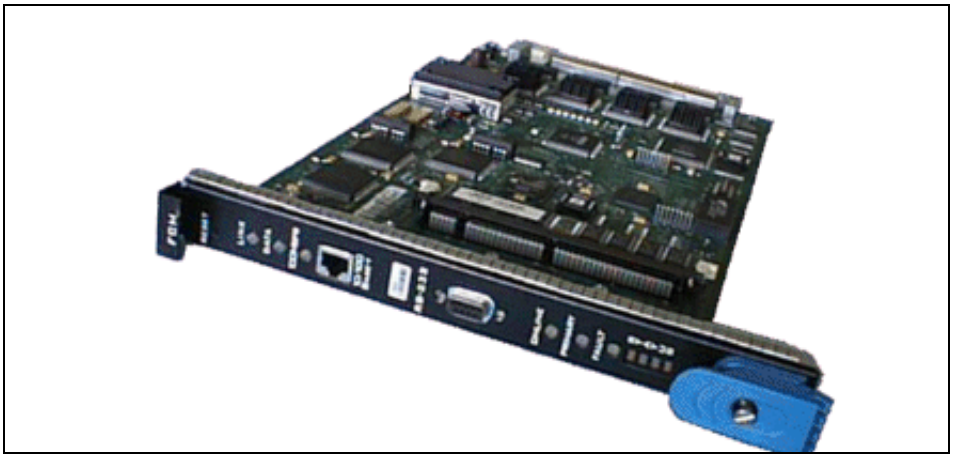


Figure 9-31 FCM module

The FCM is installed in chassis slots 13A and 13B (when configured with the High Availability Option).

Power and fan assembly

There are two hot swappable power supply assemblies that provide the CNT director with full redundancy. The second power supply is provided when the High-Availability Option (f/c 5020, 5022 or 5224 depending on the model of the director) is specified. Each power supply has an LED display indicating the power supply and DC voltage are functioning correctly.

The electrical specifications of the power supply assembly are:

- Input voltage is 220 V AC nominal; V AC input range is from 180 V AC to 264 V AC.
- Input frequency is 50/60 Hz nominal; frequency range is 47 Hz to 63 Hz.
- Output is 48 V DC nominal, plus/minus 5%.

There are four fan modules which either push or draw air to achieve the CNT director cooling. All fans can be replaced independently. In the event of a single fan failure the remaining fans will automatically adjust their speed to compensate for the failing component.

Backplane module

The 1 Gb/s based CNT director has a passive backplane that provides the connectivity for all system modules. This backplane has the capability to be extended from a 64 port model to a 128 port model. When upgrading a 64 port director to the 128 port director connectors are supplied to connect the backplanes of both 64 port directors.

The backplane module is a passive component and therefore has no moving components. If physically damaged, it is not a FRU, so the entire CNT director would need to be replaced. The physical replacement of a chassis takes 1 to 2 hours, depending on the number of ports.

ASIC

The 1 Gb/s based CNT FC/9000 directors are capable of operating dynamically in multiple modes (F, FL, TL and E_Port) or a mixture of modes. There is no requirement to swap or reconfigure ASICs if you need to alter the use of ports in your SAN fabric.

9.5.3 Upgrade Path/Extensible Core Architecture

Upgrade Path/Extensible Core Architecture (XCA) is the architecture which allows a single chassis to expand. Upgrades within a given chassis are accomplished by insertion of port cards (each supporting eight Fibre Channel ports). These port cards can be inserted while the director is operational, without causing any disruption to connectivity of ports already in operation.

These port cards are “off the shelf” items with published list prices and availability. The upgrade from 64 port to 128 port or to 256 port systems requires additional inter-chassis connections as well as upgrades to control software, in addition to the extra chassis. These are orderable via the following feature codes:

- F/c 4129 - interconnect kit for 128-port capability
- F/c 4256 - interconnect kit for 256-port capability

The XCA architecture allows the director to be viewed as a single fabric eliminating the requirement to have any principal and subordinate relationships within the director. The director represents the fabric with a single domain ID for the whole 256-port fabric.

9.5.4 Cabinet

All CNT directors must be configured within an associated cabinet, the IBM 2042-C40. A single 40U cabinet can contain up to two CNT directors, and the inVSN Enterprise Management Server. When ordering a cabinet, it is important to specify if it will be field or plant installed. To provide for additional physical security, the cabinet has a locking door. It is good practice to keep the cabinet locked and the key only available to authorized staff.

The 2042-C40 is included with the IBM 2042-128 and 2042-256 models so it is not required to be ordered separately.

These are the physical dimensions of the IBM 2042 C40 cabinet:

- Width: 61.00 cm (24.00 in)
- Depth: 99.10 cm (39.00 in)
- Height: 199.40 cm (78.50 in) with stationary feet retracted

These are the front and rear recommended service access distances:

- Cabinet installation: 81.44 cm (36 inches)
- Stand-alone chassis installation: 81.44 cm (36 inches)

9.6 Features

In the topics that follow we highlight the features.

9.6.1 Ultranet Multi-Service Director High Availability

The SAN256N is designed for high availability with active-active resiliency in the switch core, hot swap capabilities, automatic internal failover of key components,

redundancy, concurrent firmware updates, and automatic fault detection and isolation.

The SAN256N TSM modules require a minimum of three blades for high availability. Two blades will supply full non-blocking performance for 256 ports at 4 Gbps. The third blade supplies the redundancy but is also used for load balancing. The switching loads are spread across all installed TSM modules. No failover or standby blades are needed, which is an improvement over the FC/9000 design (a standby core for failover). The fourth TSM module will be needed to support 10 Gbps for all 256 ports.

In addition, the SAN256N name server function is distributed and replicated across all TFIO modules, rather than a main copy only in the control modules. This supplies additional redundancy and less interactions per I/O within the director.

9.6.2 FC/9000 availability options

When ordering the CNT director, there are two IBM configuration options available that determine the high availability configuration. These are the High-Availability Option (f/c 5020) and the Non-High Availability option (f/c 5920).

High-Availability Option (f/c 5020)

A high availability configuration for the IBM 2042-001 consists of:

- 3 to 8 x FIO/XFIO2 modules
- N+1 x FWI modules
- 2 x FCME modules
- 2 x power supply assembly
- 4 x fan module assembles
- 1 x inVSN Enterprise Manager (Management System)

Non-high Availability Option (f/c 5920)

A non-high availability configuration for the IBM 2042-001 consists of:

- 3 to 8 x FIO/XFIO2 Modules
- 2 x FWI modules
- 1 x FCME module
- 1 x power supply assembly
- 4 x fan module assembles
- 1 x inVSN Enterprise Manager (Management System)

9.6.3 Logical Domains

The Ultranet Multi-services Director offers a new feature called Logical Domains. Logical Domains offer logical segmentation of a large SAN256N into one to four separate partitions. This allows physical consolidation of smaller switches while maintaining fabric isolation. One usage is to segment the switch by division/line of business for usage-based billing. Another use would be to segment FICON mainframe environments from open systems environments. A third usage would be to separate the test environments from the production environments. This is illustrated in Figure 9-39 on page 409.

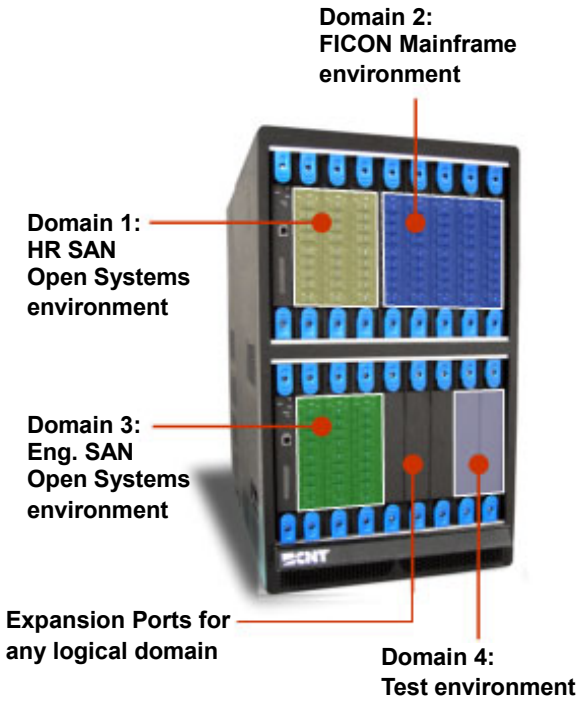


Figure 9-32 SAN256N Logical Domains

Redundant fabric services in each domain ensures continuous operation. The logical domains consist of one or more I/O modules. These must be adjacent at the 2004 GA version of SAN256N. This requirement is lifted in follow-on software versions of SAN256N.

The logical domains feature is unavailable on the FC/9000 director.

9.6.4 Trace and mirror ports

Both the FC/9000 and the SAN256N directors have two mirror ports built into the I/O modules, one for viewing the transmit stream of the port to be diagnosed called Mirror Transmit (MTx), and one for viewing the receive stream called Mirror Receive (MRx). Mirror ports transmit data only which allows connectivity to a Fibre Channel analyzer for hardware trace. These ports can be configured utilizing the Enterprise Manager to trace another port without pulling the fiber optic cable and inserting a Fibre Channel analyzer.

There are differences in the design of the mirror and trace features of the FC/9000 and the SAN256N. The FC/9000 mirroring is done at the internal bus level of the XFIO card. The FC/9000 mirror port enable can be found on the XFIO view of the Enterprise Manager. Please check with CNT and IBM support to find out which firmware version has this support enabled.

In the newer SAN256N design, the trace and mirror ports have been considerably improved. The Enterprise Manager allows the user to set triggers and filters to capture the desired amount of information flowing through the SAN256N. There are three types of trace available: hardware, firmware, and firmware setup.

The SAN256N hardware trace trigger combinations are *and, or,* and *the;* with directional control for the trigger. Once triggered, the hardware trace captures all frames.

The SAN256N firmware trace is similar to hardware trace, except that once the settings in the trigger have been met, only the frames meeting the given trigger condition defined by the filter are captured. Trigger combinations can only consist of *and* and *or*, and direction must be the same.

The SAN256N firmware startup trace depends on the base firmware capturing the current and previous states of the FCP state register changes along with other firmware trace information.

The output of the trace is stored in buffers within the SAN256N. The buffer file can be retrieved and stored on disk. The file is automatically converted to the Finisar GTX software format. This software can be downloaded from:

`http://www.finisar.com`

Once at the Finisar home page, type *GTX viewer* into the search box and Click on **Search.** Follow the instructions to download a free evaluation copy of the GTX viewer.

9.6.5 Management software inVSN Enterprise Manager

The Virtual Storage Network Enterprise Manager (inVSN) is browser based software used to manage and control one or more CNT directors. SAN256N Enterprise Manager Release 9.0 is able to concurrently manage both the SAN256N director and the FC/9000 director (requires FC/9000 R5.0 software upgrade). Some advanced features like Quality of Service and storage services are available only on the SAN256N.

The inVSN suite consists of two components:

- Server software
- Client software

The server communicates with the CNT director, while the inVSN client communicates with the inVSN server. All user interaction is performed by the client software.

inVSN management software capabilities include:

- Viewing and defining module and port configurations
- Viewing and defining zoning parameters
- Fabric management for multiple directors and switches
- Monitor alarms and system performance
- Invoking system diagnostics

inVSN Enterprise Manager (Management System) is designed to centralize and continuously monitor the management of multiple, distributed FC/9000 Fibre Channel Directors and SAN256Ns in an enterprise-wide Fibre Channel fabric backbone. The management system helps simplify the addition of managed FC/9000 and SAN256Ns as the enterprise Fibre Channel fabric grows. Utilizing a browser-based GUI, operators can use the inVSN Enterprise Manager locally or remotely from LAN-attached Manager Clients.

inVSN Enterprise Manager is designed with a flexible architecture consisting of the inVSN Enterprise Manager Console with inVSN Enterprise Manager Server software and multiple inVSN Enterprise Manager Clients. All management, monitoring, and diagnostic functions are accessed by running the inVSN Enterprise Manager Client on remote workstations. Multiple FC/9000 Directors, and the new SAN256N directors can be controlled from the same control node if they have Fibre Channel connectivity to a FC/9000 Director or a SAN256N that in turn has the Ethernet connection to the inVSN Management System.

inVSN Enterprise Manager Console

The inVSN Enterprise Manager Server software runs on the Manager Console. The inVSN Enterprise Manager Console, which is a desktop PC running Windows 2000 Professional, includes:

- A 17-inch monitor
- Internal and external modems for call home and remote diagnostics
- An 8-port Ethernet Hub with cable for a private LAN connection for multiple FC/9000 Fibre Channel Director Control Modules

The Manager Server:

- Supports continuous SAN256N and FC/9000 Fibre Channel Director monitoring, logging, and alerting.
- Centralizes log files, configuration databases, and firmware distribution.
- Supports centralized call home, pager alert, service, and support operations.

Multiple FC/9000 Control Modules and SAN256N modules can be managed from a single Manager Server, and up to 16 concurrent Manager Clients can access the Manager Server through the LAN.

If required, additional 8-port Ethernet hubs can be attached to the inVSN Management Server's private LAN to enable management of multiple FC/9000 Directors or SAN256Ns (via f/c 7603). Two 4.3 m Ethernet cables are provided with each FC/9000 Director to provide attachment to the Ethernet hub.

inVSN Enterprise Manager Server

The inVSN Enterprise Manager Server software is a Windows 2000-based GUI that helps simplify connectivity management with a common and consistent user interface. The Manager can run locally on the Manager Server or remotely on a LAN-attached inVSN Manager Client.

The inVSN Client software will run on Windows 2000 Professional, Windows 98, Windows 95, Windows XP, Linux, and Solaris servers.

The Enterprise Manager offers:

- Access security with passwords, administration of valid users, user rights, and remote access restrictions
- Enterprise management with a Fabric View showing the high-level status of all managed FC/9000 I Directors or SAN256Ns
- Scalability with the capability to add, remove, or modify managed FC/9000 Directors or SAN256Ns
- Traceability with audit logging and sessions status
- Enterprise Manager Server software

The Enterprise Manager Server Switch View is designed to provide a graphical representation of each FC/9000 and SAN256N. You can open individual windows through icons in the Fabric View. The Manager Switch View offers:

- A user-friendly interface with graphical displays of the FC/9000 Director, the SAN256N, and their components

- Centralized configuration management and control of all operating parameters
- Real-time interactive display of operational status and statistical performance information
- Automatic audit logging of all system events, configuration changes, and hardware upgrades
- Serviceability with access to all maintenance functions, such as diagnostics and firmware updates

Proactive notification of events with automatic call home alerts or pager alerts to a designated service center when service is required.ports giving the users more control over data flows.

9.6.6 inVSN Storage Network Manager

The inVSN Storage Network Manager will support the new SAN256N director and seamlessly integrate all other CNT products on the SAN/WAN for an end-to-end view of the application across the global infrastructure. The inVSN software will collect and analyze network information to aid in problem identification and isolation.

The inVSN Storage Network Manager supplies the following functions:

- Cross domain multi-protocol impact analysis
- Cross domain root cause analysis and fault correlation
- Cross domain topology visualization
- The summary of the data can be supplied to Third Party Frameworks

9.6.7 FICON

The following sections outline the FICON features.

FICON Manager Server (CUP)

Both the SAN256N and the FC/9000 support the FICON Control Unit Port (CUP) feature for in-band management of the directors for the z/OS or OS/390 operating systems.

The FICON Manager Server (f/c 7202) supports in-band management of the FC/9000 director by System Automation for z/OS or OS/390 on IBM 9672 Parallel Enterprise G5, G6, and zSeries 900 servers. This support provides a single point of control for managing connectivity in active I/O configurations. This type of dynamic in-band management is currently provided by IBM 9032 ESCON Directors. System Automation for OS/390 can now concurrently manage IBM 9032 ESCON Directors as well as FC/9000 Directors with FICON.

The SAN256N also supports in-band management with FC 7301 feature, with identical control and options as the FC/9000.

Logical Domains

With the SAN256N's new logical domain capability, no IOCP SysGen is required. Every Logical Domain has an Address 00. This makes it possible to migrate an existing physical director into a logical domain without changing the port numbering. Since the port numbers do not have to change, the existing IOCDS/IODF will not have to change.

It is possible to utilize zero cost ISLs between the logical domains within the full chassis and have no degradation in performance for cascading.

N-Port ID virtualization

The SAN256N supports 32 virtual IDs per port to provide N-port id virtualization for zSeries Linux LPARs. This feature allows for a single zSeries host port to appear to be multiple virtual N-ports. The zSeries has both normal logical partitions that include normal FICON channels and also Linux logical partitions that run FCP open systems channels. The single port can share both types of traffic with virtual N-ports for the zSeries host. This feature is expected to be utilized in the first half of 2005.

FICON cascading

FICON cascading is built on the FICON switched protocol, enabling multiple sites to be connected together to simplify FICON connectivity between sites. With FICON multi-switch cascading, there can be multiple path routes within the fabric to the FICON Control Unit. The fabric routing between the directors is handled via the Fibre Channel fabric routing, using industry standard FSPF routing (which is part of the FC-SW-2 standard).

CNT provides the required security mechanisms to ensure that E_Port ISLs between directors are safe and secure for FICON Cascade as well as for Open Systems traffic. Protocol InterMix (PIM) is the term used to describe FICON and Open Systems FCP traffic co-existing on the same Fibre Channel/FICON director at the same time. FICON Cascade utilizes underlying Fibre Channel standard technology, including E_Port for ISLs to interconnect directors as well as standard port type (N_Ports) to leverage commonality.

The FICON Cascading Activation feature (f/c 7203) is required for cascading support, or creating ISLs, between directors with FICON protocol applications. This feature activates firmware to manage FICON configurations with up to two directors between the FICON host server and the FICON storage device, as well as enabling further scalability of FICON fabrics beyond the port count limits of a

single director. For existing CNT directors, they must first be upgraded, if necessary, to the new version of hardware.

FICON cascading requires that the FC/9000 director be at the new 256-port and 2 Gb/s-ready hardware level. The director needs to be at firmware level 4.0.1.H and configured with the FICON cascading feature to support this functionality.

For the SAN256N, FICON cascading can support up to 10 Gb ISL, 100 Km distance, with policy based flow groups, and FCP/FICON intermix..

Note: FICON cascading is currently supported for dual director configurations with a single hop.

9.6.8 Intermix mode (FICON and FCP)

FC/9000

When running FICON and FCP Open Systems simultaneously (intermix mode), on the FC/9000 director, there are some guidelines which should be followed.

- Zoning with CS4.3.0 or above can be defined by:
  - Type 1: WWN (defined by WWPN)
  - Type 2: Domain and Port Id
  - Type 3: FC Address
- Enable Domain locking when using Type 2 and Type 3 zoning.
- Enterprise Manager will restrict zone type to Type 1 only with a CS4.3.0 fabric that has ISLs to a FC9000 director running CS4.1.3 or below.

SAN256N

When running FICON and FCP Open Systems simultaneously (Intermix Mode), on the SAN256N, these rules should be followed.

- Zoning with CS1.1.0 or above can be defined by:
  - Type 1: WWN (defined by wwpn)
  - Type 2: Domain and Port Id
  - Type 3: FC Address
- Do not use Type 2 - Domain and Port Id
- Enable Domain locking when using Type 3 zoning
- Enterprise Manager will restrict zone type to Type 1 only with a CS1.1.0 fabric that has ISLs to a FC9000 director running CS4.1.3 or below

Guidelines for both FC/9000 and SAN256N for intermix mode

- Use Port Groups to group similar ports together. Grouping ports will allow for smaller matrices of blocking and prohibiting assignments.
- WWN zoning should be used for controlling connectivity among FCP N-ports (“soft” enforced at the name server level). SAN256N will also offer hardware enforced WWN zoning.
- Ports defined for FCP traffic can be placed within a maximum of 256 unique zones totaling 3600 members.
- Prohibits/Allows should be used for controlling connectivity among FICON N-ports (“hard”, i.e. enforced at a traffic frame level).
- FICON ports should be placed within a single zone to provide name server separation of FICON ports from FCP ports. This ensures that FICON ports receive a RSCN for FICON port state changes.
- Port prohibits/allows should be limited to the ports defined in the FICON zone, and not used for ports in a FCP zone. Prohibits may be used for FCP N-ports, however some FCP devices/HBAs have been found to have limitations with handling a prohibit RSCN containing multiple entries. When enabling or disabling prohibits, a RSCN will be sent to all affected prohibited port(s) even if a port is in a different zone
- Use Port prohibits on all FICON ports to all FCP ports. Set prohibits on columns corresponding to FICON ports using the Enterprise Manager. Then the intersection with FICON port rows may be SET to allow FICON to FICON connections.
- Port prohibits and allows may also be set via the FICON mainframe using MVS and CUP.
- Do not set port prohibits on ISL ports.

Flow Group Sets

The Flow Group Set process provides the ability to force certain data flows to utilize a specific Inter Switch Link (ISL) and whether or not other traffic can utilize the same ISL. Consult the *Site Planning Guide for the FC/9000* or the *SAN256N* before utilizing this feature of Enterprise Manager.

9.6.9 Getting started with the FC/9000

The CNT director is delivered with the current supported level of firmware and a default TCP/IP setting of 10.1.1.51 and 10.1.1.52, and a Subnet mask setting of 255.255.255.0. There is also a default Chassis ID, Switch ID, and fabric ID. The Chassis ID is used in each domain ID defined.

It is the responsibility of an IBM Customer Engineer (CE) to reset all default addresses to reflect the environment that it is being installed into.

For added security, the TCP/IP address can only be set or reset by the CE, using an RS232 connection and entering the CE userid and password. The new TCP/IP address will be displayed in an LED panel that can be located on the FCM module.

The PC that will be used for the inVSN server needs to be connected to the FCM module using the Ethernet ports.

In Figure 9-33 we show a typical system communication setup.

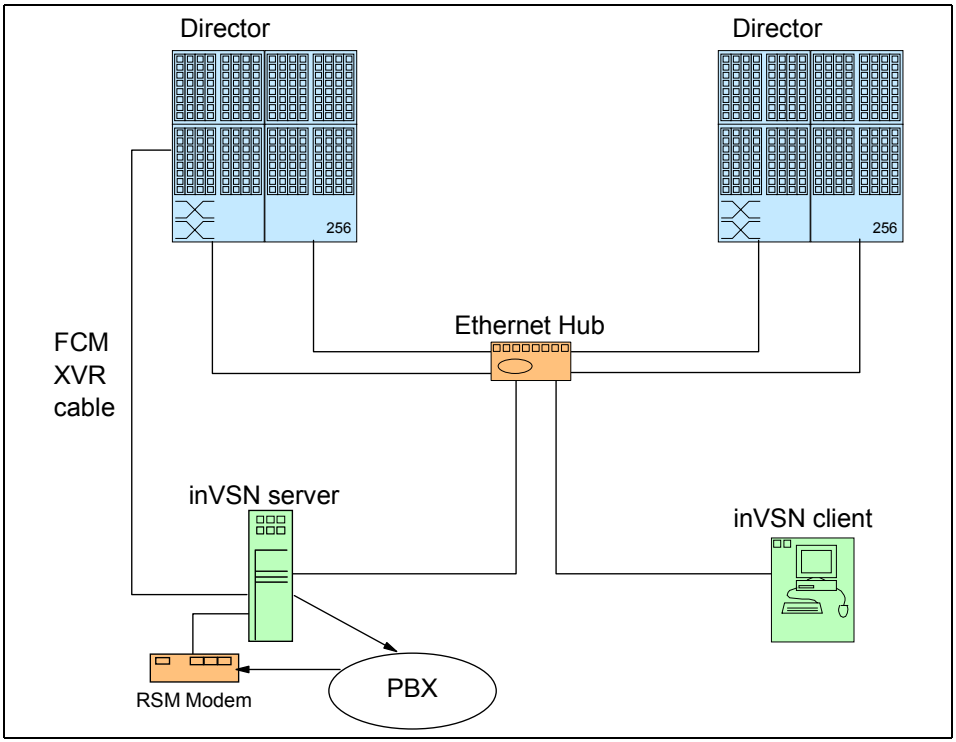


Figure 9-33 Typical CNT director communication setup

9.6.10 Setup validation

Once the Ethernet cables have ben installed, it is good practice to use the **`ping`** command from the DOS prompt of the inVSN server to determine whether the new TCP/IP addresses, and the cabling have been setup correctly.

Once communication between the server and the FCME has been established, the FCME will synchronize the times of the inVSN and modules to ensure event log timings are consistent.

9.6.11 Dial home settings

To set up the dial home information, you will need to select the configuration option on the menu bar of the inVSN server. In the configuration option, enter the contact details. External communication can be via telephone, pager, or e-mail.

9.6.12 inVSN Enterprise Manager installation

There are a number of different ways in which the inVSN software can be delivered. If no additional software feature codes are configured when ordering the CNT director, you will receive a CD that contains the inVSN server and client code. This CD is only licensed for one copy of the server and one copy of the client.

The server and client can be installed on a customer owned and designated PC, as long as it meets with the following minimum recommended specifications:

- IBM NetVista™ M41 PC or compatible PC
- 1.8 GHz Pentium 4
- 256 MB SDRAM
- 4 MB video RAM
- 40 GB hard drive
- 1.44 MB diskette drive
- CD-ROM
- 1 parallel port
- 2 x Ethernet 10Base-T/100 Base-TX ports
- IBM or equivalent mouse
- 2 external serial com ports (debug serial cable, external modem)
- 1 internal com port (internal modem)
- 17-inch 1280 x 1024 x 256 SVGA monitor
- Windows 2000 Professional
- Internal and External modem compatibility
- Analog protocol support: V.90 56kb/s ITU Standard, 3Com/US Robotic/Multi-Tech (international) hardware compatible modems recommended (do not use Winmodem software controlled modems)

Note: 256 MB memory will handle a configuration with up to 4 port blocks (or 256 ports). 768 MB memory is required for configurations with up to 12 port blocks (or 768 ports). 1 GB memory is required for configurations with up to 28 port blocks (or 1,792 ports)

If f/c 7600 is configured, the workstation pack will include a PC, monitor, Ethernet Hub and Modems. The management software comes pre-installed on the PC.

If additional inVSN client licenses are required, they can be ordered by specifying f/c 7201.

9.6.13 Installing

If installing from CD, both the server and client are installed by running **setup.exe** on the CD provided and following the on-screen instructions.

The server and client software must be kept at the same software level.

The server software needs to be started prior to the client. To start the server or client software you will need to double click on the required icon and enter a username and password.

Figure 9-34 we show the primary inVSN option screen.

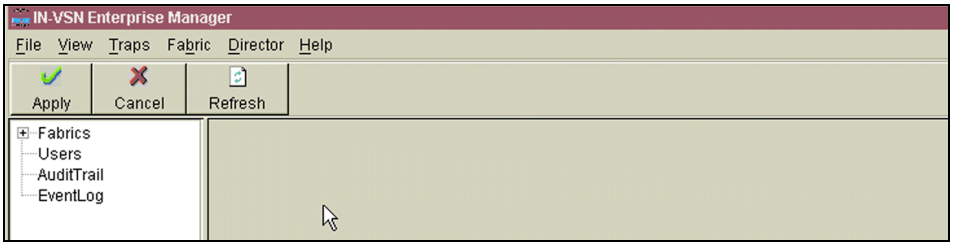


Figure 9-34 inVSN primary option screen

It is possible to run multiple inVSN clients on the same PC but, as the inVSN software is based on Java, each client version will require approximately 40 MB of memory.

If there are too many active clients, this may degrade the performance of the workstation PC. It is recommended to have only one active inVSN server and not to exceed 16 active inVSN clients. Additionally, there should be no more than four inVSN clients per workstation PC.

For high availability, we recommend that two inVSN management server PCs are installed in a high availability cluster, and only one should be active.

9.6.14 Management software

The GUI based inVSN Enterprise Manager software operates on a Windows 2000 platform using a Java Web browser and offers the following capabilities:

- Viewing and defining module and port configurations
- Managing fabric zoning
- Monitoring alarms and system performance

- Invoking system diagnostics.
- Displaying a graphical overview of port and system configuration
- Displaying port and fabric Utilization Statistics
- Displaying the status of individual modules
- Performing system diagnostics through the error and event logs

User levels

There are four levels of inVSN administrator access that have varying degrees of authority. The highest level is reserved for CE use only, as there are command options that could be destructive if used incorrectly. The three levels of authority that can be used are viewer, operator, and administrator. Refer to the *inVSN Enterprise Manager Software Installation and Operation Guide,* 9115687-203-04 for a breakdown of the privileges for each type.

Screen layout

The visual display of the inVSN Enterprise Manager software is divided into six areas. Five of these areas are shown in Figure 9-35.

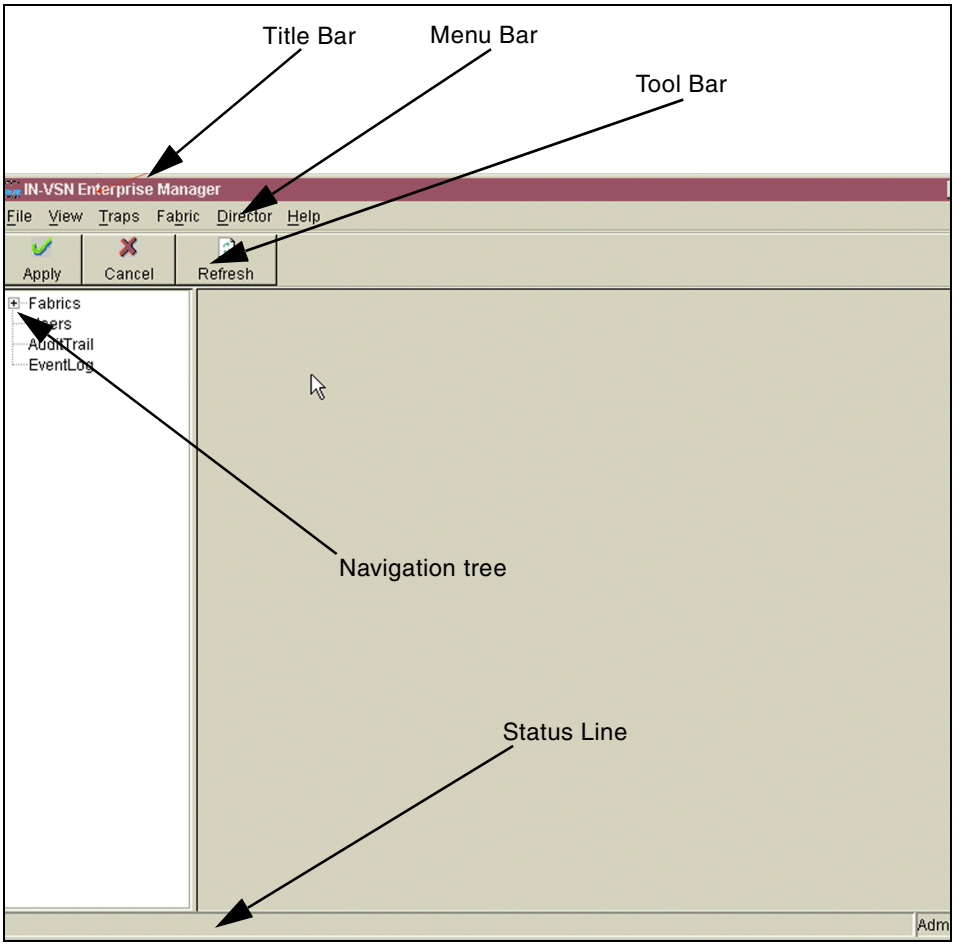


Figure 9-35 Display of inVSN available options

The title bar

The title bar displays the inVSN software version, Enterprise Manager for the client interface and Enterprise Manager Server for the server interface.

The menu bar

The menu bar contains four menu selections:

- File
- View
- Trap
- Help

Each of these menu selections represents a function of the Enterprise Manager software, and displays a menu of related options when selected.

The tool bar

The tool bar is a shortcut for navigating through the menu. It displays graphic images, referred to as buttons. Three square buttons are located in the Tool bar. The buttons allow you to perform the activity represented by a graphic image. A small *balloon help* display shows the type of activity which can be performed when the arrow of the mouse pointer is on the button. By clicking on the button, that activity will be performed.

The status line bar

The status line is a grey bar, located below the Navigation Tree and Details Panel.

The navigation tree

The navigation tree displays information specific to the server module of the inVSN Enterprise Manager which is running. You can use the Navigation Tree to select and display information specific to available Fabrics, Directors, Switches, Boards, or Ports. You may also view Event Log and Audit Trail information by using options on the Navigation Tree.

The details bar

Although not shown in Figure 9-35, at the detail bar information specific to the component you chose in the Navigation Tree is displayed here. You may also effect changes upon certain components, view and acknowledge traps, and perform other tasks related to the CNT director and software

9.6.15 inVSN Enterprise Manager database

The inVSN database can be set up for automatic backup. In the menu option of the inVSN server, select Configuration, Autobackup. You can then define the backup filename and how regularly to back it up.

To provide recovery protection, this backup copy could then be transferred across the network to another workstation or secured to diskette.

If the disaster recovery site has a separate backup fabric, to ensure that the primary and backup SAN fabrics zone settings are synchronized, there is an option to import a primary backup copy into the disaster recovery fabric.

9.6.16 Auditing and event logging

The CNT director will keep an audit trail and log events.

Audit trail

The audit trail option is accessed using the navigation tree. The audit trail can be used to track any soft changes performed and it includes timestamp and user name information.

inVSN management software allows you to monitor user activities using an Audit Trail.

All user levels (*admin*, *oper*, *viewer*) can access the audit trail.

All activities are logged and categorized into different types. Following are some examples of these operation types:

- User login
- IP address of user login
- Fabric definition
- User definitions
- Name server zoning
- Switch name changed

To look at the audit trail, click **AuditTrail** in the navigation tree, as illustrated in Figure 9-36.

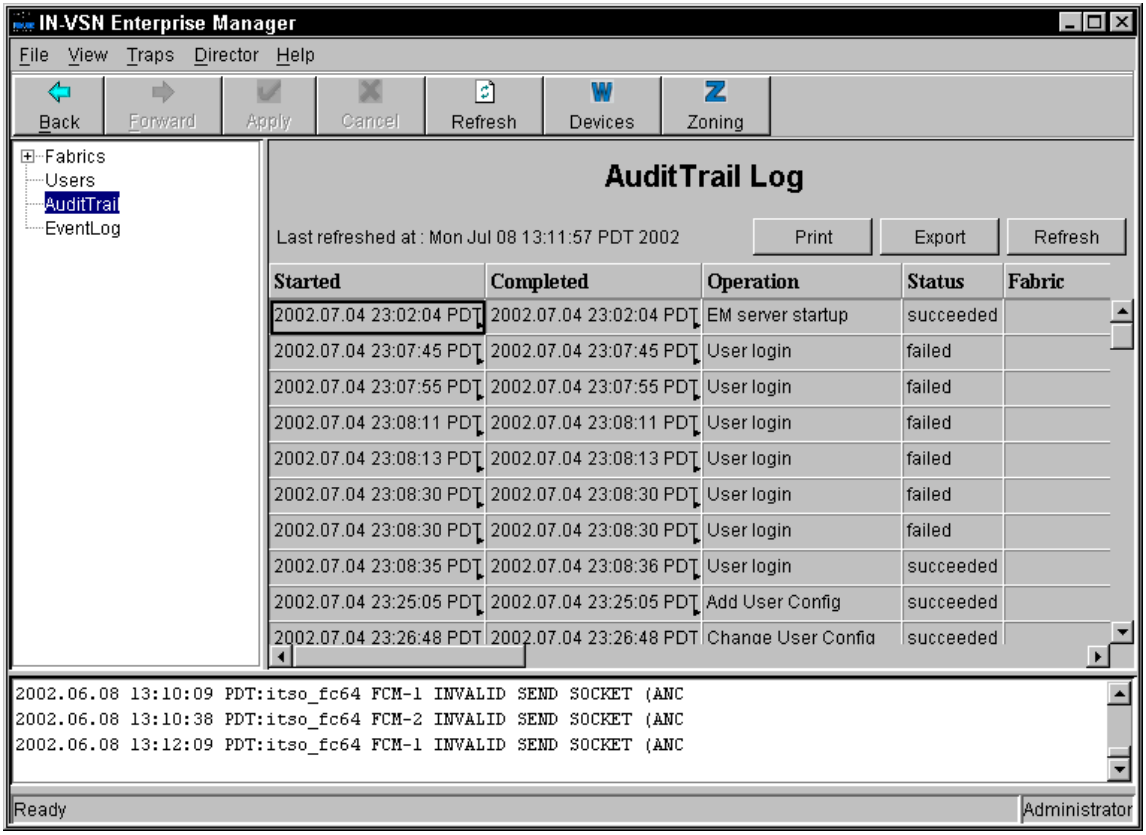


Figure 9-36 inVSN: Using CNT Audit-Trail to monitor user activities

9.6.17 Using the inVSN event log

The event log contains all important events that have occurred. This includes events triggered by users and events caused by other external or internal influences, such as FRU failures.

To read the event log, click **EventLog** in the navigation tree of inVSN, as shown in Figure 9-37.

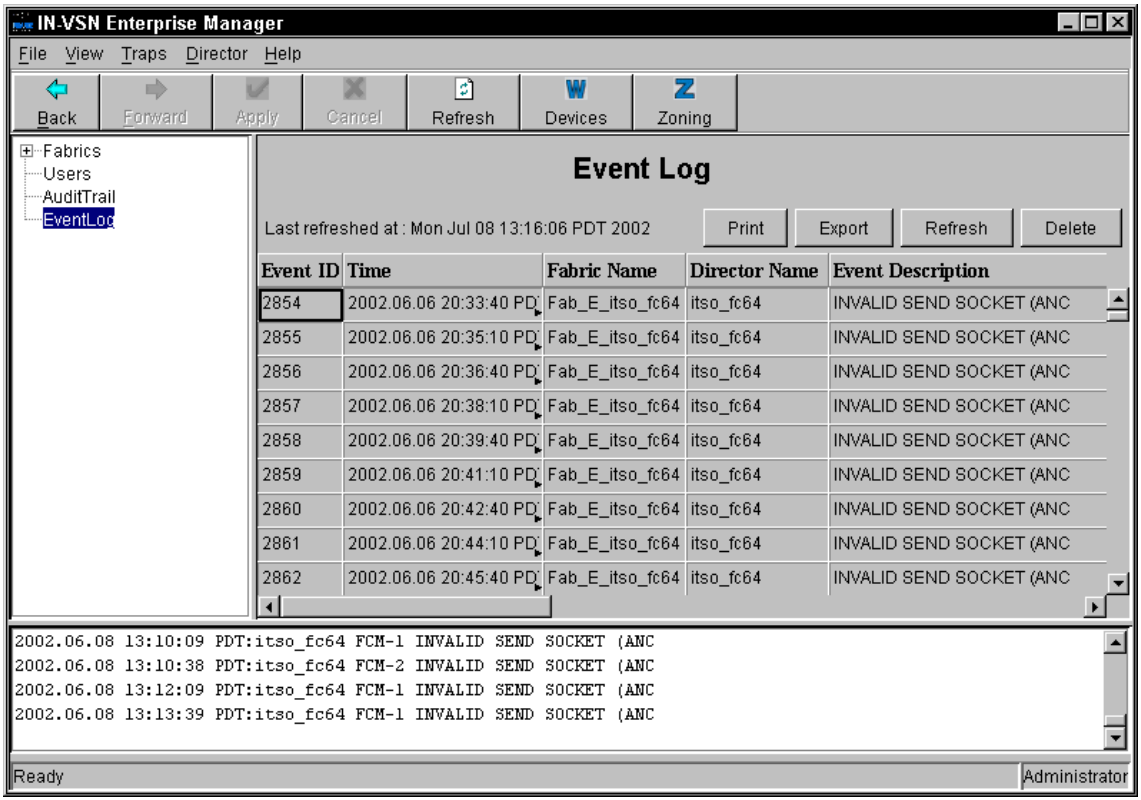


Figure 9-37 inVSN: Accessing the CNT event log

This event log is valid for all fabrics managed by this inVSN server. It includes the logs of multiple fabric devices and even multiple fabrics.

To file this log for future usage, you are able to export this log. Two file types can be used to export the log:

- Comma Separated Variable files (*.csv)
- Text files (*.txt)

9.7 Zoning

A fabric is an intelligent, shared, interconnected scheme of Fibre Channel connected servers, switches, directors, and storage nodes.

You can have up to 224 domain IDs, the first Octet in the FC address is the fabric ID and the higher 4 bits are the chassis (switch) ID. These director specific domain IDs prevent duplicate domain IDs when two directors are connected together.

In Figure we show multiple SAN switches and FC/9000 directors connected to create an individual faric.

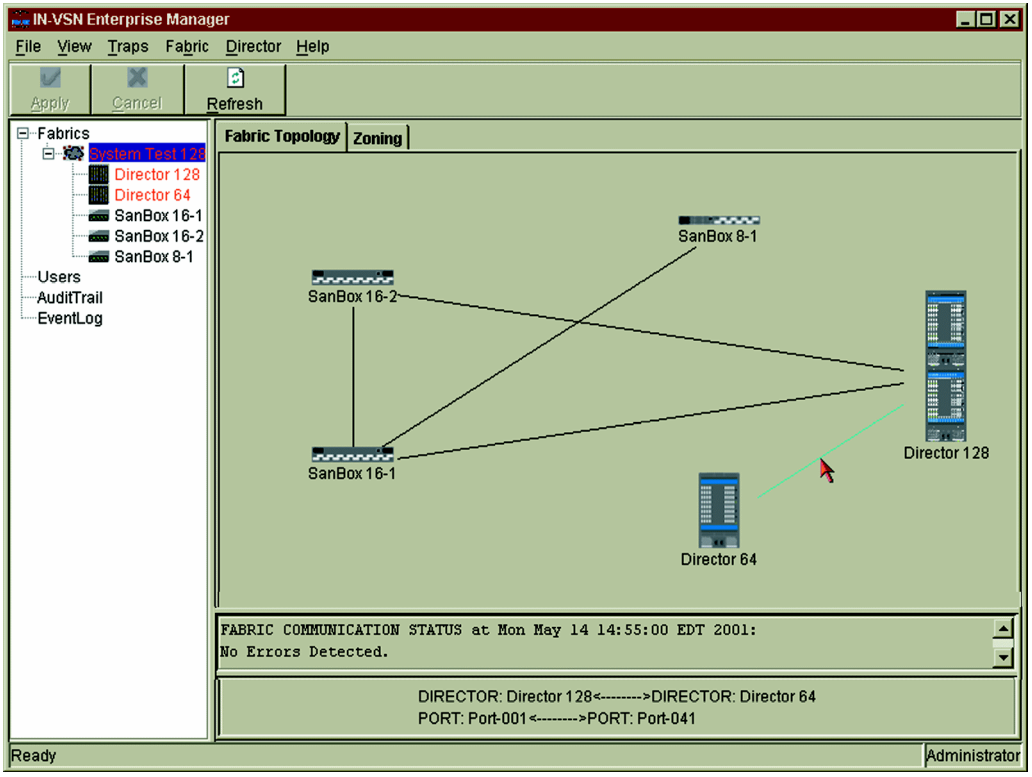


Figure 9-38 inVSN Enterprise Manager screenshot of multiple switch connected tn a fabric

9.8 Zoning

One of the basic purposes of a SAN fabric products is to enable or disable communication between the different ports (devices) attached to them.

In most cases it is helpful to limit the potential access of ports. Zoning provides an effective tool to limit and control the communication between fabric ports.

There are multiple reasons to limit access and include:

- We may want to avoid Windows servers seeing all disks in a fabric. Otherwise there would be a high risk of getting signatures written on all disks, which would then mean these disks become unusable by other operating systems.
- For security reasons, we want to limit access to disks with confidential data to only selected servers.
- We need to control the number of paths a FC host adapter has to a specific disk. This is because not all environments are flexible in their usage of multipathing software.
- Storage Virtualization with the IBM SAN Volume Controller requires that hosts cannot also see the same back-end storage that is to be virtualized.

Directors provide three kinds of zoning methods:

- Name server WWN zoning
- Hardware enforced WWN zoning (available only on the SAN256N)
- Broadcast zoning for IP over FC

9.8.1 Understanding CNT name server zoning

This zoning allows you to designate devices using their WWPN. This means you can group devices by WWNs with WWN zoning. These zones can then be grouped into *zonesets.* All zones within a zoneset are in effect at the same time and only one active zoneset can reside on the fabric.

Enterprise Manager will let you manage several zonesets across the fabric with only one of them active at a time. Note that the active zoneset cannot be modified.

One zoneset may be composed of 1 to 256 zones. That zoneset may accommodate up to 3500 member devices. Note that all zones are present in the fabric and available to any user which may have access to the fabric.

Previous versions of FC/9000 documents called E_PORT mode WWN zoning. This is the default for the current FC/9000 and is standard fabric name server zoning.

Utilizing WWN or WWPN name server zoning gives greater flexibility to manage a fabric as devices can be moved across the fabric without having to change zoning configurations.

A *name server table* is used to implement zoning on both the FC/9000 and the SAN256N. Name server tables store information about nodes that have logged into the fabric.

Name Server zoning can also be implemented utilizing domain id/port which would restrict the access of affected ports to this information. All unzoned nodes have full access to the name server table.

This means that zoned ports logging into a fabric will only get name server table information of other ports which are part of the same name server zone. Unzoned ports are considered as being part of an *orphan zone* that is not visible to users, but allows communication for all unzoned nodes.

Because this kind of zoning relies on the proper usage of FCP protocol commands by filtering RSCN commands to attached nodes, it is also widely called *soft zoning*.

However, since name server zoning relies on the correct usage of FCP commands, there is still a risk that malfunctioning nodes would affect other ports even if these other ports are not members of the same zone.

Due to its strength and flexibility, name server zoning is widely used in open systems environment.

Name server zoning rules

Name Server zoning is implemented by filtering the RSCNs to attached nodes. This is considered soft zoning.

- Zone members on the FC/9000 and the SAN256N can utilize:
  - WWN or nickname
  - Domain ID/Port
  - Node FC id (available only on the FC/9000)
- Each name server zone needs to get a unique number and a name.
- As many as 256 zones are possible in a fabric.
- Director ports which have private nodes attached to it need to be set to TL mode. Name server zoning is not effective for TL_Ports.
- TL ports are only available on the FC/9000 1 Gbps IO module.
- All nodes that are not part of any name server zone have unlimited access to the name server table. This is also true if no name server zone is

implemented at all. Consequently, by default, all attached nodes have access to the name server table. Only those nodes that are part of name server zones will have limited access to this information.

  - The only exception is a TL_Port Config list which explicitly allows access to the specified TL_Port even if name server zones exist that normally would imply otherwise. TL ports are only available on FC/9000 1 Gbps IO module.

The following rules are enforced with CNT directors:

- Name server zones cannot span logical domain boundaries (logical domains are available only on SAN256N).
- You can use name server zoning to further limit access between nodes. Name server zoning can be hardware enforced if desired when implemented in the SAN256N.
- TL-CFG lists are for TL_Ports only. They overrule the name server principles. So, TL_Port attached nodes can only communicate with ports specified in the TL-CFG lists. TL ports are only available on the FC/9000 1 Gbps IO module.

Logical domains

With the new Ultranet Multi-services director, logical domains can be utilized with the purchase of the SAN services bundle.The SAN256N physical director may be partitioned into a maximum of four logical domains (logical switches). A single domain may include a maximum of 256 ports. A logical domain can consist of two or more I/O modules. Partitioning also provides the ability to connect the logical domains together with virtual interswitch links (ISLs). Each logical partition can be managed independently, and firmware upgrades of its components can be done independently with respect to other partitions.The scenario illustrated in Figure 9-39 describes a possible layout utilizing logical domains.

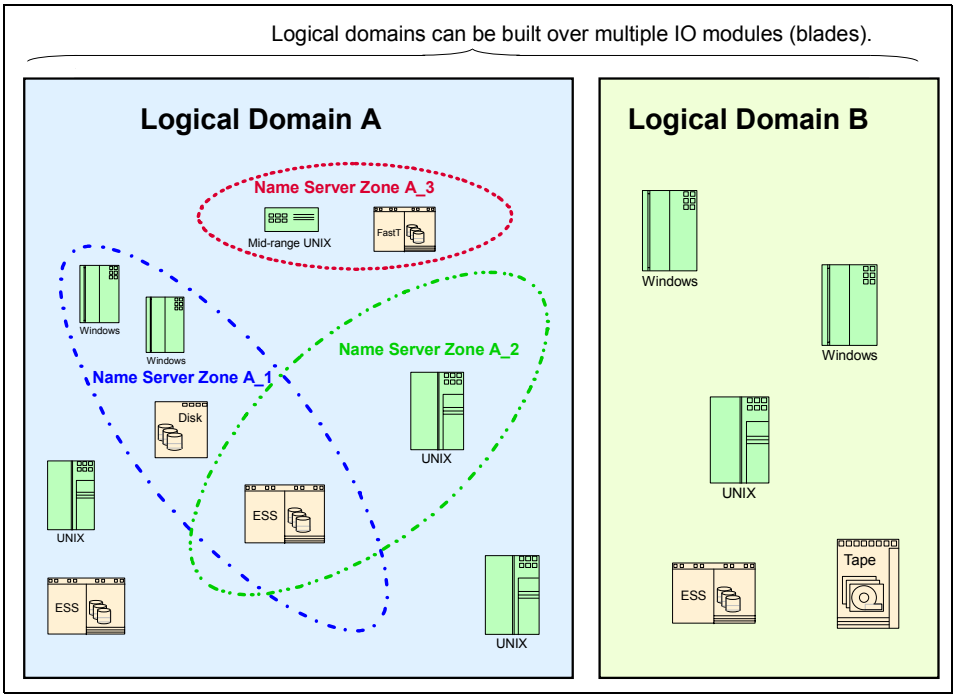


Figure 9-39 CNT SAN256N logical domains and name server zoning

In this layout:

- All three defined name server zones are within Logical Domain A. They cannot span over to Logical domain B.
- Name server zone A_1 and A_2 are overlapping. Both have an ESS node as a member.
- All members of name server zones can only access other members within the same name server zone.
- All remaining nodes in Logical Domain A which are not part of any name server zone are part of the *orphan zone*. This allows communication between these three nodes.
- There is no name server zone at all in Logical Domain B. All members can have access to any other member in Logical Domain B since they are put into the orphan zone. However, this orphan zone is limited to Logical Domain B.

9.8.2 Understanding hardware enforced WWN zoning

Hardware enforced WWN zoning utilizes special hardware within the IO modules in the SAN256N to enforce the zoning, rather than relying on filtering of RSCN commands as soft zoning does. Hardware enforced WWN zoning can be

enabled for the entire SAN256N switch, or turned on for specific zones by a special naming convention. This option is only available on the SAN256N.

9.8.3 Understanding CNT broadcast zoning

Broadcast zones focus the distribution of broadcast messages to only those targets that need to receive them. This results in reduced fabric traffic and eliminates unnecessary message processing. Broadcast zones can overlap and are also assigned by director ports.

The granularity in which broadcast zones are built is one director port.

However, broadcast zoning is used for IP networking. SAN environments do not typically exploit this kind of zoning.

9.8.4 Managing by device WWN

You can access and specify information for devices across the fabric for which Enterprise Manager has name server information. It is used to define nicknames for devices (servers and storage). For each one of them, you can set a nickname.

The nicknames can then be used in the zoning implementation, for example, in order to locate devices and hosts.

9.8.5 Cascading of directors

To further increase the amount of available FC ports, we can connect multiple directors using FC links. The FC/9000 director and the Ultranet Multi-service Director are fully interoperable.

The process of connecting directors using ISLs is known as *cascading*.

The motivation to use cascaded directors is not limited to just increasing port counts. Cascaded directors can be used to create a SAN that spans large distances.

Inter-Switch Links

Inter-Switch Links (ISLs) are commonly used in many SAN implementations as a way to scale the number of “any-to-any” ports in a fabric beyond the number of ports that a single domain can support. Since the directors support large port counts within a single domain, from 24 all the way to 256 ports, the motivation to use ISLs with the director is more likely to be for implementations involving distance, such as connecting remote data centers in a campus or metro environment.

The specific implementation of ISLs for the FC/9000 director varies, depending on software release version installed. For software versions prior to Release 3.0 the ISL ports were implemented in a proprietary fashion and are referred to as T_Ports (Trunk Ports), and for Release 3.0 and later the industry standard E_Port as defined in FC-SW-2 is implemented to provide the interconnect protocol for ISLs. These two methods provide equivalent functionality and the migration to E_Port demonstrates the industry direction of adoption of standards as they evolve. The advantage of implementing the FC-SW-2 methods for E_Port is to allow the CNT director to interoperate with a large number of Fibre Channel switches, directors, routers, and bridges that are also designed to conform to the FC-SW-2 standard.

The directors support ISLs for creating cascade, mesh, core, and multi-stage topologies for FCP as well as supporting FICON cascaded environments. When using E_Port ISL’s, WWN zoning, as defined in FC-SW-2, can be used for grouping of ports and devices in an extended fabric topology the same way they are implemented in a single domain fabric. In addition to zoning, the director provides an optional Fabric Security mode which when enabled requires an administrator to establish and configure each director with the list of “Known Domains” which have been authorized to participate in the fabric.

The distance supported for ISLs with any Fibre Channel based SAN is a function of buffer credits as well as optical infrastructure. Buffer credits are an integral part of Fibre Channel’s designed in high performance and reliability, and by definition an increasing number of buffer credits are needed as distance increases. The director has buffer credits dedicated to every port to support distances of over 100 km, without affecting other ports. Care must be taken to ensure the optical infrastructure including transceivers (GBIC, SFP’s), fiber optic cables (50, 62.5, 9 micron), and/or optical repeaters and DWDM are implemented as needed.

ISLs can be run on either shortwave GBICs or longwave GBICs. The usage of longwave GBICs enables the creation of long distance fabrics. The actual supported distance between directors can be further expanded by using channel extenders or multiplexers.

Figure 9-40 shows an example where cascaded directors are used to allow host based mirroring over long distances:

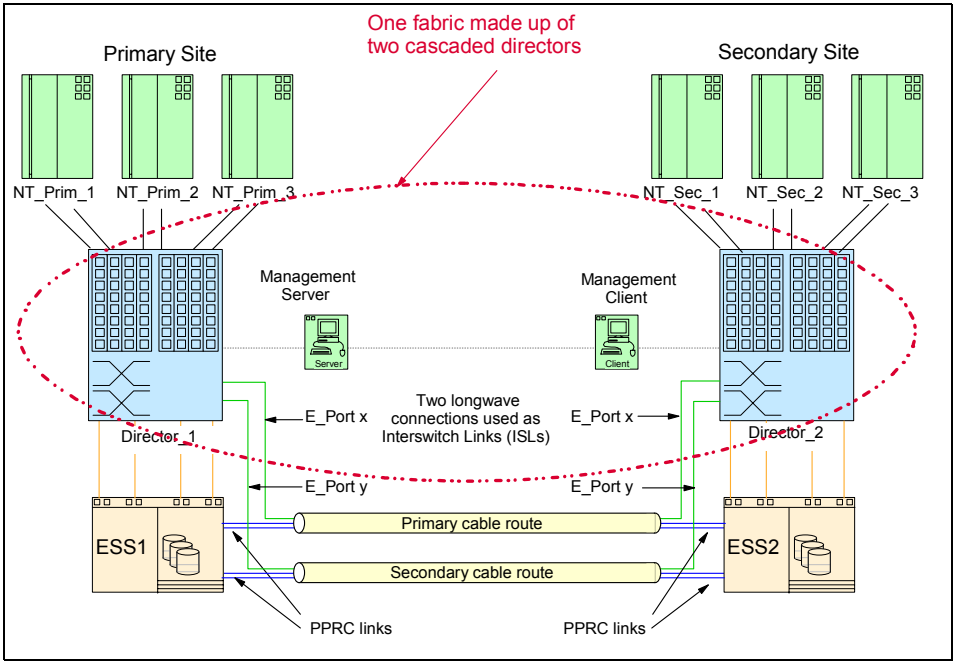


Figure 9-40 One CNT fabric consisting of two cascaded directors

Cascading is a very effective and easy-to-use method to extend distances and increase maximum port count. However, we need to consider some of the implications:

- By connecting multiple directors, they will become one fabric inheriting all the rules for one fabric, including these characteristics:
  - Name server zones are valid fabric-wide.
  - Numbers and names of name server zones must be unique fabric-wide.
  - Each director or switch in a fabric must have a unique switch ID.
  - When having no zoning implemented at all, then all nodes have access to any other node fabric-wide.
  - When actually using ISLs for I/O, there will be a delay called latency. The actual usage of a ISL is commonly called a hop. Typically up to three hops are supported for a particular I/O
  - .A logical domain in the SAN256N follows the same rules as if it were a separate switch. One advantage of using logical domains to carve smaller switches from the SAN256N is to have virtual ISLs. With virtual ISLs between two logical domains there is no latency or limit to the number of

hops. The virtual ISL has zero cost, and there is no need for ports between the domain and another domain within the same SAN256N.

- E_Ports cannot be part of name server zones.
- Up to eight directors can be cascaded to create one fabric.

9.9 Configuration guidelines

Prior to any detailed CNT director port design, a high level review should be conducted. The review should include a SAN cabling blueprint and any specific requirements, for example, availability and performance considerations, that need to be taken into account. Based on the output of this review, a preliminary port layout could be drafted.

Refer to the *inVSN FC/9000 Fibre Channel Director Site Planning Guide For E_Port Configurations* for additional information.

In Figure 9-41 we show a sample form extracted from the guide. This form could be used, in conjunction with a customer review, to design the location of the GBICs across the FIO blades.

Port Server/Device Configuration

FIO #1	Device	Model Number	O/S code Application	Interface Type	HBA	HBA Model	HBA Interface	Driver/FW
Port 1	IBM pSeries	7040-681	AIX 4.3.3	FC	Emulex	LP8000	FC	
Port 2	IBM ESS	ESS (2105-800)		FC	Emulex	LP8000	FC	LIC level 2.1.0
Port 3	IBM zSeries	2064	z/OS V1R4	FICON		f/c 2320	FICON	
Port 4	IBM ESS	ESS (2105-800)		FC	Emulex	LP8000	FC	
Port 5	Sun	450	Solaris 2.5	FC	JNI	FC64-1063-N	FC	
Port 6	Sun	420R	Solaris 2.9	FC	JNI	FCE-6460-N	FC	
Port 8	Intel	x440	Windows	FC	Qlogic	QLA2340	FC	

Figure 9-41 Port design form

9.9.1 Port positioning for performance

As the CNT director is a full non-blocking device, there is no requirement for specific port placement for baseline performance.

However, to achieve optimum performance, adjacent ports should be used, as this offers a throughput capability of less than 0.6 micro seconds. Adjacent ports would be ports 1 and 2; 3 and 4; 5 and 6; or 7 and 8; only.

In Figure 9-42 we describe the position required for a port to be an adjacent port.

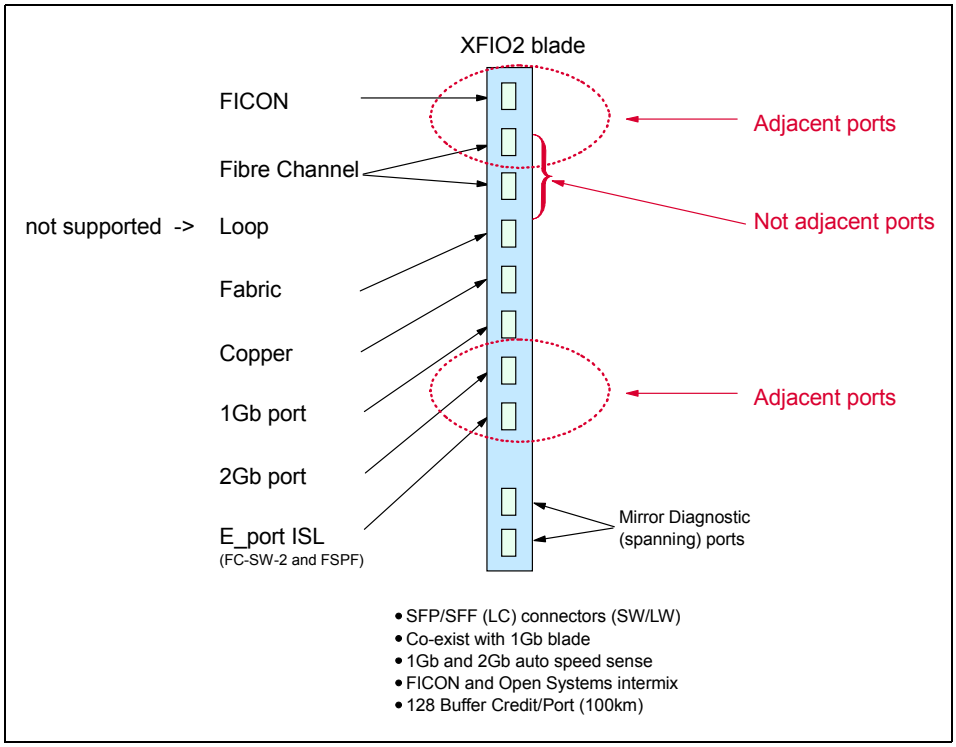


Figure 9-42 Description of an adjacent port

For any other port configuration it will be less than 2.7 microseconds.

The Ultranet Multi-services Director (SAN256N) has no optimum port placement guidelines. All I/O will go through the central switching core, and is fully non-blocking with the same performance regardless of port placement.

9.9.2 Port configuration flexibility

As all 256 ports can have the full 128 buffer credits, we are able to configure the CNT director to use a high number of longwave SFPs with no performance impact to the director. If fiber optic cables were in place, it would be possible to implement a cross site solution, as detailed in Figure 9-43.

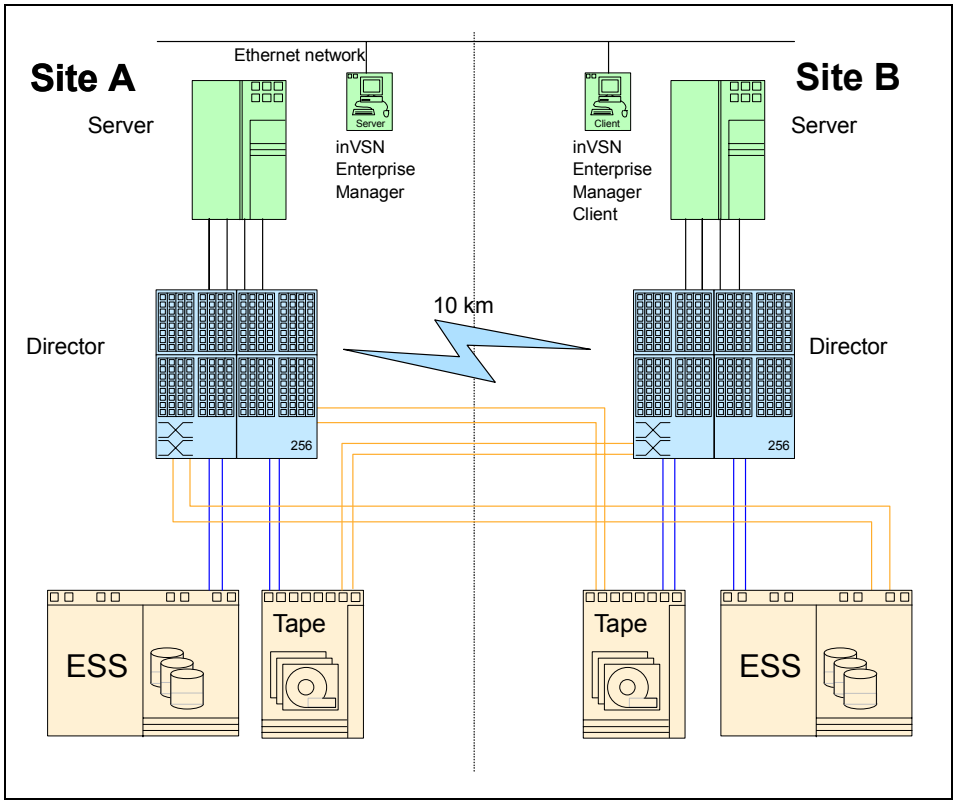


Figure 9-43 Remote site distance solution

Servers in Site A have the ability to access storage resources in Site B and we could add additional resilience by implementing clustering and remote copy.

9.9.3 Hop count

IBM currently supports a fabric of up to 8 directors with a maximum of three hops across the SAN, as shown by “Hoppy”, the hop count kangaroo, in Figure 9-44.

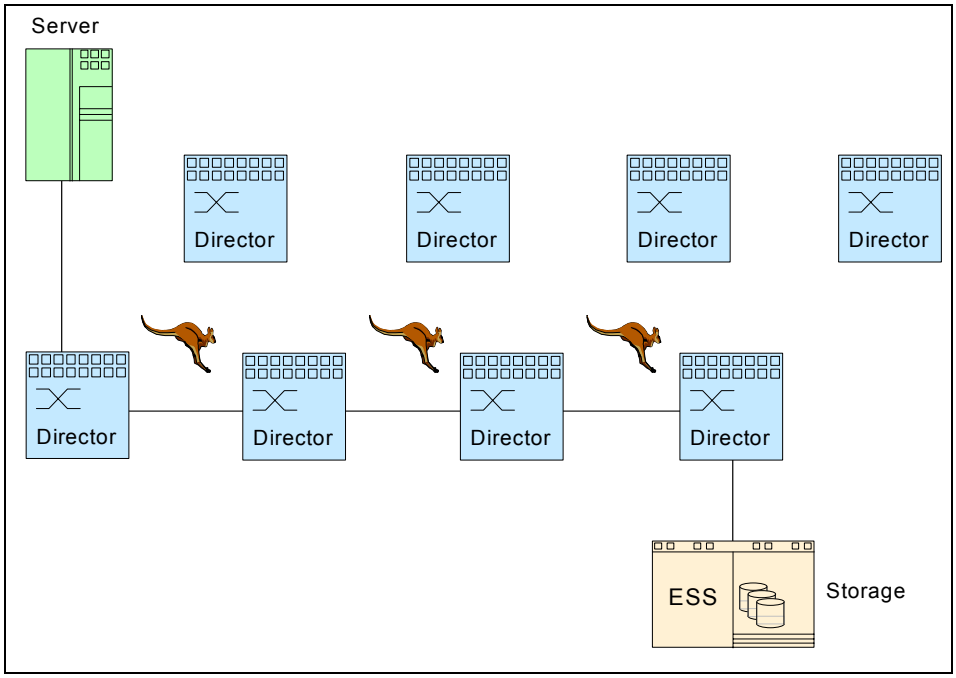


Figure 9-44 Single fabric can consist of 8 directors with maximum hop count of 3

9.9.4 FC/9000 port tuning options

There are several pull-down options available for port tuning when connecting to devices with older architecture. In Figure 9-45 we show the inVSN port display screen and pull-down option.

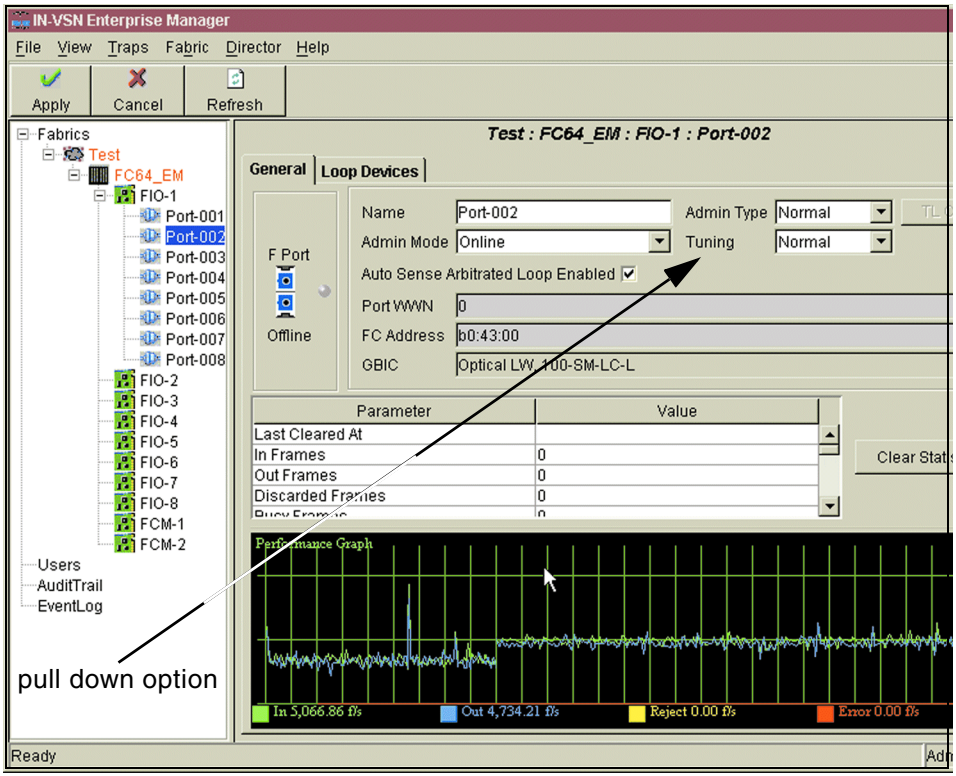


Figure 9-45 Port display screen

Non-I: Non-Interleaved option

This option prevents sequences from different sources that are bound for a single destination from being interleaved. Once a sequence has begun, the CNT director will not transmit frames from any other source other than the one which began the sequence.

This mode is recommended for Tachyon-based adapters that are being used for IP traffic. Non-I must also be used on any cross connect port for this IP traffic in a cascaded fabric.

Min-I: Minimize-Interleave option

This option, while not preventing interleaved sequences, minimizes their extent. Once a sequence has begun, the CNT director will continue to transmit from the same source, as long as frames are available for transmission, or an end-of-sequence occurs. However, if no frames are available for transmission, then a new source will be started and held until it has no frames to transmit, or an end-of-sequence occurs.

This mode is recommended for QLogic 2xxx HBAs.

Frame-L: Frame Limit option

This option limits the number of frames that can be transmitted during a single loop tenancy to 32.

This option is recommended for JNI HBA's based on Adaptec ASIC hardware, and Adaptec HBA's.

Normal option

No tuning is applied. This option is recommended for all situations that have not been mentioned in the preceding topics.

9.9.5 Interoperability and legacy device toleration

The directors now use the industry standard E_Port (in lieu of T_Port which was used to interconnect previous models of the FC/9000 director) for all FC-SW to FC-SW connections (including FC/9000 to FC/9000 and SAN256N). The E_Port implementation adheres to the FC-SW2 and FC-GS-3 specifications. This standard also includes interoperability with other vendors switches but care should be taken to understand any restrictions that may occur when connecting to different vendors equipment. Refer to the *inVSN FC/9000 Fibre Channel Director Site Planning Guide for E_Port Configurations* for more details.

9.10 Private and public device support

Before we describe the additional flexibility realized by the translative loop port (TL_Port) of the CNT director, we will describe addressing and the two types of Fibre Channel loop.

9.10.1 24-bit and 8-bit addressing

Port ids are 24 bits in length, and a 24-bit address allows 16 million addresses. That's why Fibre Channel's switched topology allows the interconnection of 16 million devices.

Arbitrated loops use a special form of a port id called the Arbitrated Loop Physical address (AL_PA). AL_PA's only use the low order 8 bits for addressing. Although 8 bits provides a raw bit space of 256 (0-255), the AL_PA's come from the 8b/10b code and must have neutral disparity. There are only 134 encoded words with neutral disparity and the AL_PA's can use 127 of these.

Arbitrated loops not connected to a switch (private loops) have zeroes for the upper 16 bits of their Port_IDs.

9.10.2 Private loop

Private loop is the simplest Fibre Channel protocol. Private loops are generally built using hubs, and communication is handled by passing messages around the loop from one device to the next in order to gain control of the loop and to communicate on it. Because each device is dependent upon the device next to it for communication, private loops tend to be unstable.

For example, adding, removing, or even rebooting a device causes other devices on the loop to pause. This pause forces every device to examine what changed in the environment and to adjust for that change. This can be extremely disruptive when data is being transferred at the same time a change of state takes place on an arbitrated loop.

In a controlled, stable and consistent environment, private loop can perform adequately for many applications. The main reason for its initial popularity is due to its relatively low cost. However, when performance, stability and management are required, a higher performing Fibre Channel communication mode is necessary.

Importantly, a private loop is only known to itself and in some cases it is not possible, unless some form of translation takes place, for a private device to see, or be seen by, others.

9.10.3 Public loop

Public loop implies Fibre Channel arbitrated loop with hubs, switch or switches. Public loop adds a number of key advantages over private:

- **Name server:** As described in the Fibre Channel standard, the addition of a name server adds instant control and stability to a loop because each device

registers with the name server on login. The name server then takes over the need for devices to pause and pass messages to one another each time an event happens on the loop. The switch internally controls the name server.

- **Management:** With a name server and switch management software, devices on a switch can be queried, examined, and brought online and offline. Errors can be logged and trapped with messages sent to an administrator who can then take any appropriate action.
- **Zoning:** With zoning, a SAN can be made more secure by separating and segregating data and devices. Zoning can be done by individual switch ports, using name server entries or by each devices unique world wide name.
- **Ability to address more than 126 devices:** In a loop only configuration, it is possible to address a maximum of 126 devices, not including the initiator. Through connection to a fabric, that range can be increased. This is in comparison to a private loop's 8-bit address which limits a fabrics size.

9.10.4 Translative loop mode

While there are clear advantages to public loop, there are situations where a private device cannot be upgraded to public loop. Changing a device from private to public simply requires a different firmware load, but there are devices which have been deployed which cannot support public loop either by design, or because there is no public loop firmware available. There are also situations where it may not be desirable to make a firmware change to public loop, but private devices still need to be able to communicate with fabric devices on a switch. In order for a switch to adequately provide bridging from public to private, it must be able to support the following three types of translation:

- Public initiator to or from a private target
- Private initiator to or from a public target
- Private initiator through the fabric to or from a private target on a different loop

Generally speaking, an initiator is a server and a target is a disk or tape device.

Public initiator to or from a private target

In this type of translation, the switch port that the private loop is connecting to must be assigned as a translative TL_Port, so it can expect a private loop to be attached to it. The private loop targets then log into this port.

When a public initiator sends a message to a private target, it first looks up the address of the private target. Then the upper 16 bits of its address are stripped off. This is necessary because a private target expects an 8-bit address.

Private initiator to or from a public target

Here the opposite occurs. The private initiator keeps track of public addresses using tables. When sending data to a public target, the upper 16 bits are added by performing a look up in these tables. This is necessary because a public target expects to receive a 24-bit address. To the private initiator, it will appear as if the public target is on its private loop.

There are some limitations in the number of devices that can be addressed when translating public and private loop. In general these methods are adequate for most implementations. To achieve a desirable and required performance, any translating must be done by the hardware.

Ideally, all Fibre Channel environments are comprised of fabric aware (non-private) devices. The benefits of this are the stability of having a name server, the control of much greater systems management as well as the ability to build large fabrics and to zone them for security. Unfortunately this cannot always be the case. The CNT director TL_Port available on the FIO 1G/s module provides for the seamless translation of data between all types and modes.

FC-AL support

The FC/9000 2 Gbps FIO2 modules and the new SAN256N IO modules do not support loop connections. Connecting FC-AL legacy devices to the FC/9000 and SAN256N requires an edge switch with FC-AL support.

9.10.5 Firmware level

The CNT director is shipped with the latest supported version of firmware. To locate the firmware version you are running, click on the **Help** option on the menu bar of the inVSN client primary menu screen and then select **About**.

Recording this information can be useful, as technical support will usually request firmware levels of the SAN components when placing a call.

To find out the latest level of firmware available, refer to the following Web site and select the Interoperability Matrix:

`http://www.storage.ibm.com/ibmsan/products/directors/index.html`

9.10.6 Trivial File Transfer Protocol

Both CNT director and the SAN256N utilize the Trivial File Transfer Protocol (TFTP), a standard IP utility to provide a non-disruptive firmware code load. The code update requires the CE userid and password.

TFTP allows you to:

- Load new director firmware
- Load new management control firmware into the director
- Retrieve and modify the current management configuration of the director management processes, including the IP network configuration parameters, and the SNMP configuration parameters

TFTP is for CE use only, it cannot be accessed by any inVSN interface options.

9.11 Power-up process

The power-up process for a fully populated 64 port CNT director is approximately three minutes.

At power-up, the attributes of all devices in a fabric are *discovered* using a firmware protocol embedded in the director's I/O processor. Under this protocol, standard Fibre Channel initialization mechanisms are applied sequentially until an appropriate response is received and a port configuration assigned.

In Figure 9-46 we show the CNT respond-mode *wait cycle*, which enables the director to adapt to typical environmental changes without compromising the integrity of the fabric.

Loop and fabric devices can be interchanged, additional directors and nodes added, failed devices removed and restored, and individual port components (for example, GBICs) replaced while the rest of the director continues normal operations.

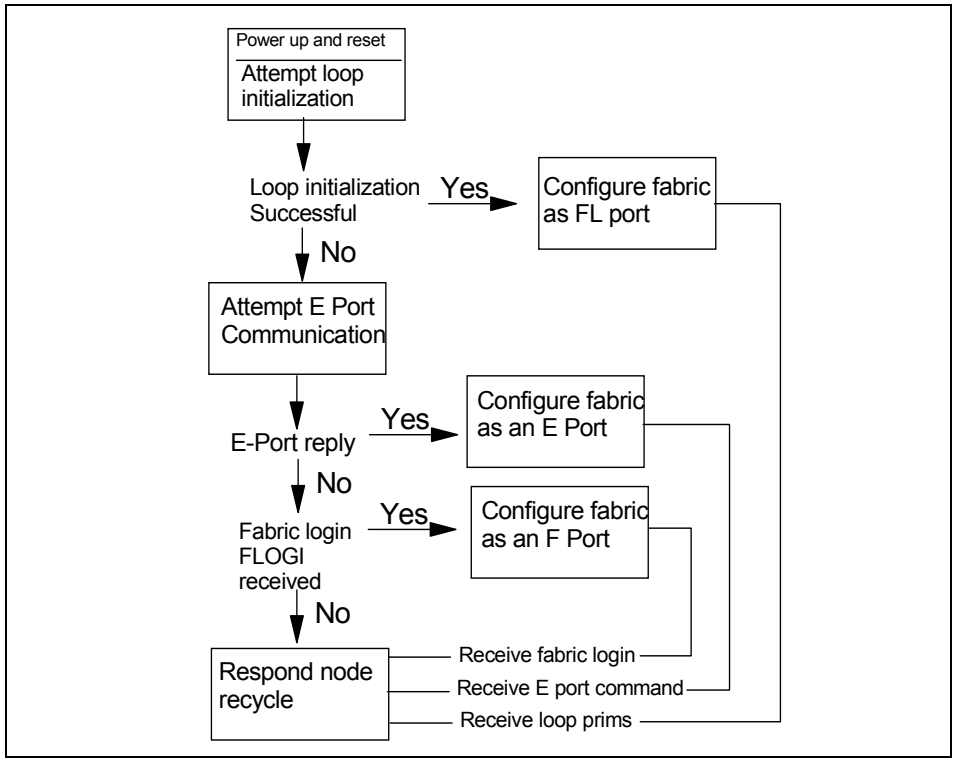


Figure 9-46 CNT wait cycle process

Power-up order

- Whenever possible, it is good practice to perform a site power-up in the following order:
  - Director
  - Targets
  - Initiators

9.12 Interoperability

For the current interoperability matrix, refer to:

http://www.storage.ibm.com/ibmsan/products/directors/index.html

Selecting the *Interoperability Matrix* will display a complete list of servers, disk and tape systems that are supported. This list also contains supported operating system versions and links to other Web sites that document the required HBA levels.

9.12.1 Standards

The FC/9000 director is SANmark certified for E_Port and FC-SW2 standards. The CNT director supports the following Fibre Channel Standards:

- FC-PH
- FC-PH-2
- FC-PH-3
- FC-AL
- FC-AL2
- FC-FLA
- FC-GS-2
- FC-FG
- FC-PLDA
- FC-Tape
- FC-VI
- FC-SB-2
- FC-SW
- FC-SW-2
- FICON
- Fibre Channel Element MIB
- FICON CUP
- Ethernet
- SONET
- GS-3-ANSI
- SMI-S-SNIA
- FC-SWAPI-ANSI:
- SNMP

The new 2 Gbps family eliminates FC-AL support. An edge switch will have to be used to provide this support if required.

Additional standards the SAN256N director supports are

- FC-SP-ANSI standards for security protocols
- FC-SW-3

- iSCSI
- FCIP

9.13 Feature codes

The following information is extracted from the IBM/CNT support web sites. Consult the web site for the latest products and features.

http://www.storage.ibm.com/ibmsan/products/directors

9.13.1 Ultranet Multi-Service Director

- 2042 Model C40—CNT FC/9000 Director Cabinet with space for two CNT SAN256Ns; two FC/9000-64 Directors; one FC/9000-128 Director or one FC/9000-256 Director with two adjacent cabinets
- 2042 Model C80—CNT SAN256N Cabinet with space for one or two SAN256Ns
- 2042 Model N16—CNT SAN256N with sixteen I/O module slots, dual control modules, three core switching modules, dual power supplies and redundant fans per chassis.
- FC 2214—Shortwave 2Gbps SFP transceiver 4 pack
- FC 2224—Longwave 2Gbps SFP transceiver 4 pack
- FC 2235—Extended longwave 2Gbps SFP transceiver (35 kilometers)
- FC 2280—Extended longwave 2Gbps SFP transceiver (80 kilometers)
- FC 5316—SAN256N 16-port 2Gbps I/O module
- FC 7301—FICON CUP
- FC 7302—FICON cascading
- FC 7303—SAN services bundle
- FC 7600—Enterprise manager server bundle
- FC 7603—Ethernet hub, 8 ports
- FC 7604—Ethernet hub, 24 ports
- FC 7640—Vertical fiber cable management set

9.13.2 FC/9000 Fibre Channel Director

- 2042 Model 256—CNT FC/9000-256 Fibre Channel Director with I/O Modules (32ports), inVSN Enterprise Manager and one inVSN Enterprise Client
- 2042 Model 128—CNT FC/9000-128 Fibre Channel Director with I/O Modules (24 ports)
- 2042 Model C40—CNT FC/9000 40U Cabinet
- inVSN Enterprise Manager and one inVSN Enterprise
- FC 2010—1 Gbps shortwave GBIC (multi-mode fiber)

- FC 2020—1 Gbps longwave GBIC (single-mode fiber)
- FC 2030—1 Gbps extended longwave GBIC (80 kilometers)
- FC 2080—Extended Distance longwave 80 Km SFP (2 Gbps)
- FC 2210—2 Gbps shortwave SFP transceiver (multimode fiber)
- FC 2220—2 Gbps longwave SFP transceiver (single-mode fiber)
- FC 2235—2 Gbps extended longwave SFP transceiver (35 kilometers)
- FC 2308—CWDM Extended Distance Module, 8 Port Mux
- FC 2316—CWDM Extended Distance Module, 16 Port Mux
- FC 2321—1470 nm Wavelength Specific SFP transceiver, Gray
- FC 2322—1490 nm Wavelength Specific SFP transceiver, Violet
- FC 2323—1510 nm Wavelength Specific SFP transceiver, Blue
- FC 2324—1530 nm Wavelength Specific SFP transceiver, Green
- FC 2325—1550 nm Wavelength Specific SFP transceiver, Yellow
- FC 2326—1570 nm Wavelength Specific SFP transceiver, Orange
- FC 2327—1590 nm Wavelength Specific SFP transceiver, Red
- FC 2328—1610 nm Wavelength Specific SFP transceiver, Brown
- FC 5010—1 Gbps I/O (FIO) Module with eight ports
- FC 5210—2Gbps I/O (XFIO2) Module with eight ports
- FC 5220—FCME Upper (Replacement for 1 Gbps FCM)
- FC 5221—FCME Lower (Replacement for 1 Gbps FCM)
- FC 5224—HA Option/256 ports
- FC 5910—FIO Slot Filler
- FC 5924—Non-HA Filler Modules/256 ports with N+1 redundant control, switch and power modules
- FC 8003—Warranty Extension (2 years)
- FC 8013—Warranty Extension (3 years)
- FC 7201—Additional inVSN Enterprise Manager Client license
- FC 7202—FICON Management Server
- FC 7203—FICON Cascading Activation
- FC 7600—inVSN Enterprise Manager
- FC 7603—Additional Ethernet Hub, 8 ports
- FC 7604—Additional Ethernet Hub, 24 ports
- FC 7610—PC Replacement Upgrade for EM Server
- FC 7621—512 MB Memory Card
- FC 7625—Advanced Remote Serviceability Kit

- 2042 Model 128—CNT FC/9000-128 Fibre Channel Director with I/O Modules (24 ports), inVSN Enterprise Manager and one inVSN Enterprise

- 2042 Model 001—CNT FC/9000-64 Fibre Channel Director with I/O Modules (16 ports), inVSN Enterprise Manager and one inVSN Enterprise Client
  - FC 4128—Interconnection kit for 1 Gbps128 port upgrade (non-256 capable)
  - FC 4001—Upgrade Model 001 to 2 Gbps and 256 capable
  - FC 4002—Upgrade Model 001 HA to 2 Gbps and 256 capable
  - FC 4008—Upgrade Model 128 to 2 Gbps and 256 capable

- FC 4009—Upgrade Model 128 HA to 2 Gbps and 256 capable
- FC 4129—Interconnection Kit for 2 Gbps 128 port upgrade
- FC 4256—Interconnection Kit for 256 port upgrade
- FC 5020—High-Availability Option with N+1 redundant control, switch and power modules
- FC 5210—2 Gbps I/O (XFIO2) module with eight ports
- FC 5220—FCME Upper (Replacement for 1 Gbps FCM)
- FC 5221—FCME Lower (Replacement for 1 Gbps FCM)
- FC 5910—FIO Slot Filler
- FC 8001—Warranty Extension (2 years)
- FC 8011—Warranty Extension (3 yea

10

IBM TotalStorage SAN m-type family

On October 5, 2004, IBM announced that it will offer IBM TotalStorage branded SAN switch and director products from McDATA. A new OEM agreement will provide an expanded portfolio of IBM TotalStorage infrastructure simplification and business continuity solutions to customers worldwide. These products will replace McDATA branded products being resold by IBM. This announcement is expected to provide customers more choices and flexibility as they build storage area networks (SANs). IBM has announced:

- IBM TotalStorage SAN12M-1 entry fabric switch (2026-E12); scalable from 4 to 12-ports, with 4-port FlexPort feature, for entry IBM TotalStorage SAN solutions. The SAN12M-1 switch replaces the McDATA 4300 Fabric Switch (2031-212).
- IBM TotalStorage SAN24M-1 midrange fabric switch (2026-224); scalable from 8 to 24-ports, with 8-port FlexPort feature, for midrange IBM TotalStorage SAN solutions. The SAN24M-1 switch replaces the McDATA 4500 Fabric Switch (2031-224 and E24).
- IBM TotalStorage SAN32M-1 enterprise fabric switch (2027-232); scalable from 8 to 32-ports, with affordable four-pack features, for enterprise mainframe FICON and open system IBM TotalStorage SAN solutions. The SAN32M-1 switch replaces the McDATA 3232 Fabric Switch (2031-232).
- IBM TotalStorage SAN140M enterprise director (2027-140), scalable from 16 to 128 ports for large enterprise mainframe FICON and open system IBM

© Copyright IBM Corp. 2004. All rights reserved.

TotalStorage SAN solutions. The SAN140M director replaces the McDATA 6140 Enterprise Director (2032-140) and the McDATA 6064 Enterprise Director (2032-064).

IBM will continue to offer field upgrade features for McDATA products resold by IBM. This includes warranty extension, port-card and future features. Since IBM branded products are technically the same as McDATA branded products, they may be added to existing McDATA switch networks and be managed under McDATA management software. This helps protect customer McDATA switch investments.

The core-to-edge family of connectivity products fully complements these initiatives, allowing users to begin to build a small SAN environment and still be able to expand to a full enterprise-wide SAN.

In this chapter, we cover the full IBM portfolio of McDATA including products currently offered, recently offered, and supported by IBM. Further details of these can be obtained at the following Web site:

http://www-1.ibm.com/servers/storage/san/m_type/

We also cover those McDATA products that are being replaced and these will be referred to by their McDATA designation.

10.1 Product description

The IBM TotalStorage SAN m-type family support e-business and other mission-critical applications that require the highest levels of system availability, including 24x7 business requirements. The directors' high availability features complement the high availability features of the IBM TotalStorage Enterprise Storage Server (ESS). With their FICON switching capabilities, directors and the McDATA Sphereon 3232 Fabric Switch also support IBM 9672 Parallel Enterprise G5, G6 and zSeries Servers with FICON Channel Cards.

10.1.1 Machine type and model number changes

With the OEM agreement, changes have also been made to the machine type and model numbers. In Table 10-1we show the old and new designations:

Table 10-1 Machine type and model number changes

Old	New
2031-212	2026-E12
2031-224	2026-224
2031-232	2027-232
2032-140	2027-140

10.1.2 McDATA Sphereon 4300 Fabric Switch

The McDATA Sphereon 4300 Fabric Switch is the IBM 2026-E12. It is an entry level switch in a 1U high design and offers up to twelve non-blocking longwave or shortwave ports providing 1 and 2 Gb/s Fibre Channel Arbitrated Loop (FC-AL) and Fabric (FC-SW) operation. The switch utilizes auto-sensing and auto-negotiating ports and allows customers to purchase connectivity in four-port increments. The entry version does not support full fabric connectivity, but can be upgraded with a software feature to provide such support. The switch may be non-rack installed (desktop) or installed into an FC-512 Cabinet or an industry standard 19" rack. The power supplies are input rated at 90 to 265 volts alternating current (VAC), at 47-63Hz.

Figure 10-1 McDATA Sphereon 4300 Fabric Switch

Scalability

The entry version consists of four, eight or twelve shortwave ports. Each port is self-configuring as a fabric (F_port) or fabric loop port (FL_port).

The full fabric version additionally supports longwave ports, and the ports will also self-configure as expansion ports (E_port). The longwave SFP transceivers support connections up to 10 km.

The switch provides scalable upgrades, in 4-port increments, without fabric disruption. Each FlexPort upgrade consists of four shortwave SFP transceivers and an activation key which adds four ports to the fabric switch. Longwave transceivers are purchased individually.

Availability

Being an entry level switch, the 2026-E12 is not designed to be as highly available as the 2026-224, and as such only has a single fixed power supply. Three fans are installed, two of which are required for machine operation. Hot-pluggable optical transceivers can be replaced without taking the switch offline. Firmware upgrades can be downloaded and activated while the fabric switch remains operational.

Serviceability

The switch provides the following error detection, reporting, and serviceability features:

- Light-emitting diodes (LEDs) on switch FRUs and adjacent to Fibre Channel ports that provide visual indicators of hardware status or malfunctions.
- A design that enables quick removal and replacement of SFP transceivers without the use of tools or equipment.
- System alerts and logs that display switch, Ethernet link, and Fibre Channel link status at the SANpilot interface.

- Diagnostic software that performs power-on self-tests (POSTs) and port diagnostics (loopback tests).
- An RS-232 maintenance port at the rear of the switch (port access is password protected) that enables installation or service personnel to change the switch's IP address, subnet mask, and gateway address. These parameters can also be changed through a Telnet session, access for which is provided through a local or remote PC with an Internet connection to the switch.
- Data collection through the SANpilot interface application to help isolate system problems. The data includes a memory dump file and audit, hardware, and engineering logs.
- Beaconing to assist service personnel in locating a specific port or switch. When port beaconing is enabled, the amber LED associated with the port flashes. When unit beaconing is enabled, the system error indicator on the front panel flashes. Beaconing does not affect port or switch operation.

Restriction: The 2026-E12 does not provide an Element Manager feature, and hence cannot be fully managed by the EFCM application. Launching the 2026-E12 from EFCM will result in the SANpilot web interface opening.

10.1.3 McDATA Sphereon 4500 Fabric Switch

The McDATA Sphereon 4500 Fabric Switch is the IBM 2026- 224. It provides storage consolidation using a high-port density 1U high design, ports for longwave and shortwave transceivers, offers up to twenty-four non-blocking ports providing 1 and 2 Gb/s Fibre Channel Arbitrated Loop (FC-AL) and Fabric (FC-SW) operation. The switch utilizes auto-sensing and auto-negotiating ports, allows customers to purchase connectivity in eight-port increments, and provides integrated support for full fabric and FC-AL tape attachment to core fabric switches and directors. The switch may be non-rack installed (desktop) or installed into an FC-512 Cabinet or an industry standard 19" rack. The power supplies are input rated at 100 to 240 volts alternating current (VAC), at 47-63Hz.

Figure 10-2 McDATA Sphereon 4500 Fabric Switch

Scalability

The switch versions include an entry 8-port, a midrange 16-port and enterprise 24-port edge switch. The entry switch version consists of eight shortwave ports. Each port is self-configuring as a fabric, fabric loop or expansion port. Longwave SFP transceivers may be added to the first four ports for 2 Gb/s connections up to 10 km. The switch provides scalable upgrades, in 8-port increments, without fabric disruption. Each FlexPort upgrade consists of eight shortwave SFP transceivers and an activation key which adds eight ports to the fabric switch.

Availability

The 2026-224 provides hot-swappable, load-sharing dual power supplies that allow the switch to remain online if one supply fails. Dual power cords enable attachment to separate power sources for improved availability. Hot-swappable power and cooling components eliminate downtime for service when replacing a failed component and eliminates the risk of erroneously cabling a replacement switch because of a simple component failure. Failed power supplies and fans can be replaced without special tools. Hot-pluggable optical transceivers can be replaced without taking the switch offline. Firmware upgrades can be downloaded and activated while the fabric switch remains operational.

Serviceability

The switch provides the following error detection, reporting, and serviceability features:

- Light-emitting diodes (LEDs) on switch FRUs and adjacent to Fibre Channel ports that provide visual indicators of hardware status or malfunctions.
- Redundant FRUs (SFP transceivers and integrated cooling fan and power supply assemblies) that are removed or replaced without disrupting switch or Fibre Channel link operation.
- A modular design that enables quick removal and replacement of FRUs without the use of tools or equipment.

- System alerts and logs that display switch, Ethernet link, and Fibre Channel link status at the SANpilot interface, EFC Server, customer-supplied server (running the EFCM Lite application), or remote workstation.
- Diagnostic software that performs power-on self-tests (POSTs) and port diagnostics (loopback tests).
- An RS-232 maintenance port at the rear of the switch (port access is password protected) that enables installation or service personnel to change the switch's IP address, subnet mask, and gateway address. These parameters can also be changed through a Telnet session, access for which is provided through a local or remote PC with an Internet connection to the switch.
- Data collection through the SANpilot interface or Element Manager application to help isolate system problems. The data includes a memory dump file and audit, hardware, and engineering logs.
- Beaconing to assist service personnel in locating a specific port or switch. When port beaconing is enabled, the amber LED associated with the port flashes. When unit beaconing is enabled, the system error indicator on the front panel flashes. Beaconing does not affect port or switch operation.
- An external modem for use by support personnel to dial-in to the EFC Server (optional) for event notification and to perform remote diagnostics.

10.1.4 McDATA Sphereon 3232 Fabric Switch

The McDATA Sphereon 3232 Fabric Switch, which is the IBM 2027-232, is a 2 Gb/s Fabric Switch intended for departmental Fibre Channel SAN applications and connections to SAN backbones utilizing directors. The switch shown in Figure 10-3 is 1.5U high and can be mounted in the FC-512 Cabinet, an IBM 2101 or 7014 rack, an industry standard 19" rack, or used in a stand-alone table-top configuration. The power supplies are input rated at 100 to 240 volts alternating current (VAC), at 47-63Hz.

Figure 10-3 McDATA Sphereon 3232 Fabric Switch switch

Scalability

Each fabric switch is capable of providing up to 32 ports of non-blocking Fibre Channel switching capability, featuring hot-pluggable SFP transceivers. The switch ships with redundant hot-swappable power supplies and cooling units.

The minimum configuration contains 16 shortwave transceivers. You can add up to sixteen additional transceivers, either shortwave or longwave, for device interconnection up to 10km using the longwave transceiver. Extended distance longwave transceivers are available for interconnection up to 35 km.

Generic ports (G_Port) automatically determine the port type when connected to a node port (N_Port) or an expansion port (E_Port). Any port can function as an F_Port when connected to a device or as an E_Port when connected to another switch. This switch does not support direct connection of arbitrated loop devices. If you plan to use arbitrated loop, it is recommended that you consider the 2026-224 switch.

Availability features

The switch is initialized, configured and controlled by a control processor (CTP) card. The CTP card contains microprocessor and an application specific integrated circuit (ASIC) subsystem that provides port communication functions and enables frame transmission between switch ports without software intervention.

The CTP card also provides non-volatile memory for storing firmware (two memory regions to be able to store two firmware versions), switch configuration information, persistent operating parameters and memory dump files.

There is also a 10/100 Mb/s Ethernet port and an RS-232 maintenance port controlled by the CTP card.

Note: The CTP is not a FRU and if it fails the entire switch must be replaced.

The features that ensure high availability for the switches are covered in the following topics.

Power supplies

Two redundant power supplies share the operating load. If one supply fails, the other supply handles the full load. The failed power supply can be replaced concurrently. There are separate receptacles at the rear of the switch for input power connection. For full redundancy each input should come from a different power source.

Fans

The switches have six fans. Two on each power supply and two in the center section of the switch. If a single fan fails, the redundant fans provide cooling until it is replaced. If two or more fan fails they must be replaced immediately.

Spare ports

Unused ports can be used as spare ports. In case of a port failure the cable can be moved to a spare port to continue switch operation. Care should be taken when zoning is configured specifying port numbers since any affected zone(s) may need to be re-configured. Depending on the operating system, the path may need to be re-configured to be able to continue operation on a new port.

Concurrent firmware upgrade

The CTP card provides two nonvolatile memory regions for storing firmware. Storing two firmware versions allow firmware upgrades to be performed concurrently without disrupting switch operation. This includes non-disruptive activation of the new code.

Serviceability

The switch provides the following error detection, reporting, and serviceability features:

- LEDs on switch FRUs and next to each Fibre Channel port that provide visual indication of status or failures
- System alerts that display at the EFC Server or a remote workstation connected to it
- Event logs, audit logs, link incident logs, and hardware logs
- Diagnostic software that performs power on self tests (POSTs) and port diagnostics, including internal and external loopback wrap tests
- Automatic notification to support personnel or administrators by e-mail messages
- Automatic notification to service support center by the call home feature
- Dial-in capabilities for use by service personnel to monitor or perform remote diagnostics
- RS-232 maintenance port that is password protected and allows service personnel to change the switch network address
- Redundant FRUs (power supplies and fans) that are removed and replaced without affecting switch operations; no special tools needed to remove and replace FRUs

- SFP transceivers that are removed and replaced without affecting operation of other ports
- Beaconing for quick identification of a switch or specific port by a flashing LED without affecting operation
- Data collection through the Element Manager application to help isolate problems
- Unsolicited SNMP trap messages indicating operational state changes and failure conditions sent to authorized workstations

10.1.5 McDATA Intrepid 6140 Director

The McDATA Intrepid 6140 Director is the IBM 2027-140. The director is a 140-port product that provides dynamic switched connections between Fibre Channel servers and devices in a SAN environment. It is 12U high, so up to three can be configured in an FC-512 Fabricenter equipment cabinet, which can provide up to 420 ports in a single cabinet.

The McDATA Intrepid 6140 Director, shown in Figure 10-4, provides 140-port, 2 Gb/s, high availability switching and enterprise-level scalability for data center class core/edge fabrics, and long transmission distances (up to 35 km, or up to 100 km with repeaters).

.

Figure 10-4 McDATA Intrepid 6140 Director

Scalability

Each director comes with a minimum of four 4-port UPM (Universal Port Modules) consisting of 16 G_Ports. The McDATA Intrepid 6140 Director is capable of supporting from 16 up to 140 ports by adding additional UPMs.

The ability to support a number of different port types aids in building a scalable environment. A G_Port is a generic port that can function as either an F_Port or an E_Port. When the director is connected with an N_Port (node device), the G_Port state changes to an F_Port (fabric port). When a G_Port is interconnected with another director, the port state on each director changes to an E_Port. E_Ports are used for Inter-Switch Link (ISL) connections.

An arbitrated loop topology connects multiple device node loop (NL_Ports) in a loop (or hub) configuration without benefit of a multi-switch fabric. Although the director does not support direct connection of arbitrated loop devices, such devices can communicate with the director via an interconnect with the McDATA Sphereon 4500 Fabric Switch.

For shortwave ports, the maximum distance is 500 m at 1 Gb/s and 300 m at 2 Gb/s using 50 micron fiber. For longwave ports the maximum distance to a device is 20 km at 1 Gb/s and 10 km at 2 Gb/s using 9 micron fiber. Using longwave ports and four repeaters spaced 20 km each, distances of up to 100 km can be reached. There is an extended distance option that can be configured on a port basis by port basis. The extended distance option is used to assign additional buffers (60) to the specified port in order to support operation at distances of up to 100 km using repeaters.

Connectivity

The director contains ports at the front and the rear of the director. The ports on the front are numbered from 0-127 and continue in the rear from 132-143. Ports 128-131 are not available ports.

In Figure 10-5 we show the numbering scheme of UPM cards and the associated fiber ports for the front of the director. On the bottom, the port count starts at the right most UPM and goes from the top to the bottom on each UPM. On the top, the port count continues from the right most UPM but the count now starts from the bottom to the top of each UPM. This is because the cards on the top are physically installed upside-down compared to the bottom cards.

Tip: The large, bold, hexadecimal numbers are the Link Port Addresses used for FICON IOCP configurations on zSeries processors.

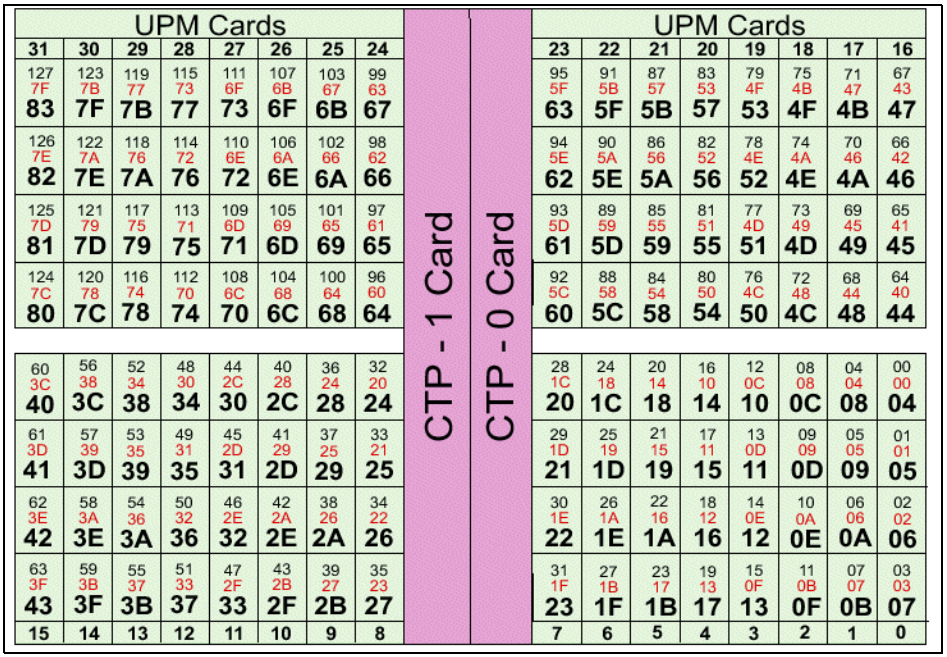


Figure 10-5 ED-6140 port map (front)

In Figure 10-6 we show the numbering scheme for the rear ports. This scheme is slightly different. On the bottom left UPM, the ports count from right to left; the next sequential UPM is on the top right card, where the ports count from left to right; and finally, the top left card, where the ports count from right to left.

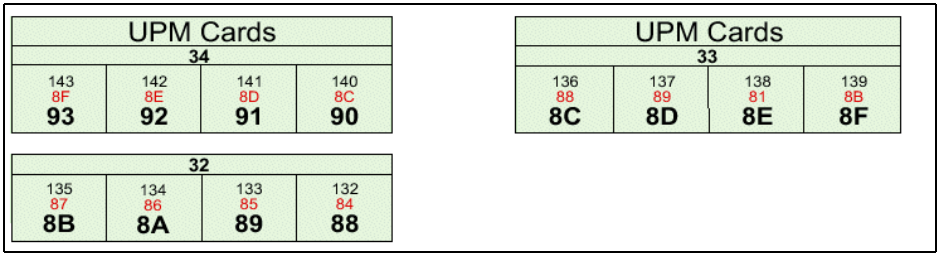


Figure 10-6 ED-6140 port map (rear)

For availability purposes it is recommended that you spread your storage ports across multiple cards. Servers with multiple HBAs connected to the director should also be connected to ports spread across multiple cards, as should any ISLs to another director or switch. In the event of a UPM card failure, only a single link to a given storage device or server will be impacted, which will minimize any performance degradation.

Availability

Pairs of critical field replaceable units (FRUs) installed in the director provide redundancy in case a FRU fails. When an active FRU fails, the backup FRU takes over operation automatically (by failover processing) to maintain director and Fibre Channel link operation.

A standard availability director has all possible FRUs installed and is fully redundant. Standard redundancy is provided through dual sets of FRUs and spare (unused) ports on UPM cards The director offers excellent redundancy and maintenance capabilities such as:

- All active components are redundant
- Active components provide support for automatic failover
- Redundant power and cooling
- Hot swapping of all field replaceable units
- Automatic fault detection and isolation
- Non-disruptive firmware updates

The director provides a modular design that enables quick removal and replacement of components.

Backplane

The backplane provides 48 VDC power distribution and connections for all logic cards. The backplane is a non-concurrent FRU. The director must be powered off prior to FRU removal and replacement.

CTP2 card

The director is delivered with two CTP2 cards. The active CTP2 card initializes and configures the director after power on and contains the microprocessor and associated logic that coordinate director operation. A CTP2 card provides an initial machine load (IML) button on the faceplate. When the button is pressed and held for three seconds, the director reloads firmware and resets the CTP2 card without switching off power or affecting operational fiber-optic links.

Each CTP2 card also provides a 10/100 Mb/s RJ-45 twisted pair connector on the faceplate that attaches to an Ethernet local area network (LAN) to communicate with the EFC Server or a simple network management protocol (SNMP) management station. During an IML, this Ethernet connection will also drop.

Each CTP2 card provides system services processor (SSP) and embedded port (EP) subsystems. The SSP subsystem runs director applications and the underlying operating system, communicates with director ports, and controls the RS-232 maintenance port and 10/100 Mb/s Ethernet port. The EP subsystem provides Class F and exception frame processing, and manages frame

transmission to and from the SBAR assembly. In addition, a CTP2 card provides non-volatile memory for storing firmware, director configuration information, persistent operating parameters, and memory dump files. Director firmware is upgradable concurrently (without disrupting operation).

The backup CTP2 card takes over operation if the active card fails. Failover from a faulty card to the backup card is transparent to attached devices, and includes transferal of the TCP/IP address using the same media access control (MAC) address as the original interface.

Each card faceplate contains a green LED that illuminates if the card is operational and active, and an amber LED that illuminates if the card fails. Both LEDs are extinguished on an operational backup card. The amber LED blinks if FRU beaconing is enabled.

UPM card

A UPM card is a concurrent FRU and can be added or replaced while the director is powered on and operating. Each UPM card provides four full-duplex generic ports (G_Ports) that transmit and receive data at 2 Gb/s or 1 Gb/s. UPM cards use non-open fiber control (OFC) Class 1 laser transceivers.

Port cards do not automatically failover and continue link operation after a port card failure. To continue device operation, the fiber optic cable from a failed port must be physically moved to an unused operational port. Hence it is advisable to reserve sufficient spare ports in a director to allow for this possibility. When a cable is moved, additional SAN configuration may be necessary for continued data availability.

In Figure 10-7, we show a front view of the director containing the CTP2 and UPM cards.

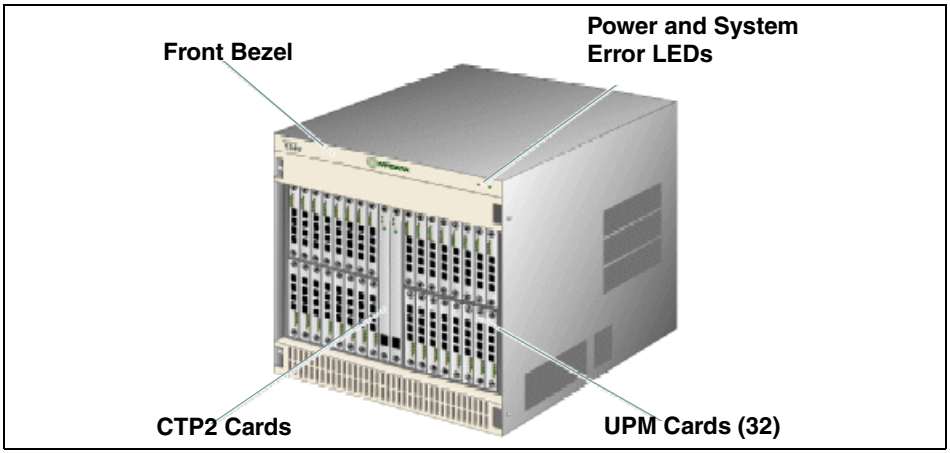


Figure 10-7 McDATA Intrepid 6140 Director hardware (front view)

Fan module

Three fan modules, each containing one system fan (three system fans total), provide cooling for director FRUs, as well as redundancy for continued operation if a fan fails. A fan module can be replaced while the director is powered on and operating, provided the module is replaced within ten minutes (after which software powers off the director). An amber LED for reach fan module illuminates if one or more fans fail or rotate at insufficient angular velocity.

Power supply modules

The McDATA Intrepid 6140 Director contains two redundant, load-sharing power supply modules which are installed in slot positions 1 and 0 (left to right). They provide 48-volt direct current (VDC) power to the director FRUs. The power supplies also provide over-voltage and over-current protection. Either power supply can be replaced while the director is powered on and operational. Each power supply has a separate backplane connection to allow for different AC power sources which is recommended for full power redundancy. The power supplies are input rated at 200 to 240 volts alternating current (VAC), at 47-63Hz.

AC module

The alternating current (AC) module is located at the bottom rear of the director. Either AC module can be replaced while the director is powered on and operational. The module provides:

- Two single-phase, 220 VAC, power connectors.
- An input filter and AC system harness (internal to the FRU) that provides the wiring to connect the AC power connectors to the power supplies (through the backplane).

SBAR assembly

The director is delivered with two serial crossbars (SBAR) assemblies. The active SBAR is responsible for Fibre Channel frame transmission from any director port to any other director port. Connections are established without software intervention. The assembly accepts a connection request from a port, determines if a connection can be established, and establishes the connection if the destination port is available. The assembly also stores busy, source connection, and error status for each director port.

The backup SBAR takes over operation if the active assembly fails, and provides the ability to maintain connectivity and data frame transmission without interruption. Failover to the backup assembly is transparent to attached devices.

Each SBAR assembly consists of a card and steel carriage that mounts flush on the backplane. The carriage provides protection for the back of the card, distributes cooling airflow, and assists in aligning the assembly during installation. The rear of the carriage contains a green LED that illuminates if the assembly is operational and active, and an amber LED that illuminates if the assembly fails. Both LEDs are extinguished on an operational backup assembly. The amber LED blinks if FRU beaconing is enabled.

In Figure 10-8 we show a rear view of the director, including three additional UPM cards.

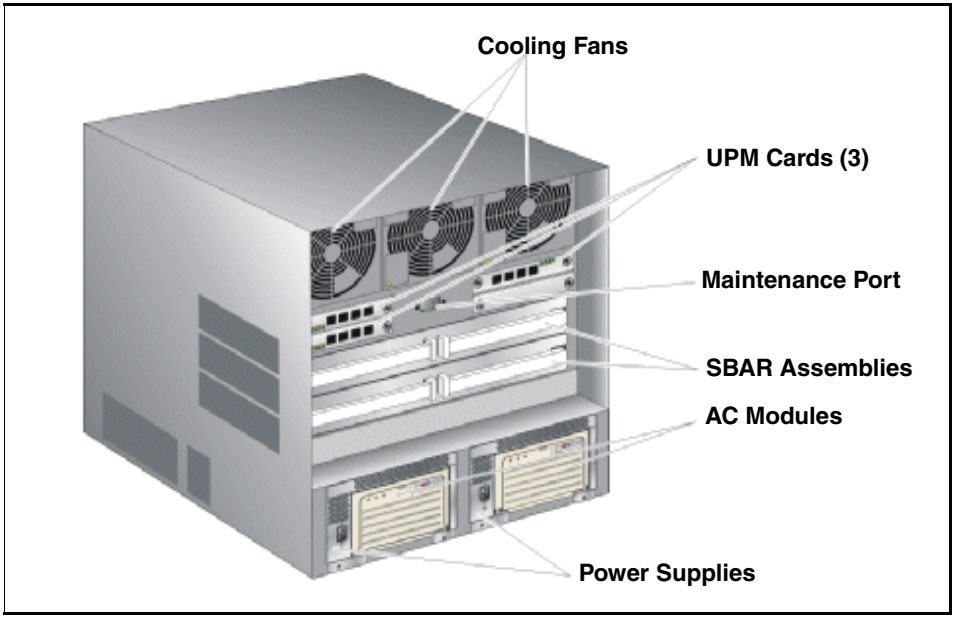


Figure 10-8 McDATA Intrepid 6140 Director hardware (rear view)

Serviceability

The director, with its associated software and hardware, provides the following error detection, reporting, and serviceability features:

- Light-emitting diodes (LEDs) on director FRUs and the front bezel that provide visual indicators of hardware status or malfunctions.
- System and threshold alerts, and event logs, audit logs, link incident logs, threshold alert logs, and hardware logs that display director, Ethernet link, and Fibre Channel link status at the EFC Server, customer-supplied server (running the EFCM Lite application), or remote workstation.
- Diagnostic software that performs power-on self-tests (POSTs) and port diagnostics (internal loopback, external loopback, and Fibre Channel (FC) wrap tests). The FC wrap test applies only when the director is configured to operate in S/390 mode.
- Automatic notification of significant system events (to support personnel or administrators) through e-mail messages or the call-home feature. The call-home feature may not be available if the EFC Manager application (EFCM Lite) is installed on a customer-supplied PC.
- An external modem for use by support personnel to dial-in to the EFC Server for event notification and to perform remote diagnostics.
- An RS-232 maintenance port at the rear of the director (port access is password protected) that enables installation or service personnel to change the director's internet protocol (IP) address, subnet mask, and gateway address; or to run diagnostics and isolate system problems through a local or remote terminal.
- Redundant FRUs (logic cards, power supplies, and cooling fans) that are removed or replaced without disrupting director or Fibre Channel link operation.
- A modular design that enables quick removal and replacement of FRUs without the use of special tools or equipment.
- Concurrent port maintenance. UPM cards are added or replaced and fiber-optic cables are attached to ports without interrupting other ports or director operation.
- Beaconing to assist service personnel in locating a specific port, FRU, or director in a multi-switch environment. When port beaconing is enabled, the amber LED associated with the port flashes. When FRU beaconing is enabled, the amber (service required) LED on the FRU flashes. When unit beaconing is enabled, the system error indicator on the front bezel flashes. Beaconing does not affect port, FRU, or director operation.

- Data collection through the Element Manager application to help isolate system problems. The data includes a memory dump file and audit, hardware, and engineering logs.
- Status monitoring of redundant FRUs and alternate Fibre Channel data paths to ensure continued director availability in case of failover. The EFC Manager application queries the status of each backup FRU daily. A backup FRU failure is indicated by an illuminated amber LED.

10.1.6 McDATA Intrepid 6064 Director

The McDATA Intrepid 6064 Director is the IBM 2032 Model 064 (2032-064). The ED-6064 director is a 64-port product that provides dynamic switched connections between Fibre Channel servers and devices in a SAN environment. It is 9U high, so up to four ED-6064 directors can be installed in an FC-512 equipment cabinet, which can provide up to 256 ports in a single cabinet. The director implements Fibre Channel technology that provides a scalable bandwidth (2 Gb/s), redundant switched data paths, and long transmission distances (up to 35 km, or up to 100 km with repeaters).

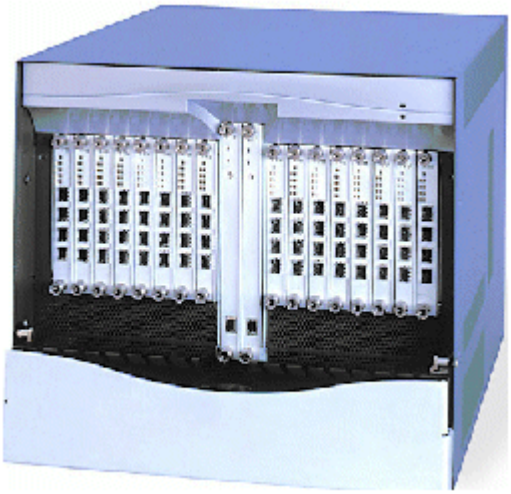


Figure 10-9 McDATA Intrepid 6064 Director

Scalability

The McDATA Intrepid 6064 Director is capable of supporting from 16 up to 64 ports by adding additional UPM cards. Each director comes with a minimum of four 4-port UPMs (Universal Port Modules), consisting of 16 G_Ports.

The ability to support a number of different port types aids in building a scalable environment. A G_Port is a generic port that can function as either an F_Port or an E_Port. When the director is connected with an N_Port (node device), the G_Port state changes to an F_Port (fabric port). When a G_Port is interconnected with another director or switch, the port state changes to an E_Port. E_Ports are used for Inter-Switch Link (ISL) connections.

An arbitrated loop topology connects multiple device node loop (NL_Ports) in a loop configuration without benefit of a multi-switch fabric. Although the director does not support direct connection of arbitrated loop devices, such devices can communicate with the director via an interconnect with the McDATA Sphereon 4500 Fabric Switch.

For shortwave ports the maximum distance is 500 m at 1 Gb/s and 300 m at 2 Gb/s using 50 micron fiber. For longwave ports the maximum distance to a device is 20 km at 1 Gb/s and 10 km at 2 Gb/s using 9 micron fiber. Using longwave laser ports and four repeaters spaced 20 km each, distances of up to 100 km can be reached. There is an extended distance option that can be configured on a port by port basis. The extended distance option is used to assign additional buffers (60) to the specified port in order to support operation at distances of up to 100 km.

Note: To obtain 2 Gb/s port speeds, all port cards must be UPMs and both CTPs must be CTP2s. If any port cards are FPMs (as shipped with early ED-6064s), then the whole director will run at 1 Gb/s.

Connectivity

In Figure 10-10 we show the numbering scheme of UPM cards and associated fiber ports. UPM cards are numbered from right to left and ports on each UPM count from bottom to top.

UPM Cards										UPM Cards							
15	14	13	12	11	10	9	8	CTP2 - 1 Card	CTP2 - 0 Card	7	6	5	4	3	2	1	0
63	59	55	51	47	43	39	35			31	27	23	19	15	11	07	03
62	58	54	50	46	42	38	34			30	26	22	18	14	10	06	02
61	57	53	49	45	41	37	33			29	25	21	17	13	09	05	01
60	56	52	48	44	40	36	32			28	24	20	16	12	08	04	00

Figure 10-10 ED-6064 port map

For availability purposes it is recommended that you spread your storage ports across multiple cards. Servers with multiple HBAs connected to the same director should also be connected to ports spread across multiple cards, as should any ISLs to another director or switch. In the event of a UPM card failure, only a single link to a given storage device or server will be impacted, which will minimize any performance degradation.

Availability

Pairs of critical field replaceable units (FRUs) installed in the director provide redundancy in case a FRU fails. When an active FRU fails, the backup FRU takes over operation automatically (failover) to maintain director and Fibre Channel link operation.

A standard availability director has all possible FRUs installed and is fully redundant. Standard redundancy is provided through dual sets of FRUs and spare (unused) ports on UPM cards. The McDATA director offers excellent redundancy and maintenance capabilities, such as these:

- All active components are redundant
- Active components provide support for automatic failover
- Redundant power and cooling
- Hot swapping of all field replaceable units
- Automatic fault detection and isolation
- Non-disruptive firmware updates

The director provides a modular design that enables quick removal and replacement of components.

Backplane

The backplane provides 48 VDC power distribution and connections for all logic cards. The backplane is a non-concurrent FRU. The director must be powered off prior to FRU removal and replacement. The backplane in the director supports 2 Gb/s operation.

CTP2 cards

The director is delivered with two CTP2 cards. The active CTP2 card initializes and configures the director after power on and contains the microprocessor and associated logic that coordinate director operation. A CTP2 card provides an initial machine load (IML) button on the faceplate. When the button is pressed and held for three seconds, the director reloads firmware and resets the CTP2 card without switching off power or affecting operational fiber-optic links. Each CTP2 card also provides a 10/100 Mb/s RJ-45 twisted pair connector on the faceplate that attaches to an Ethernet local area network (LAN) to communicate

with the EFC Server or a simple network management protocol (SNMP) management station. During an IML, this Ethernet connection will also drop.

Each CTP2 card provides system services processor (SSP) and embedded port (EP) subsystems. The SSP subsystem runs director applications and the underlying operating system, communicates with director ports, and controls the RS-232 maintenance port and 10/100 Mb/s Ethernet port. The EP subsystem provides Class F and exception frame processing, and manages frame transmission to and from the SBAR assembly. In addition, a CTP2 card provides nonvolatile memory for storing firmware, director configuration information, persistent operating parameters, and memory dump files. Director firmware is upgradable concurrently (without disrupting operation).

The backup CTP2 card takes over operation if the active card fails. Failover from a faulty card to the backup card is transparent to attached devices, and includes transferal of the TCP/IP address using the same media access control (MAC) address as the original interface. Each card faceplate contains a green LED that illuminates if the card is operational and active, and an amber LED that illuminates if the card fails. Both LEDs are extinguished on an operational backup card. The amber LED blinks if FRU beaconing is enabled.

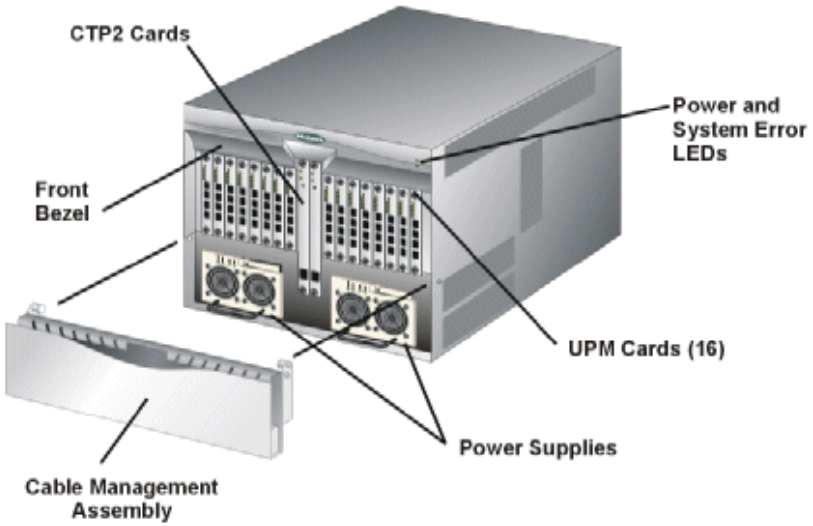


Figure 10-11 McDATA Intrepid 6064 Director hardware (front view)

UPM cards

A UPM card is a concurrent FRU and can be added or replaced while the director is powered on and operating. Each UPM card provides four full-duplex generic

ports (G_Ports) that transmit and receive data at 2 Gb/s or 1 Gb/s. UPM cards use non-open fiber control (OFC) Class 1 laser transceivers.

Port cards do not automatically failover and continue link operation after a port card failure. To continue device operation, the fiber optic cable from a failed port must be physically moved to an unused operational port. Hence it is advisable to reserve sufficient spare ports in a director to allow for this possibility. When a cable is moved, additional SAN configuration may be necessary for continued data availability.

Fan modules

Two fan modules, each containing three fans (six fans total), provide cooling for director FRUs, as well as redundancy for continued operation if a fan fails. A fan module can be replaced while the director is powered on and operating, provided the module is replaced within ten minutes (after which software powers off the director). An amber LED for each fan module illuminates if one or more fans fail or rotate insufficiently.

Power supply module

Redundant, load-sharing power supplies step down and rectify input power to provide 48-volt direct current (VDC) power to director FRUs. The power supplies also provide over-voltage and over-current protection. Either power supply can be replaced while the director is powered on and operational. Each power supply has a separate backplane connection to allow for different alternating current (AC) power sources. The power supplies are input rated at 100 to 240 volts alternating current (VAC), at 47-63Hz.

Power module assembly

The power module assembly is located at the bottom rear of the director. The module is a non-concurrent FRU, and the director must be powered off prior to scheduled removal and replacement. The module provides:

- Two single-phase AC power connectors. Each connector is input rated at 100 to 240 VAC.
- A power switch (circuit breaker) that controls AC power distribution to both power supplies. The breaker is set manually, or is automatically tripped by internal software if thermal sensors indicate the director is overheated.
- A 9-pin maintenance port that provides a connection for a local terminal or dial-in connection for a remote terminal. Although the port is typically used by maintenance personnel, operations personnel use the port to configure network addresses.

- An input filter and AC system harness (internal to the FRU) that provides the wiring to connect the AC power connectors to the power switch and power supplies (through the backplane).

Figure 10-12 shows a rear view of the ED-6064.

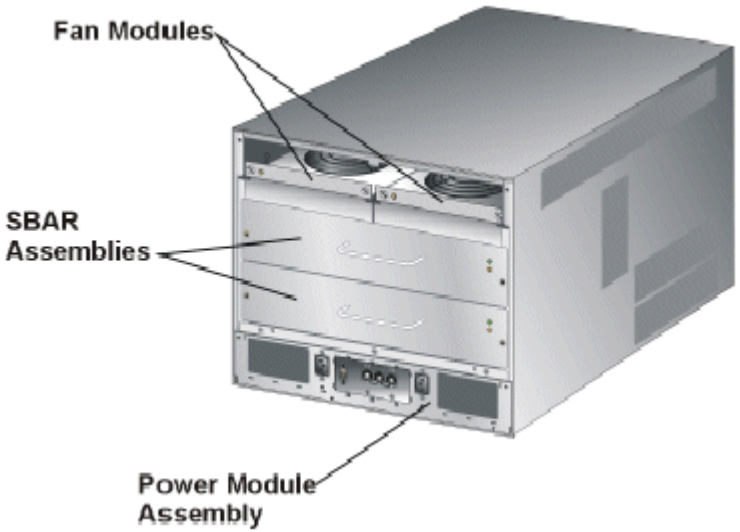


Figure 10-12 McDATA Intrepid 6064 Director hardware (rear view)

SBAR assemblies

The director contains two serial crossbar (SBAR) assemblies. Each SBAR card is responsible for Fibre Channel frame transmission from any director port to any other director port. Connections are established without software intervention. The card accepts a connection request from a port, determines if a connection can be established, and establishes the connection if the destination port is available. The card also stores busy, source connection, and error status for each director port.

The redundant serial SBAR assembly ensures uninterrupted transmission and receipt of Fibre Channel frames between ports if the active SBAR card fails. Failover to the backup card is transparent to attached devices.

RFI Shield

The RFI shield covers and provides RFI protection for all rear-access FRUs except the power module assembly. The RFI shield is concurrent and can be removed or replaced while the director is powered on and operating.

Serviceability

The director is configured with reporting, and serviceability features. The director provides the following error detection, reporting, and serviceability features:

- Light-emitting diodes (LEDs) on director FRUs and the front bezel that provide visual indicators of hardware status or malfunctions.
- System and threshold alerts, and event logs, audit logs, link incident logs, threshold alert logs, and hardware logs that display director, Ethernet link, and Fibre Channel link status at the EFC Server, customer-supplied server (running the EFCM Lite application), or remote workstation.
- Diagnostic software that performs power-on self-tests (POSTs) and port diagnostics (internal loopback, external loopback, and FC wrap tests). The FC wrap test applies only when the director is configured to operate in S/390 mode.
- Automatic notification of significant system events (to support personnel or administrators) through e-mail messages or the call-home feature. The call-home feature may not be available if the EFC Manager application (EFCM Lite) is installed on a customer-supplied PC.
- An external modem for use by support personnel to dial-in to the EFC Server for event notification and to perform remote diagnostics.
- An RS-232 maintenance port at the rear of the director (port access is password protected) that enables installation or service personnel to change the director's internet protocol (IP) address, subnet mask, and gateway address; or to run diagnostics and isolate system problems through a local or remote terminal.
- Redundant FRUs (logic cards, power supplies, and cooling fans) that are removed or replaced without disrupting director or Fibre Channel link operation.
- A modular design that enables quick removal and replacement of FRUs without the use of special tools or equipment.
- Concurrent port maintenance. UPM cards are added or replaced and fiber-optic cables are attached to ports without interrupting other ports or director operation.
- Beaconing to assist service personnel in locating a specific port, FRU, or director in a multi-switch environment. When port beaconing is enabled, the amber LED associated with the port flashes. When FRU beaconing is enabled, the amber (service required) LED on the FRU flashes. When unit beaconing is enabled, the system error indicator on the front bezel flashes. Beaconing does not affect port, FRU, or director operation.

- Data collection through the Element Manager application to help isolate system problems. The data includes a memory dump file and audit, hardware, and engineering logs.
- Status monitoring of redundant FRUs and alternate Fibre Channel data paths to ensure continued director availability in case of failover. The EFC Manager application queries the status of each backup FRU daily. A backup FRU failure is indicated by an illuminated amber LED.

10.2 Fabric planning

In this section we discuss some of the considerations that are important to ensure maximum availability.

10.2.1 Dual fabrics and directors

One of the first points to consider is that although the director is a highly available device, for maximum protection we recommend splitting the connections into two separate directors. Each director should be in a separate fabric, and all servers and storage devices connected to each fabric. In the event of a fabric or director failure each device, server, or critical application would still have access through the alternate fabric.

10.2.2 Server to storage ratio

Server-to-storage ratio is another important point that needs to be carefully considered. The output of most host devices is bursty in nature; most devices do not sustain full-bandwidth output, and it is uncommon for the output of multiple devices to peak simultaneously. These variations are why multiple hosts can be serviced by a single storage port.

This device sharing leads to the concept of fan-out ratio. Device fan-out ratio is defined as the storage or array port IOPS divided by the attached host IOPS, rounded down to the nearest whole number. A more simplistic definition for device fan out is the ratio of host ports to a single storage port. Fan-out ratios are typically device dependent.

In general, the maximum practical device fan-out ratio is 12 to 1. If there is not enough information to know the estimated throughput requirements, an initial setting that works well is a ratio of 6 server connections per storage connection for high performance profile servers (usually UNIX boxes), and a 12:1 ratio for low performance profile servers (usually Windows) so as not to overload the box with storage connections.

10.2.3 ISLs

When cascading directors, it is important to consider the number of ISLs in order to avoid oversubscription, which may result in performance degradation. When multiple hops are required, the directors assign the shortest path between two devices (minimum hop counts), also known as Fabric Shortest Path First (FSPF).

We should consider the traffic patterns between dual directors and the need for alternate paths in case of link failures, and then configure enough ISLs to allow the required bandwidth according to our performance expectations. Each ISL will increase the bandwidth for traffic between directors, but reduces the number of ports available for device connections and also introduces blocking between the switches.

When multiple ISLs are available, the director will try to balance the load between them by assigning the same number of connections on each path. Each path is assigned when a device logs into the fabric. When there are multiple paths with the minimum number of hop counts (minimum cost) to a device the director tries to assign them to different ISLs. This minimizes the possibility of a single ISL affecting all paths to a device.

10.2.4 Load balancing

We must also take into account that load balancing is a static balance. Each path is determined when a device logs into the fabric. The same number of ISL connections may not result in the same traffic later on. There may be some ISLs oversubscribed, while other ISLs may have unused bandwidth available. Connections are assigned as devices log in to the fabric and it is an automatic process without manual intervention or configuration (unless The Preferred Path Option or Open Trunking features are installed). The only possible interaction is to power on or connect the devices in a given sequence, but this does not provide for consistent results in the case of a re-initialization with all the devices up.

One way to force a re-configuration is to add an additional ISL. When a new ISL is detected, the path calculation and load balancing calculation is performed again. The re-configuration is done within the time-out periods so it should not impact current traffic. This could be a way to solve an ISL oversubscription problem. One way to have a device select a different ISL would be to reboot that device or disconnect/reconnect it from the fabric. This way there is a chance another ISL would be selected. However, there is no guarantee, and it is possible that the same ISL would be selected.

Note: Careful consideration should be given to those devices affected by heavy workload and critical applications. If at all possible the connections to these devices should be routed directly through a director without going through ISLs.

10.2.5 Principal switch selection

This value determines the principal switch for the multi-switch fabric. Select either **Principal** (highest priority), **Default**, or **Never Principal** (lowest priority) from the Switch Priority drop-down list. If all fabric elements are set to Principal or Default, the director or switch with the highest priority and the lowest WWN becomes the principal switch. The switch must be offline to change this setting.

The following are some examples of principal switch selection:

- If you have three fabric elements and set all to Default, the director or switch with the lowest WWN become the principal switch.
- If you have three fabric elements and set two to Principal and one to Default, the element with the Principal setting that has the lowest WWN becomes the principal switch.
- If you have three fabric elements and set two to Default and one to Never Principal, the element with the Default setting and the lowest WWN becomes the principal switch.

Note: Note that at least one director or switch in a multi-switch fabric needs to be set as Principal or Default. If all the fabric elements are set to Never Principal, all ISLs will segment. If all but one element is set to Never Principal and the element that was Principal goes offline, then all of the other ISLs will segment. It is recommended to configure the switch priority as Default, although if you have multiple directors and switches in the same fabric, you may wish to favour the directors over the switches.

Domain ID assignment

Each director or switch in a multi-switch fabric is identified by a unique domain ID that ranges between 1 and 31. A domain ID of 0 is invalid. Domain IDs are used in 24-bit Fibre Channel addresses that uniquely identify source and destination ports in a fabric. The switch must be offline to change this setting.

Note: Although 31 domain IDs are available, the maximum tested and supported in a single fabric by the McDATA and IBM agreement is 24.

Each fabric element is configured through the Element Manager application with a preferred domain ID. When a director or switch powers on and comes online, it requests a domain ID from the fabric's principal switch (indicating its preferred value as part of the request).

- If the requested domain ID is not allocated in the fabric, the domain ID is assigned to the requesting director or switch.
- If the requested domain ID is already allocated, an unused domain ID is assigned, unless the Insistent box is checked, in which case the switch will segment from the rest of the fabric.

If two operational fabrics join, they determine if any domain ID conflicts exist between the fabrics. If one or more conflicts exist, the interconnecting ISL E_Ports segment to prevent the fabrics from joining. To prevent this problem, it is recommended that all directors and switches be assigned a unique preferred domain ID. This is particularly important if zoning is implemented through port number (and by default domain ID) rather than by WWN.

Frame delivery order

When directors or switches calculate a new least-cost data transfer path through a fabric, routing tables immediately implement that path. This may result in Fibre Channel frames being delivered to a destination device out of order, because frames transmitted over the new (shorter) path may arrive ahead of previously transmitted frames that traverse the old (longer) path. This can cause problems because many Fibre Channel devices cannot receive frames in the incorrect order.

A rerouting delay parameter can be enabled at the Element Manager application to ensure the director or switch provides correct frame order delivery. The delay period is equal to the error detect time-out value (E_D_TOV) specified in the Element Manager application. Class 2 frames transmitted into the fabric during this delay period are rejected; Class 3 frames are discarded without notification. By default, the rerouting delay parameter is disabled.

Note: To prevent E_Port segmentation, the same error detect time-out value E_D_TOV and resource allocation time-out value (R_A_TOV) must be specified for each fabric element.

E_Port segmentation

When an ISL activates, the two fabric elements exchange operating parameters to determine if they are compatible and can join to form a single fabric. If the elements are incompatible, the connecting E_Port at each director or switch segments to prevent the creation of a single fabric. A segmented link transmits

only Class F traffic; the link does not transmit Class 2 or Class 3 traffic. The following conditions cause E_Ports to segment:

- **Incompatible operating parameters:** Either the R_A_TOV or E_D_TOV is inconsistent between the two fabric elements.
- **Duplicate domain IDs:** One or more domain ID conflicts are detected.
- **Incompatible zoning configurations:** Zoning configurations for the two fabric elements are not compatible. For an explanation, refer to “Merging fabrics” on page 481.
- **Build fabric protocol error:** A protocol error is detected during the process of forming the fabric.
- **No principal switch:** No director or switch in the fabric is capable of becoming the principal switch.
- **No response from attached switch:** After a fabric is created, each element in the fabric periodically verifies operation of all attached switches and directors. An ISL segments if a switch or director does not respond to a verification request.
- **ELP retransmission failure time-out**: A director or switch that exhibits a hardware failure or connectivity problem cannot transmit or receive Class F frames. The director or switch did not receive a response to multiple exchange link parameters (ELP) frames, did not receive a fabric login (FLOGI) frame, and cannot join an operational fabric.

10.2.6 Special considerations

Special consideration and attention should be given to these items:

- Ordering enough port cards to have spare ports available. If both shortwave and longwave are going to be used, spares of both types should be considered. If only a few longwave ports are needed, the SFPs can be redistributed among different UPMs on directors.
- Providing independent power sources for the two line cords (except for the 2026-E12).
- Assigning the required LAN addresses.
- Providing the communications facilities for the call home feature.
- Verifying devices restrictions and requirements.
- Arranging cable connections. We recommend that multiple cable connections from a device or application not be attached to a single port card. Spreading connections among multiple port cards eliminates singles point of failure in case a port card fails and has to be replaced. This recommendation applies to

all connection types: servers, devices or ISLs when coupling to other switches.

- Labeling cables and assigning port names that allow easy identification of cable connections.
- Assigning a unique “Preferred Domain ID”, even if it is not a multi-switch fabric. The range is 1 to 31. This will make it easy to interconnect directors in the future, since by having all different preferred domain IDs, we can know in advance the domain ID that each director will have.
- Set “Insistent Domain” on all switches and directors so that the domain ID cannot change unexpectedly.
- Assigning a Switch Priority value of **Principal** to the director intended to be the principal director. The Principal switch assumes Domain Address Manager (DAM) functions and controls distribution and allocation of domain IDs to all switches in the fabric. When more than one director share the lowest switch priority value, the director with the lowest WWN gets the principal switch assignment.
- Upgrade firmware to the latest level.
- FICON Management Server (CUP) supports management of the 2027-232, ED-6064 and ED-6140 by System Automation for the IBM zSeries servers. This provides a single point of control for managing connectivity in active I/O configurations.
- Only the 2026-224 and 2026-E12 support FC-AL connections.
- All current directors and switches use LC SFP connectors.
- Firmware version 5.1 and EFCM 7.1 are required for Open Trunking.
- Zone all servers and storage, and disable the default zone.

10.2.7 Open Fabric

McDATA supports OEM interoperability through the use of McDATA Open Fabric (Interop-mode). Although WWN zoning is available, port zoning is not. Features which are implemented differently by each vendor might be unavailable. McDATA/Brocade interoperability is not supported by IBM except via an RPQ.

10.2.8 Supported devices, servers and HBAs

The list of supported devices, servers, and HBAs is constantly being updated as new configurations are certified and compatibility issues are fixed.

The current list of IBM supported configurations can be obtained at the following Web sites:

http://www.storage.ibm.com/ibmsan/products/2032/library.html#matrix

10.3 Features of directors and switches

Several optional software features may be licensed at additional cost. These are described in the following sections.

10.3.1 Element Manager

Element Manager is an optional feature for all switches (except the 2026-E12) and directors which works with the EFCM server to provide the user interface to each switch. If an EFCM server is being used to manage a multi-switch fabric, then every switch and director should be licensed for Element Manager.

For a complete explanation of the function and features of Element Manager see the *McDATA® Intrepid 6064 and 6140 Directors Element Manager User Manual*, 620-000172 manual.

10.3.2 FICON Management Server

This feature allows host control and inband management of the director or 2027-232 through an IBM System/390 or zSeries host attached to a switch port via a FICON channel. As of E/OS 6.0, this feature can coexist with the OSMS feature. See “FICON support” on page 465 for more detail. It is not available for other switches.

10.3.3 Full Volatility Option

McDATA products do not retain any Fibre Channel data frames following a power off. The hardware and associated buffers do not have the ability to retain frames. Any frames that are routed to the embedded port (for any reason) are also discarded upon power off.

In the default configuration, if the director should experience a fault condition it will capture a dump of the embedded memory space into non-volatile memory. The dump will retain recent frames transmitted from and to the embedded port. This non-volatile memory is physically on the CTP cards. While this information would be very difficult to extract and parse, it could exist in the default configuration.

To eliminate the dump, the Full Volatility option can be configured on products (except the 2026-E12) so that the dump capture capability is turned off. This would disable the memory capture process if a fault condition occurs. Obviously

this limits the amount of diagnostic information available for potential problem resolution, however the vast majority of problems are resolved without the use of the dump files, so this should not be a problem.

10.3.4 Open Systems Management Server

The OSMS feature allows host control and inband management of the switch or director (except 2026-E12) through a management application that resides on an open-systems interconnection (OSI) device. This device is attached to a switch port. The device communicates with the switch through Fibre Channel common transport (FC-CT) protocol. As of E/OS 6.0, this feature can coexist with the FICON Management feature.

OSMS is an ANSI-based feature and hence is supported by third-party SAN management software such as IBM Tivoli SAN Manager.

See “Fabric management” on page 465 for more detail on inband management.

10.3.5 Open Trunking

McDATA has developed a software implemented, more comprehensive solution than load balancing called Open Trunking, which is available for all directors and switches except for the 2026-E12.

Open Trunking is an optional, user-purchasable software feature that provides automatic, dynamic, statistical outbound traffic load balancing across ISLs in a fabric environment. This feature is available with E/OS 5.1 and EFCM 7.1 and can be enabled on a per-switch basis. It operates transparently to the existing FSPF algorithms for path selection within a fabric. It employs Template Registers in the port hardware and measures flow data rates and ISL loading — it uses these numbers to optimize use of the ISL bandwidth. The feature controls Fibre Channel traffic at a flow level, rather than at a per frame level in order to achieve optimal throughput. This feature may be used on McDATA switches in homogeneous as well as heterogeneous fabrics. This feature complies with current Fibre Channel ANSI standards.

Note: In a heterogeneous environment, you will be able to use the trunking feature from a McDATA switch to another vendors switch, but the return traffic from another vendors switch would not be able to be trunked.

Statistics

Open Trunking is performed using the FSPF shortest-path routing database. Traffic statistics are collected and periodically examined to determine which

traffic needs to be rerouted from heavily loaded ISLs to less loaded ISLs. Open Trunking measures these three statistics:

- The long-term (approximately a minute) statistical rates of data transmission between each ingress port (ISL or N_Port) and each destination domain.
- The long-term statistical loading of each ISL, measured in the same time-span as the above.
- The long-term average percentage of time spent with zero transmit BB_Credits for each ISL.

It should be noted that the zero BB_Credit statistic is not just the portion of time spent unable to transmit due to credit starvation. It also includes the portion of time spent transmitting with no more transmit credits. Since a credit is consumed at the start of a frame, not at the end of a frame, an ISL that is transmitting may have no transmit BB_Credits. It is common for an ISL to be 100% loaded and still have a zero transmit BB_Credit statistic of close to 100%.

Unfortunately, Open Trunking cannot do much to help ISLs that spend a lot of time unable to transmit due to lack of BB_Credits. This condition is normally caused by overloaded ISLs or poor-performing N_Ports elsewhere in the fabric, not at the local switch. The zero BB_Credit statistic is primarily used to ensure that Open Trunking doesn't make things worse by rerouting traffic onto ISLs that are lightly used but have little or no excess bandwidth due to credit starvation.

Rerouting and cost function

Rerouting is accomplished by modifying the switch hardware forwarding tables. Traffic may be rerouted from an ISL of one capacity to an ISL of another capacity if there would be an improvement to the overall balance of traffic. Whenever traffic is rerouted, there is a possibility of out-of-order packet delivery. Therefore, the algorithms used are extremely cautious and are based on long-term stable usage statistics. A significant change in traffic patterns must last for roughly a minute or longer, depending on the situation, before Open Trunking can be expected to react to it.

For example, as shown in Figure 10-13, if Open Trunking recognizes that ISL1 is 99% loaded and has traffic from HBA1 and HBA2, while ISL2 is 10% loaded with traffic from HBA3, it might reroute either the flow from HBA1 or HBA2 onto ISL2. The choice is determined by flow statistics: If the flow from HBA1 to SW1 is 1.9 Gb/s, it wouldn't reroute that flow, because doing so would overload ISL2. In that case it could only reroute the flow from HBA2 to ISL2.

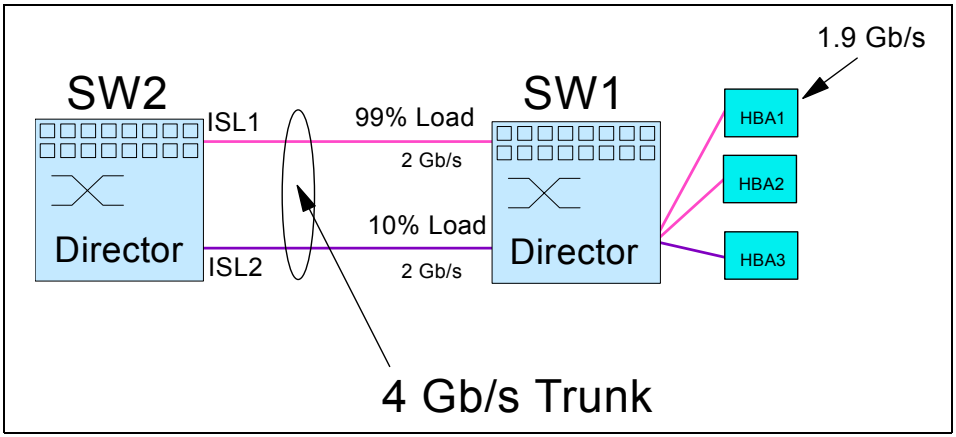


Figure 10-13 McDATA Open Trunking

At the heart of Open Trunking is a cost function that computes a theoretical cost of routing data on an ISL. It is the cost function that makes it possible to compare loading levels of links with different bandwidth, 1 Gb/s versus 2 Gb/s: a 1 Gb/s ISL with 0.9 Gb of traffic is not equally as loaded as a 2 Gb/s ISL with 0.9 Gb/s of traffic. The cost function is based on the ISL loading and the link bandwidth.

All rerouting decisions are made in such a way as to minimize the cost function. This means that a flow is rerouted from ISL X to ISL Y only if the expected decrease in the cost function for ISL X, computed by subtracting the flow's data-rate from ISL X's data-rate, is greater than the expected increase in the cost function for ISL Y. In fact, to enhance stability of the system, the expected increase in the cost function for ISL Y must be at least 10% less than the expected decrease in the cost function for ISL X.

The cost functions are kept in pre-compiled tables, one for each variety of ISL (currently 2 Gb/s and 1 Gb/s). The 10% differential mentioned previously is hard-coded in the tables. The cost function is mainly needed because of the difficulty of making rerouting decisions among ISLs of differing bandwidths; without that requirement, Open Trunking could just reroute in such a way as to minimize the maximum ISL loading.

Multiple checks have been implemented on the rerouting selection process to prevent off-loading traffic from a lightly-loaded ISL onto an even more lightly-loaded ISL. User configurable options can be changed by EFCM, CLI, or SANpilot configuration.

- Two versions of the ISL statistical data-rate are kept, one designed to underestimate the actual data-rate and the other designed to over-estimate it.

When making a rerouting decision, the statistics are used in such a way as to result in the most conservative (least likely to reroute) decision.

- No flow is rerouted from an ISL unless the ISLs utilization is above a minimum threshold called the “off-loading bandwidth consumption threshold” or unless it spends more than “low BB_Credit threshold” portion of its time unable to transmit due to lack of BB_Credits. In the absence of one of these conditions, there really isn’t any condition that justifies the cost of rerouting. Both of these parameters are user-configurable.
- No flow is rerouted to an ISL unless the ISLs expected utilization, computed by adding the flow’s data-rate to the ISLs current data-rate, is less than an “onloading bandwidth consumption threshold”. There is an onloading bandwidth consumption threshold for each ISL capacity. This threshold is not user-configurable.
- No flow may be rerouted if it has been rerouted recently. A period of “flow reroute latency” must expire between successive reroutes of the same flow. This latency is not user-configurable.

Periodically, every “load-balancing period”, a rerouting task runs that scans all flows and decides which ones to reroute using the criteria discussed above. The load-balancing period is not user-configurable.

10.3.6 Preferred Path

The Preferred Path feature enables you to influence the route of data traffic when traversing multiple switches or directors in a fabric. If more than one ISL connects switches in your SAN, this feature will be useful for specifying an ISL preference for a particular flow. The data path consists of the source port of the switch or director being configured, the exit port of that switch or director, and the domain ID of the destination switch or director. Each switch or director must be configured for its part of the desired path in order to achieve optimal performance. You may need to configure Preferred Paths for all switches or directors along the desired path for proper multi-hop Preferred Path operation.

It is not available for the 2026-E12.

Attention: Activation of a new Preferred Path will cause a reroute to occur if the Preferred Path is different from the current path.

10.3.7 SANtegrity Binding

SANtegrity Binding is an optional part of the SANtegrity Security Suite (see “Security” on page 483). It enables fabric and director/switch binding (along with port binding) that essentially locks downs the fabric so that the appropriate

devices are connected in the manner you defined, and not changed by accidental or unauthorized access to the fabric. It is not available for the 2026-E12.

This feature can enhance the security of both open system FCP and zSeries FICON environments, and requires E/OS version 4.0 and EFC Management Server version 6.1. It can be used with the Open System Management Server (OSMS).

Fabric binding controls which switches are allowed to attach to a fabric. This provides security from accidental fabric merges and potential fabric disruption when fabrics become segmented because they cannot merge. Switches are defined by their WWN and Domain ID, so the Insistent Domain ID option must be enabled on all switches.

Switch binding controls which devices and switches can be connected to director or switch ports. It is defined by the WWN of the attached switch or device port. You can choose to restrict just F_Ports (devices), E_Ports (ISLs) or both.

If you wish to implement FICON cascading (source and destination ports on different directors connected via ISLs) with zSeries, then a high integrity fabric is required. This is achieved with SANtegrity Binding, and requires the fabric binding feature to be used. A special *Enterprise Fabric Mode* can be enabled, which will enable all features required for high integrity. Note that some changes may require the directors to be taken offline, and hence are disruptive, so should be carefully planned.

10.3.8 Feature activation

A feature key is a varying length string of alphanumeric characters consisting of both uppercase and lowercase (e.g. `XxXx-XXxX-xxXX-xX`).

Note: The key is case sensitive and it must be entered exactly, including the dashes.

The feature key, which is encoded with a product's serial number, can only be configured on the director or switch to which it is assigned. Encoded within the key are all the features which have been licensed for the product (e.g. Element Manager, OSMS, Open Trunking).

When a new feature is purchased, a numeric transaction code (in the format `xxx-xxx-xxx`) is issued, and this is used to generate a new key which is then used to replace the existing key. If several features are purchased at the same time for the same product, a matching number of transaction codes will be received.

If a feature key has not been supplied, or a new one needs to be generated, go to the McDATA Product Feature Enablement web site (registration required) https://mcdata.getkeys.com/ and enter the product serial number and transaction codes. A feature key will then be generated.

Important: Do not lose the transaction codes in case you ever have to regenerate the feature key.

10.4 FICON support

FICON Intermix enables customers to simultaneously run mainframe (FICON) and Open Systems (FCP) data traffic through a shared McDATA director or 2027-232 switch.

In an intermixed environment, zoning must be established so that open hosts can only communicate with FCP storage devices, and zSeries hosts with FICON storage devices. For FICON, only a single zone containing all hosts and devices is required, since device access is managed via IOCP, rather than name server discovery as with open systems.

FICON cascading currently allows customers to scale their mainframe storage environments by supporting single hop, two director FICON fabrics. It requires the full SANtegrity security suite. See “SANtegrity Security Suite” on page 484 and “SANtegrity Binding” on page 463 for more details.

10.5 Fabric management

In the topics that follow we talk about the methods for managing the fabric.

10.5.1 In-band management

In-band management was introduced in firmware release 3.0 and was available in two mutually exclusive versions. As of E/OS 6.0, the two features may be installed at the same time, and either an Open Systems or FICON Management Style view chosen:

- **Open Systems Management Server (OSMS):** OSMS is an ANSI-based feature that supports SAN management software from vendors such as IBM Tivoli. OSMS extends the switch's capability to include in-band management by an open systems host-based application. OSMS allows the Fabric Switch and devices attached to it to be discovered or seen in a fabric through a framework software application.

- **FICON Management Server (FMS):** The FICON management server is an in-band management feature developed by IBM that identifies an entity known as the Control Unit Port (CUP), which can always be accessed from any port on the switch.

In-band management console access (through a Fibre Channel port) is provided by enabling user specified features that allow OSMS or FMS host control of the director.

When the Element Manager application is configured for Open Systems operating mode, control and management of the director is provided by a director software subsystem (management server) that communicates with an application client. When implementing in-band director management through a Fibre Channel connection, we should consider the following minimum host requirements:

- Connectivity to an open systems interconnections (OSI) server with a director-compatible host bus adapter (HBA) that communicates through the Fibre Channel common transport (FC-CT) protocol.
- Installation of a SAN management application on the OSI server. Management applications include IBM Tivoli SAN Manager or Veritas SANPoint Control.

There is also a FICON Management Server (FMS) to support in-band management of the director by System Automation for OS/390. This is an optional feature. System Automation for OS/390 provides a single point of control for managing connectivity in active I/O configurations. It takes an active role in detecting unusual I/O conditions and let a customer view and change paths between a processor and an I/O device using the dynamic switching capabilities of the director.

10.5.2 Out-of-band management

The McDATA portfolio provides for out-of-band management access in the following ways:

- Through the EFC Server attached to the switch or director's Ethernet port.
- Through a remote personal computer (PC) or workstation connected to the EFC Server through the customer intranet.
- Through a simple network management protocol (SNMP) management workstation connected through the director LAN segment or customer intranet.

- Through a PC with a direct serial connection to the director maintenance port (rear of the director chassis). The maintenance port is used by installation personnel to configure switch network addresses.
- Through a PC with a modem connection to the EFC Server. The modem is for use by support center personnel only.
- Through a PC with a Web browser and Internet connection to the director through a LAN segment for SANpilot access.

Out-of-band management is performed through the EFC Server, either by dedicated applications like EFC Manager, SANpilot or by remote SNMP workstations that access an SNMP agent running on the EFC Server.

10.5.3 EFC Server

The EFC Server is a 1U high rack mounted Intel server running Windows 2000, with a single power cord, and includes an internal modem and CD-RW drive. The server supports up to 48 McDATA directors or switches (managed products). The server is used to run the Enterprise Fabric Connectivity Manager (EFC Manager) and Element Manager applications which monitor product operation, change configurations, download firmware updates, backup configurations, and initiate diagnostics.

Unlike the previous laptop version of the EFC Server, the current one is not supplied with a keyboard, mouse or display. As such, initial network configuration is done via a front panel display and cursor buttons (see Figure 10-14). An alternative method is to connect a keyboard, mouse and display to the rear of the server (see Figure 10-15 on page 468) and reboot. A normal Windows logon panel will then appear and configuration can be carried out using the normal GUI method. Once the network is working, VNC can be used for remote console access.

A server failure does not affect port connections or functions of an operational director or switch. The only operating effect of a server failure is loss of remote access, configuration, management, monitoring functions, and call-home.

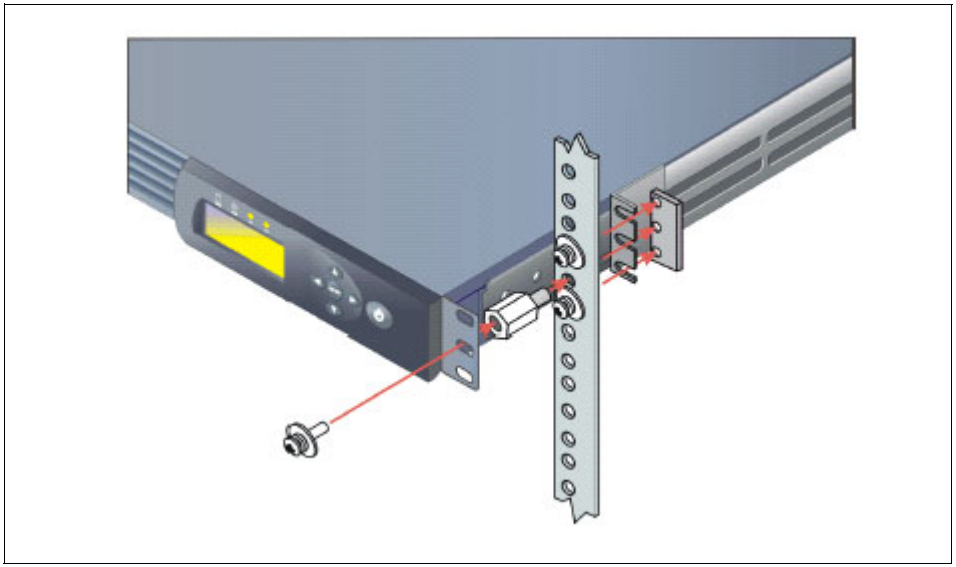


Figure 10-14 LCD panel on front of McDATA 1U EFC Management Server

Connectivity

The EFC Server provides an auto-detecting 10/100 Base-T Ethernet interface that connects to the 24-port hub mounted at the top of the Fabricenter equipment cabinet. This provides access to the switch product network. If an optional customer intranet is used for LAN connections, the EFC Server provides a second auto-detecting 10/100 Base-T Ethernet connection. This interface is used for remote workstation access.

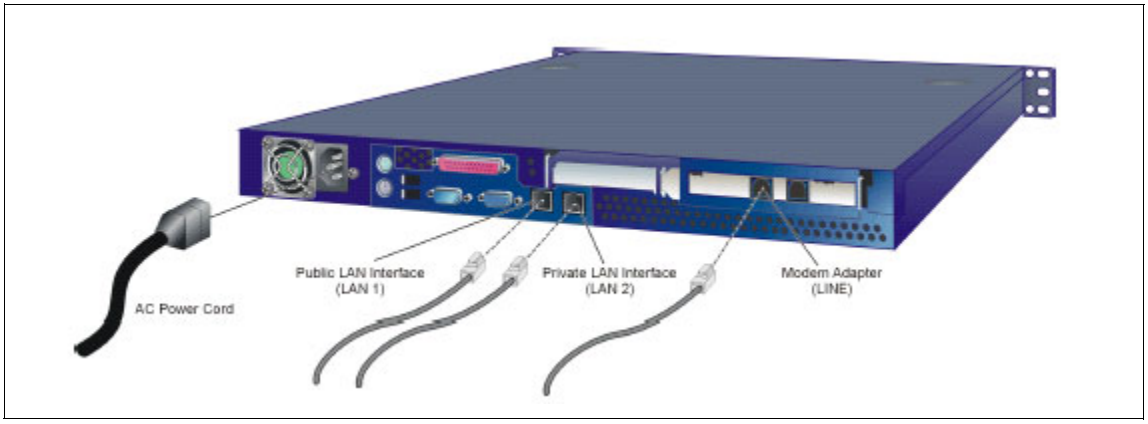


Figure 10-15 Rear view of McDATA 1U EFC Management Server

The EFC Server has an internal modem for service and support of managed products. The modem provides a dial-in capability that allows authorized service personnel to communicate with the EFC Server and operate the EFC Manager and Element Manager applications remotely. The modem is also used to automatically dial out to an authorized support center (to report the occurrence of significant system events) using a call-home feature. The call-home feature is enabled in the Element Manager application and configured through the dial-up networking feature of Windows.

Connectivity planning considerations

Directors, switches, and the EFC Server are delivered in a cabinet mount configuration in accordance with customer specifications. Because Ethernet cables that connect the managed products, the hub, and the EFC Server are factory installed, connectivity planning is not required for a stand-alone cabinet installation. However, consider the following Ethernet connectivity issues when:

- **Installing additional directors/switches:** Ensure cable lengths provide sufficient cable inside the cabinet to connect to product Ethernet ports such as an Ethernet Hub.
- **Interconnecting Fabricenter cabinets:** To increase the number of products managed by one EFC Server, Ethernet hubs in one or more equipment cabinets must be connected. In addition to planning for an Ethernet cable length that will connect the two cabinets, also plan for an additional 1.5 m (5 feet) of cable outside the cabinet to provide slack for service clearance, limited cabinet movement, or inadvertent cable pulls.
- **Consolidating EFC Server operation:** For control and efficiency, all directors and switches in a multi-switch fabric should be managed by one EFC Server. When products in two or more cabinets are joined to form a fabric, the PC environment should be consolidated to one server and one or more clients. Plan for Ethernet cabling to interconnect cabinets and ensure all directors, switches, and PC platforms participating in the fabric have unique IP addresses.

Remote user workstations

Administrators may restrict remote access to selected workstations by configuring the TCP/IP addresses of those workstations through the EFC Manager application. They may also specify the maximum number of simultaneously active sessions. Remote workstations must have access to the LAN segment on which the EFC Server is installed. Product administrative functions are accessed through the LAN and EFC Server.

The LAN interface can be part of the dedicated 10/100 Mb/s LAN segment that provides access to managed products. This Ethernet connection is part of the equipment cabinet installation and is required. Connection of remote

workstations through the hub is optional. This type of network configuration using one Ethernet connection through the EFC Server is shown in Figure 10-16.

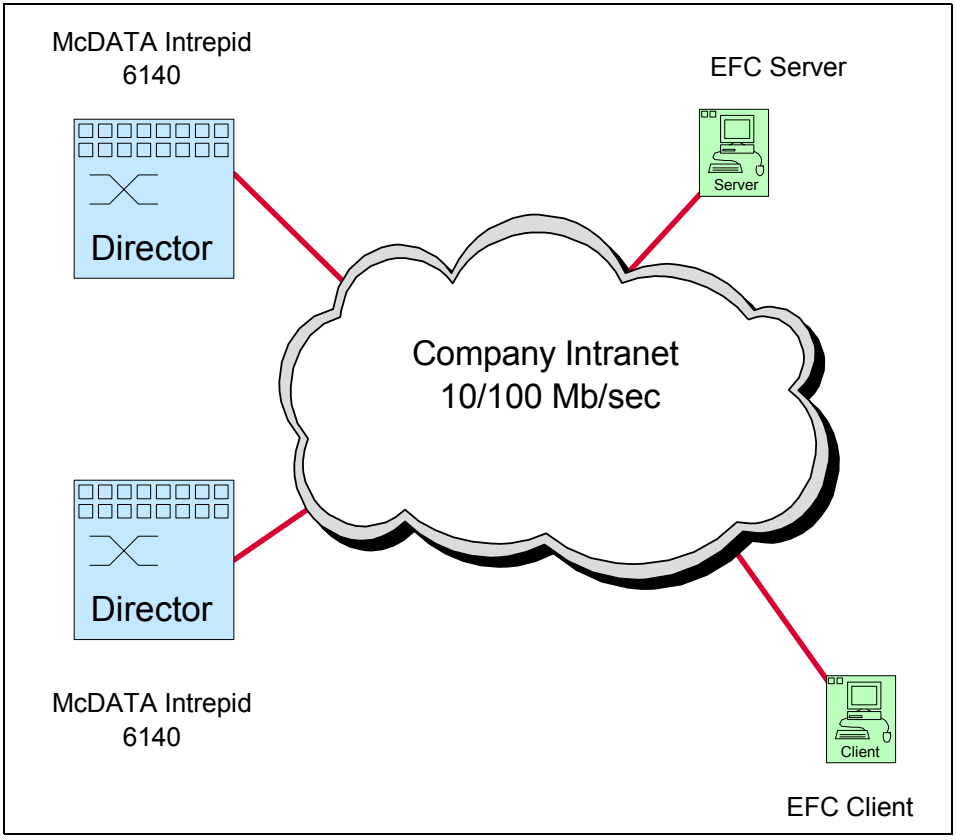


Figure 10-16 EFC Server public intranet (one ethernet connection)

A typical network configuration (with two Ethernet connections and if only one EFC Server connection is used) is provided through the customer intranet, with all functions provided by the EFC Server and available to users throughout the enterprise. The purpose for dual LAN connections is to provide a dedicated LAN segment that isolates the EFC Server and managed products from other users in the enterprise. Both Ethernet adapters in the EFC Server provide auto-detecting 10/100 Mb/s connections.

Figure 10-17 shows an example of a network configuration using both Ethernet connections on the EFC Server.

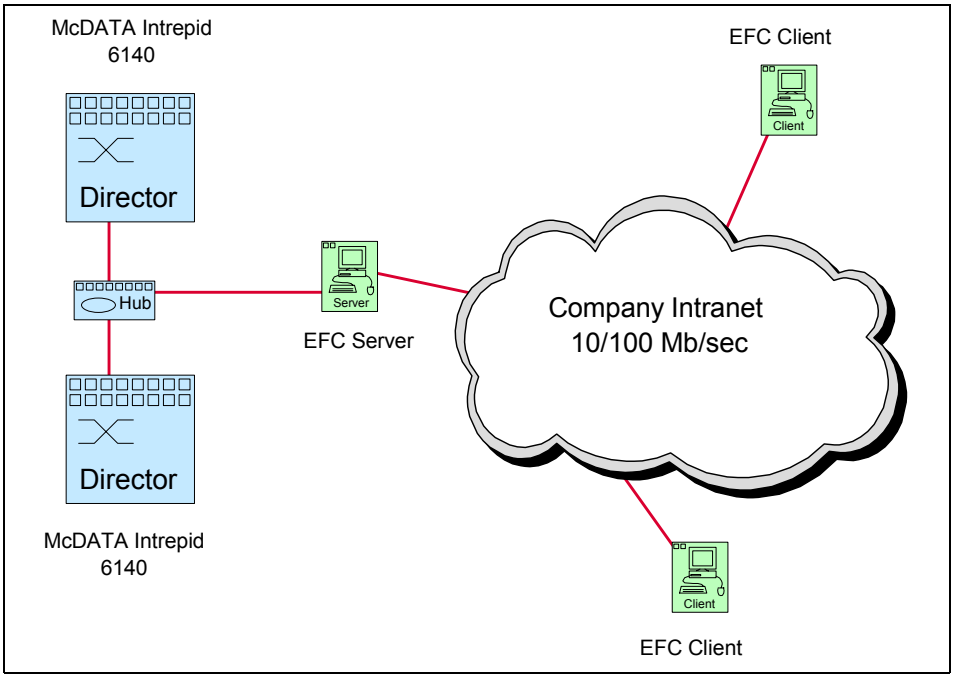


Figure 10-17 EFC Server private network (two ethernet connections)

The EFC Manager and Element Manager applications download and install to remote workstations from the EFC Server, using a standard Web browser. The applications operate on PC, Linux or UNIX workstations.

Server migration

If upgrading an existing EFC Manager from the old laptop to the new 1U server, the following should be considered:

- The laptop only runs EFCM versions up to 7.1, or 8.1b.
- The 1U server only runs EFCM 7.2 or 8.0 (or later).
- EFCM 8.x requires a serial number (from the CD jewel case) and a license key. For existing installations these are obtained by ordering FC 8001, which is a no charge feature.
- The laptop must be at least EFCM 6.0 to upgrade.
- All switch and director products to be managed by the EFCM must be at E/OS 5.1 or greater for EFCM 7.1, and ideally E/OS 6.0 for EFCM 8.x.
- All switch (except 2026-E12) and director products should be licensed for Element Manager.

- Whenever the EFCM server is upgraded, the remote clients must be upgraded to match.
- See the *McDATA® EFCM 7.x to 8.1 Transition Guide*, 620-000183 for full migration details.

10.5.4 EFC Manager

The EFC Manager application provides a common Java-based GUI for all managed McDATA products which support the Element Manager feature. It is intended to give a fabric wide view of the SAN, and can discover non-McDATA switches, provided the principal switch in a fabric is a McDATA switch. The application is accessed on the EFC Server through a network connection from a remote user workstation. The application operates independently from the director, switch, or other product managed by the EFC Server. Users can perform the following common product functions:

- Configure new McDATA products and their associated network addresses (or product names) to the EFC Server for access through the EFC Manager and Element Manager applications.
- Display product icons that provide operational status and other information for each managed McDATA product.
- Open an instance of the Element Manager application to manage and monitor a specific McDATA product.
- Open the Fabrics View to display managed fabrics, manage and monitor fabric topologies, manage and monitor zones and zone sets, and show routes (data paths) between end devices attached to a multi-switch fabric.
- Define and configure user names, nicknames, passwords, SNMP agents, and user rights for access to the EFC Server, EFC Manager application, and managed McDATA products, either locally or from remote user workstations.
- Configure Ethernet events, e-mail notification for system events, and call-home notification for system events.
- Display EFC audit, EFC event, session, product status, and fabric logs.

With EFCM 8.0 and later, there is greater control over user privileges than in earlier versions. The new version has the same look and feel as the SANavigator product, as shown in Figure 10-18, and includes some of the function that was previously only available with that product.

Important: EFCM 8.x requires a serial number (available from the EFCM CD jewel case) and a license key to install.

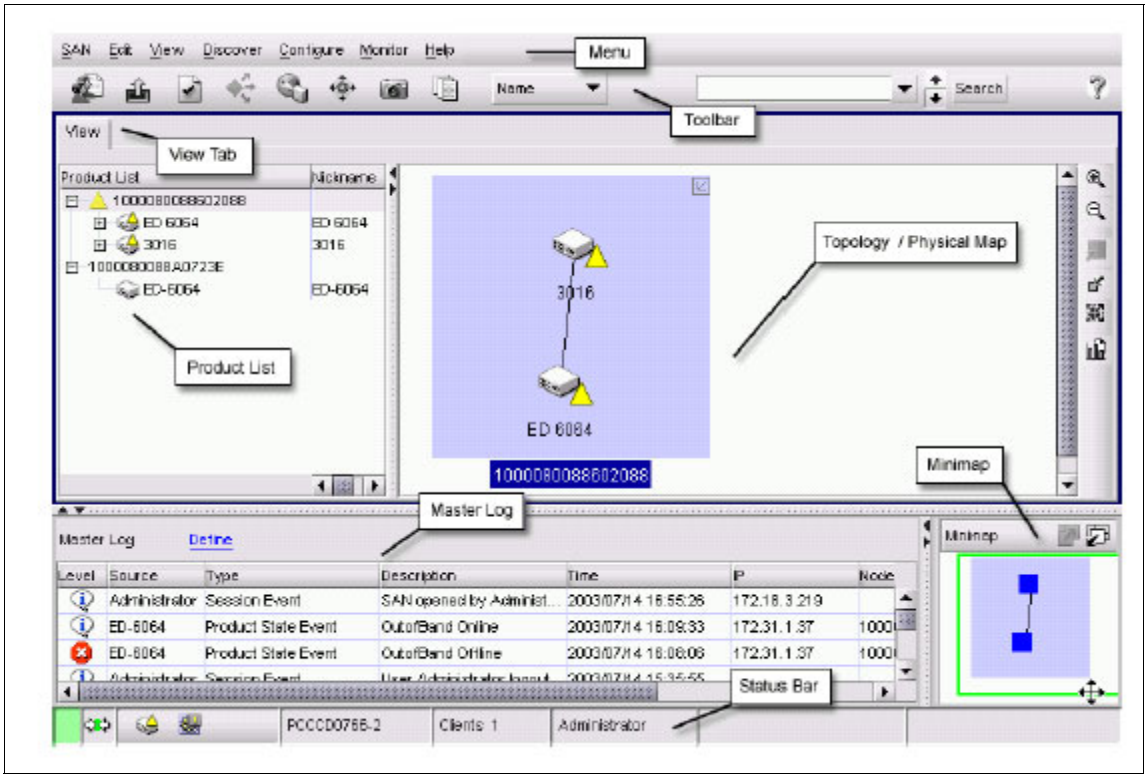


Figure 10-18 EFCM 8.0 main window

Optional features

There are currently two optional and chargeable features:

Performance and Event Management Module

Performance Monitoring allows you to measure the current performance statistics, historic metrics and future trends of every switch port on the SAN.

Event Management provides the ability to automate routine tasks and reduce the amount of manual intervention necessary for the management of the SAN.

SAN Planning Module

The tools available in the Planning module help evaluate the effects of a new device deployment on an existing SAN or plan for a completely new storage network using a set of best practice configuration rules.

10.5.5 Troubleshooting

When it is necessary to perform fabric problem determination, usually the first step will be to check for any alerts. If alerts are detected, the alert details should be checked. After this the appropriate logs should be examined. Some logs are part of the EFCM application, whilst each director or switch will also have its own logs viewable via the Element Manager.

EFCM logs

The EFCM has six logs:

Audit log

Displays a history of user actions performed through the application (except login/logout).

Event log

The EFC Manager's Event Log displays errors related to SNMP traps and Client-Server communications.

Session log

Displays the users who have logged in and out of the Server.

Product status log

Displays operational status changes of managed products.

Fabric log

This log displays events that have occurred for a selected fabric. To display the log, you must have persisted the fabric through the Persist Fabric dialog box. You must also select the persisted fabric from the Physical Map before selecting Fabric Log from the menu.

Master log

The Master Log, which displays in the lower left area of the main desktop, lists all events from the Element Manager and EFCM logs that occurred throughout the SAN in the past 48 hours. These include user actions, client/server communications, SNMP trap errors, product hardware errors, product link incident and threshold errors, and Ethernet events. This log combines entries from all other EFC Manager and Element Manager logs.

Note: The Master Log is not available with the 8.xb (low RAM) version.

Element Manager logs

The Audit, Event, Hardware, Link Incident, and Threshold Alert logs store up to 1000 entries each. The most recent entry displays at the top of the log. After 1000 entries are stored, new entries overwrite the oldest entries.

Audit log

The audit log displays a history of all configuration changes applied to the director or switch from any source such as Element Manager, SNMP management stations, or host.

Event log

The event log provides a record of significant events that have occurred on the director or switch, such as hardware failures, degraded operation, port problems, FRU failures, FRU removals and replacements, port problems, Fibre Channel link incidents, and communication problems between the director and the server platform. The information is useful to maintenance personnel for fault isolation and repair verification.

Hardware log

The hardware log displays information on FRUs inserted and removed from the director.

Link Incident log

The link incident log displays a thousand of the most recent link incidents. The information is useful to maintenance personnel for isolating port problems (particularly expansion port (E_Port) segmentation problems) and repair verification.

Threshold Alert log

This log provides details of threshold alert notifications. Besides the date and time that the alert occurred, the log also displays details about the alert as configured through the Configure Threshold Alert(s) option under the Configure menu.

Open Trunking log

This log displays only if the optional Open Trunking feature is installed, and provides details on flow rerouting that is occurring through director ports.

10.5.6 SANpilot interface

The SANpilot interface is a standard, no charge, feature of all switches and directors. It is a management tool consisting of an embedded Web server which enables administrators and operators with an internet browser to monitor and

manage individual switches or directors. The SANpilot interface does not replace nor offer the management capability of the EFC Manager and Element Manager applications (for example, the Web server does not support all director maintenance functions).

SANpilot users can perform the following operations:

- Display the operational status of the director, FRUs, and Fibre Channel ports, and display director operating parameters.
- Configure the director (identification, date and time, operating parameters, and network parameters), ports, SNMP trap message recipients, fabric zones and zone sets, and user rights (administrator and operator).
- Monitor port status, port statistics, and the active zone set, and display the event log and node list.
- Perform director firmware upgrades and port diagnostics, reset ports, enable port beaconing, and set the director online or offline.

The SANpilot interface can be opened from a standard Web browser running Netscape Navigator 4.6 or higher, or Microsoft Internet Explorer 4.0 or higher. At the browser, enter the IP address of the director or switch.

10.5.7 Command line interface

The command line interface (CLI) provides a director and switch management alternative to the EFC Manager, Element Manager, and SANpilot user interfaces. The interface allows users to access application functions by entering commands through a PC-attached telnet session. Any platform that supports telnet client software can be used. The primary purpose of the CLI is to automate management of several directors or switches using scripts.

Although the CLI is designed for use in a host-based scripting environment, basic commands (**`config`**, **`maint`**, **`perf`**, and **`show`**) can be entered directly at a DOS-style command prompt. The CLI is not an interactive interface; no checking is done for pre-existing conditions, and a user prompt does not display to guide users through tasks. For additional information, refer to the *McDATA® Enterprise Operating System Command Line Interface User Manual*, 620-000134.

Tip: To use automated scripts, you must have TCP/IP connectivity from a management server. If the switches have been defined on a separate private network, you will need a management server on that subnet, or you will need to establish a VLAN which includes a management server on your companies Intranet.

10.5.8 SNMP

For SNMP communication, the standard MIBs supported by McDATA products are:

- MIB-II (Internet MIB) as described in RFC 1213: supported by all switches and directors.
- Fibre Alliance (FCMGMT) MIB, version 3.1.
- Fibre Channel Fabric Element (FCFE), version 1.10: supported by all switches and directors.

Configuration of the SNMP agent is accomplished through the Embedded Web Server, CLI, and EFC Element Manager. The switch and director resident SNMP agents:

- Support SNMPv1 manager.
- Enable access to variables in the standard MIB-II definition, the Fibre Channel Fabric Element MIB, and switch or director Private MIB. All groups and variables in the supported MIBs are read only by SNMP management stations unless noted otherwise.
- Enable the switch or director to send unsolicited trap messages to the network management station when specific events occur on the switch or director.

The traps supported are:

- Standard generic traps
- Switch or director enterprise-specific traps

For a detailed description on how to configure SNMP management please refer to the *McDATA SNMP Support Manual*, 620-000131.

10.6 Zoning

McDATA switches implement zoning by WWN or port number, or a combination of the two. Enforcement of WWN zoning depends on the release of E/OS on the switch.

Software-enforced zoning

For earlier versions of director or switch firmware (prior to Version 5.0), the device configuration at a fabric element enforces zoning by limiting access to name server information in response to a device query. Only devices in the same zone as the requesting device are returned in the query response. This type of zoning is also called name server zoning or soft zoning.

Hardware-enforced zoning

For later versions of director or switch firmware (Version 5.0 and later), the device configuration at a fabric element enforces zoning by programming route tables that strictly prevent Fibre Channel traffic between devices that are not in the same zone. E/OS 6.0 and later extended hardware enforcement to FL_Ports on 2026-E12 and 2026-224 switches. This type of zoning is also called hard zoning.

Zone characteristics

- Each device port that belongs to a zone is called a zone member
- The same device can belong to more than one zone (overlapping zones)
- Zones can spread through multiple directors in a multi-switch fabric
- ISLs are not specified as zone members, only device ports

10.6.1 Configuring zones

Zoning is configured through the Fabric Manager application. Members can be configured by specifying the director ID and port number, or by the eight byte WWN of the device. Nicknames can be assigned to each WWN by means of the EFC Manager Configure Nicknames dialog box. We recommend using meaningful nicknames to make it easier to define zoning, since we can use the nickname instead of the WWN of a device. It is also recommended that a standard naming convention be deployed to provide conformity.

There are a maximum 4096 zone members, but the exact number of zone members that can be defined is bounded by the available nonvolatile random access memory (NVRAM) in the director or switch and depends on the number of zones defined, length of zone names and other factors.

Zoning by WWN

Defining members by WWN or nickname has the advantage that the zone definition will not change if we move the port in the director. This is useful when rearranging ports or moving to a spare port because of a port failure. The disadvantage is that removing or replacing a device HBA and thus changing its WWN disrupts zone operation and may incorrectly exclude or include devices until the zone is re-configured with the new WWN.

In order to make it easy to reconfigure WWN or nicknames in affected zones there are Find, Remove and Replace WWN/Nickname dialog boxes available among the Zoning Tasks.

Note: Some devices such as the IBM TotalStorage Enterprise Storage Server avoid this problem by effectively preserving the WWN on replacement host adapters.

Zoning by port number

By using port numbers to define zone members, any device attached to that port can connect with the others in the same zone. It has the advantage that we do not have to worry about redefining the WWN if an HBA needs to be replaced. A disadvantage is that somebody could rearrange the port connections to allow the possibility of gaining access to devices that you did not intend them to have access to, and losing access to correct devices.

To provide a higher level of security, you can also configure the port binding feature to bind a WWN to a given port which will not allow any other device to plug into the port. See “Controlling access at the switch” on page 484 for more details.

Note: In Open Fabric (interop) mode, port zoning is not supported.

Zone sets

Zones are grouped in zone sets. A zone set is a group of zones that can be activated or deactivated as a single entity across all managed products either in a single switch fabric or in a multiple switch fabric. There can be a maximum of 1024 zones in a zone set (1023 plus the default zone) and up to 64 zone sets can be defined in the zone library.

A **default** zone groups all devices not defined as members of the currently active zone set. The devices in the default zone can communicate with each other, but they cannot communicate with the members of any other zone. The default zone can be enabled or disabled independently of the active zone by the Default Zone option of the Configure Menu. Default zoning comes enabled by default.

It is always wise to be careful when activating zone sets, as any one of the following events could occur, whether by design or by accident:

- When the default zone is disabled, the devices that are not members of the active zone set become isolated and cannot communicate.
- When no zone set is active, then all devices are considered to be in the default zone. If no zone set is active and default zone is disabled then no device can communicate.

- Activating a new zone set will replace the currently active zone set. Be sure you have the correct zone set for the fabric you are currently updating, if your EFC Manager manages multiple fabrics.

  Note: EFC Manager 6.0 and higher provides a new feature which provides a difference check against the current active zone set. Any differences will be highlighted to alert you of any potential inconsistencies. This should eliminate the chance of an incorrect zone set activation.

- Deactivating the currently active zone set will make all devices members of the default zone if default zoning is enabled. If default zoning is disabled, all communication will stop.
- Zones defined through the Fabric Manager are saved in a zone library. Any zone in the zone library can be displayed, modified, and selected to be part of a zone set.

Tip: It is strongly recommended that all devices be properly zoned, and the default zone set disabled, as this will improve fabric security.

Saving zone information

Zoning information is saved in the **`EfcData`** directory of the EFC Server and from there it is backed up to a CD-RW Drive which comes with the EFC Server. This makes it possible to carry and replicate the zoning configuration onto a separate fabric controlled by a different EFC Server. To restore the configuration in a different director, it must have the same TCP/IP address as the director the configuration was saved on.

Zone change notification

A fabric format Registered State Change Notifications (RSCN) service request is sent to all N_Ports when the zoning configuration is changed, unless the “Suppress RSCN’s on zone set activations” option is checked in the Switch Operating Parameters dialogue box.

Broadcast frames are transmitted to all N_Ports, regardless of the zone they belong to.

10.6.2 Zoning and LUN masking

Zoning allows us to specify which ports can connect with each other. When we are connecting to storage arrays or storage subsystems, like the IBM TotalStorage Enterprise Storage Server, with multiple LUNs defined, we still

need to perform LUN masking at the storage subsystem level, so each host is only allowed to access its own LUNs.

10.6.3 Persistent binding

Server-level access control is called persistent binding. Persistent binding uses configuration information stored on the server, and is implemented through the server's HBA driver. The process binds a server device name to a specific Fibre Channel storage volume or logical unit number (LUN), through a specific HBA and storage port WWN.

Some operating systems (e.g. Solaris) require persistent binding in order to define the host to LUN access. This should be checked as part of your host configuration.

10.6.4 Blocking a port

In addition to zoning, access to a port can be disabled at any time by "blocking" the port from the Element Manager application. A blocked port can not accept connections, it only transmits Off Line sequences. Blocking can be done individually by selecting the Port Menu options, or all four ports in a card at once using the Port Card Menu options.

Tip: All unused ports should be blocked, as this reduces the risk of unplanned devices being able to connect to the fabric.

10.6.5 Merging fabrics

Two or more directors or switches can be interconnected using ISLs to form a single fabric. In a multi-switch fabric, the zoning configuration applies to the entire fabric.

Merging zone information

When products join, active zone information is interchanged between adjacent units to determine if they can merge. If one or both of the fabrics are not zoned, then they will merge successfully, and if an active zone exists, it will be propagated to the joining fabric. If both fabrics are zoned, the compatibility requirements in order to merge are:

- Zone names must be unique to each fabric.
- If a zone name exists on both fabrics, then they should have the same members.
- Domain IDs must be unique.

If configurations cannot merge, then the E_Ports on each product become segmented and the associated ISL cannot carry traffic from attached devices, but can carry management and control traffic.

Unique domain IDs

When merged, each director in the resulting fabric must have a unique domain ID. If there are duplicated IDs, the directors will not merge. Duplicated IDs are another reason for E_Port segmentation.

Once the switches are connected together, domain IDs are assigned at power on by the principal switch in the fabric. At power on a director requests its ID from the principal switch and reports its preferred ID. If the preferred ID is available, it is assigned, otherwise the principal switch assigns an available ID. If we have directors in the fabric with the same preferred ID, results may be inconsistent. In the case of a power down, it all depends on which switch is powered up first (see “Principal switch selection” on page 455).

If any zone member was defined by port number and domain ID, the zone must be re-configured when the domain ID changes.

For these reasons we strongly recommend assigning unique preferred domain IDs to each director, and using the Insistent domain option. In order to alter the operating parameters the director must be taken offline, so it is important to configure correctly at installation time.

Time-out values

Another requirement is to have the same R_A_TOV, E_D_TOV values, otherwise the ISLs will become segmented. This should only be a concern for extended distances; for direct connection the default values of R_A_TOV and E_D_TOV do not need to be changed. For going beyond 10 km, the time-out values may need to be adjusted according to the distance and the equipment used for the extended link. In this case we have to make sure that all directors in the fabric have the same values.

10.7 Performance

Each port on a director has 16 buffer credits assigned by default. Ports configured for extended distances (10 - 100 km) are assigned 60 buffer credits.

McDATA directors use a single-stage switching architecture which provides a consistently low latency for all ports. For 1 Gb/s traffic, the latency is less than 2.5 microseconds, whilst for 2 Gb/s it is less than 2.0 microseconds.

If we have too many server ports trying to talk to the same storage port we can have the server ports not being able to run at full speed since the storage port is oversubscribed. This means that the fan-in ratio of the storage port is too high. The same thing can happen when we are cascading directors, and if we do not have enough ISLs they can become congested.

The Element Manager offers a real-time Performance view that can help detect these situations. At the top of the view there is a graphical display of performance for all ports. Each bar graph in the upper portion of the main panel displays the level of transmit/receive activity for the port. This information updates every five seconds. Each bar graph shows the percentage link utilization for the port. A red arrow marks the highest utilization since the opening of the Performance view.

Clicking in the bar graph for a port, the statistics counters for that port are shown below the graph bars. These are cumulative port and error statistics. The counter can be cleared by clicking on the Clear action button and confirming the operation.

Ports that appear at 100% utilization and have a high number of discarded frames may be a good indication of oversubscription.

10.8 Security

These are some of the security recommendations and McDATA security features along with the importance of implementing them to avoid exposures.

10.8.1 Restricting access to those that need it

It is important to ensure that all default passwords have been changed for the following interfaces:

- EFC Manager
- SANpilot for each director and switch
- CLI (Telnet) for each director and switch

Tip: It is recommended that each authorised user of EFCM have their own account and password, rather than sharing a common account and password.

Logon from remote workstations can be limited to specific network addresses or completely suppressed using the Configure Session Options dialog box with System Administrator rights.

In order to perform the management functions, the EFC Manager should be able to communicate with the director. This is done through the director and EFC Server Ethernet connection. In order to keep this connection independent of the customer network, it is advisable to build a private LAN with directors, switches, EFC Servers and finally a remote workstation.

Just as important as securing the LAN connection is avoiding any accidental changes of the network address settings in the director itself. The TCP/IP address is set from the maintenance port, and normal site security precautions should be in place to avoid physical access to the director by anyone other than authorized personnel. Provided physical controls are in place, then it is not normally necessary to change the default maintenance port passwords. The same physical security precautions should be considered for the Ethernet hub where directors and EFC Servers are connected.

10.8.2 Controlling access at the switch

A port binding feature is available on switches and directors that allows you to “bind” a specific switch or director port to the WWN of an attached device for exclusive communication. If another HBA is plugged into a port that has previously been bound to a WWN, it will not work. Likewise, if after a WWN has been bound to a given port, it cannot be plugged into a different port.

This Port Binding feature is available through the **Configure Ports** option in the Element Manager application’s Configure menu. This feature is also available through the pop-up menu when you right click on a port in the **Hardware View, Port Card View, Port List View, and Performance View.**

10.8.3 SANtegrity Security Suite

See “SANtegrity Security Suite” on page 484 for information on fabric and switch binding.

10.9 SANavigator

SANavigator is a SAN management application capable of managing multi-vendor fabrics. Some of the functions previously only available with SANavigator, such as Performance and Event Management, and Planning are now available as options with EFCM 8.0 and later.

Other features are still only available with SANavigator. These are:

LUN Management

The SANavigator LUN Management tool gives you the ability to manage the LUNs on storage arrays from EMC, HDS and IBM. With this tool, you get the increased data security features associated with LUN masking as well as the ability to create new LUNs on supported storage arrays.

Provisioning

The SANavigator Provisioning module transforms the task of manual storage provisioning into a consistent, secure process. It allows you to automate and control systems operations, policy management and workflow processes according to your best practices, thereby increasing productivity and operational efficiency.

OPENconnectors

McDATA OPENconnectors act as an interface, providing the necessary infrastructure for SANavigator to share its knowledge of the fabric with higher level network management framework applications. Available connectors: HP OpenView NNM, BMC Patrol, CA Unicenter TNG, Tivoli Netview.

10.10 Licensing

All of the E/OS optional features, plus EFCM V8.0 and later, require chargeable licence keys to activate (see “Feature activation” on page 464 for feature activation details). License keys are tied to product serial numbers, and so must be ordered for each product requiring the feature.

10.10.1 Warranties

All products (hardware and software) have a standard 13 month warranty, which starts from the date of shipment. Warranties may be extended by one, two, three or four years, provided the extension is purchased before the existing warranty expires.

10.11 IBM feature codes

The following are the feature codes for the IBM offerings from McDATA.

FC-512 cabinet

- **FC 3768:** 1U rack-mount management server
- **FC 3183:** 2027-232 Sphereon 2027-232 fabric switch
- **FC 4583:** 2026-224 Sphereon 2026-224 fabric switch

- **FC 6083:** 2032-064 Intrepid 6064 director
- **FC 6283:** 2027-140 Intrepid 6140 director

Intrepid ED-6140

- **FC 3220:** Extended longwave 20 km transceiver
- **FC 3235:** Extended longwave 35 km transceiver
- **FC 6210:** 2 Gb/s Fibre 4-Port Modules with 4 shortwave transceivers (LC SFP)
- **FC 6211:** 2 Gb/s Fibre 4-Port Modules with 4 longwave transceivers (LC SFP)
- **FC 6002:** Open System Management Software
- **FC 6007:** SANtegrity binding
- **FC 6120:** Open Trunking software
- **FC 6140:** Element Manager
- **FC 6150:** Full volatility option
- **FC 6303:** FICON In-Band Management Server
- **FC 6422:** Preferred path option
- **FC 3765:** 1U rack-mount management server
- **FC 5030:** Ethernet hub with 24 ports
- **FC 8000:** EFCM fabric manager
- **FC 8001:** EFCM fabric manager upgrade
- **FC 8005:** Performance and event management module (512 ports)
- **FC 8007:** SAN planning module
- **FC 8015:** Performance and event management module (additional 64 ports)

Intrepid ED-6064

- **FC 3220:** Extended longwave 20 km transceiver
- **FC 3235:** Extended longwave 35 km transceiver
- **FC 6010:** Fibre 4-Port Modules with 4 shortwave transceivers (LC SFP)
- **FC 6011:** Fibre 4-Port Modules with 4 longwave transceivers (LC SFP)
- **FC 6210:** 2 Gb/s Fibre 4-Port Modules with 4 shortwave transceivers (LC SFP)
- **FC 6211:** 2 Gb/s Fibre 4-Port Modules with 4 longwave transceivers (LC SFP)
- **FC 6220:** 2 Gb/s director upgrade
- **FC 6221:** 2 Gb/s director upgrade with CTP2 processors
- **FC 6000:** Open System Management Software
- **FC 6006:** SANtegrity binding
- **FC 6020:** Open Trunking software
- **FC 6022:** Preferred path option
- **FC 6050:** Full volatility option
- **FC 6300:** Element Manager
- **FC 6302:** FICON In-Band Management Server
- **FC 3765:** 1U rack-mount management server

- **FC 5030:** Ethernet hub with 24 ports
- **FC 8000:** EFCM fabric manager
- **FC 8001:** EFCM fabric manager upgrade
- **FC 8005:** Performance and event management module (512 ports)
- **FC 8007:** SAN planning module
- **FC 8015:** Performance and event management module (additional 64 ports)

Sphereon 2026-224

- **FC 4501:** Open Systems Management Software
- **FC 4502:** Rail kit for rack installation (non-desktop)
- **FC 4503:** SANtegrity binding software
- **FC 4522:** Preferred path option
- **FC 4550:** Full volatility option
- **FC 4508:** FlexPort, additional 8-port enablement
- **FC 4510:** Sphereon 2026-224 element manager software
- **FC 4520:** Open Trunking software
- **FC 3750:** EFC management software
- **FC 3765:** 1U rack-mount management server
- **FC 3230:** 2 Gb/s longwave transceiver (LC SFP)
- **FC 3220:** Extended longwave 20 km transceiver
- **FC 3235:** Extended longwave 35 km transceiver
- **FC 5030:** Ethernet hub with 24 ports
- **FC 8000:** EFCM fabric manager
- **FC 8001:** EFCM fabric manager upgrade
- **FC 8005:** Performance and event management module (512 ports)
- **FC 8007:** SAN planning module
- **FC 8015:** Performance and event management module (additional 64 ports)

Sphereon 2026-E12

- **FC 4301:** Full fabric capability activation
- **FC 4304:** FlexPort, additional 4-port enablement
- **FC 4630:** 2Gb/s longwave transceiver (LC SFP)

Sphereon 2027-232

- **FC 3002:** Open Systems Management Server
- **FC 3007:** SANtegrity binding software
- **FC 3008:** FICON management server
- **FC 3020:** Open Trunking software
- **FC 3022:** Preferred path option
- **FC 3050:** Full volatility option
- **FC 3779:** Element manager
- **FC 3026:** Rail kit for FC-512 or other 19” rack
- **FC 3028:** Rail kit for IBM rack (2100, 7014)

- **FC 3204:** 2 Gb/s shortwave SFP transceiver 4-pack
- **FC 3205:** 2 Gb/s (10 km) longwave SFP transceiver 4-pack
- **FC 3230:** 2 Gb/s longwave transceiver (LC SFP)
- **FC 3231:** 2 Gb/s shortwave transceiver (LC SFP)
- **FC 3220:** Extended longwave 20 km transceiver
- **FC 3235:** Extended longwave 35 km transceiver
- **FC 3765:** 1U rack mount management server
- **FC 5030:** Ethernet hub with 24 ports
- **FC 8000:** EFCM fabric manager
- **FC 8001:** EFCM fabric manager upgrade
- **FC 8005:** Performance and event management module (512 ports)
- **FC 8007:** SAN planning module
- **FC 8015:** Performance and event management module (additional 64 ports)

11

Cisco switches and directors

The Cisco portfolio is based on multilayer network and storage intelligence. The switches allow for the configuration of scalable solutions that can help address the need for high performance and reliability in environments ranging from small workgroups to large, integrated enterprise SANs.

The MDS 9000 family introduces SAN capabilities, including Cisco's Virtual SAN (VSAN) technology, designed to enable efficient SAN utilization by creating hardware-based isolated environments within a single SAN fabric. Each VSAN can be zoned as a typical SAN and maintains its own fabric services for added scalability and resilience.

Overall, the MDS 9000 Family is designed for investment protection, flexibility, scalability, robustness, advanced diagnostics, and integration between low cost workgroup, and enterprise SANs.

© Copyright IBM Corp. 2004. All rights reserved.

11.1 Product description

The Cisco Systems MDS 9120 (2061-020), MDS 9140 (2061-040), MDS 9216 (2062-D01) Multilayer Fabric Switches and the MDS 9506 (2062- D04 / T04), MDS 9509 (2062-D07 / T07) Multilayer Directors are available from IBM and Authorized IBM Business Partners. The Cisco MDS 9000 family provides 1 to 2 Gbps Fibre Channel capability and performance with advanced intelligence to help address the demands for security, performance, and manageability required to consolidate geographically dispersed SAN islands into a large enterprise SAN. They are designed to be fully interoperable with other members of the Cisco MDS 9000 family, and are based on a multilayer network and storage intelligence.

The following Cisco products are available as part of the IBM Storage Networking solutions portfolio.

11.1.1 MDS 9120 and 9140 Multilayer Switches

The Cisco MDS 9120 Multilayer Fabric Switch (IBM 2061-020) and Cisco MDS 9140 Multilayer Fabric Switch (IBM 2061-040) are 1 RU (rack-unit) fabric switches that can support 20 or 40 shortwave or longwave SFP fiber optic transceivers. Some of these ports operate with a four to one fanout and are referred to as host optimized ports.

The MDS 9120 has a total of 20 ports. The first group of four ports on the left-hand side are full bandwidth ports, and are identified by a white border. The remaining four groups of ports are host optimized port groups.

Shown in Figure 11-1 is the MDS 9120 switch.

Figure 11-1 MDS 9120 Multilayer Switch (IBM 2061-020)

The MDS 9140 has a total of 40 ports. The first eight ports on the left-hand side are full bandwidth ports, and are identified by a white border. The remaining eight groups of ports are host optimized port groups.

Shown in Figure 11-2 on page 491 is the MDS 9140 switch.

Figure 11-2 MDS 9140 Multilayer Switch (IBM 2061-040)

The switches are configured with dual redundant power supplies either of which can supply power for the whole switch. They also include a hot-swappable fan tray to manage the cooling and airflow for the entire switch.

The 91n0 switches share a common firmware architecture with the Cisco MDS 9500 series of multilayer directors, making it an intelligent and flexible fabric switch.

11.1.2 MDS 9216 Multilayer Switch

The Cisco MDS 9216 Model D01 (IBM 2062-D01) is a 3 RU (rack-unit), 2 slot fabric switch that can support from 16 to 48 shortwave or longwave SFP fiber optic transceivers. These ports fully support either 1 Gbps or 2 Gbps Fibre Channel and are auto-sensing.

Shown in Figure 11-3 is the MDS 9216 switch.

Figure 11-3 MDS 9216 Multilayer Switch (IBM 2062-D01)

The switch is configured with dual redundant 845W AC power supplies, either of which can supply power for the whole switch. It also includes a hot-swappable fan tray with four fans to manage the cooling and airflow for the entire switch. If an individual fan (or fans) within the assembly fails, the Fan Status LED turns red. Individual fans cannot be replaced; however, the fan assembly can be replaced. The MDS 9216 continues to run if the fan assembly is removed, as long as preset temperature thresholds are not exceeded. This allows the fan assembly to be replaced without having to bring the systems down. The switch is designed with side-to-side airflow, which is commonly used in LAN switching environments. Sufficient space between racks is required to provide adequate cooling.

The chassis consists of two slots, the first slot must contain the **Supervisor module.** This provides the control and management functions for the 9216, and includes 16 standard Fibre Channel ports. It contains 2 GB of DRAM and has one internal CompactFlash card that provides 256 MB of storage for the firmware images.

The second slot may contain any one of the modules described in 11.3.2, “Optional modules” on page 505.

The MDS 9216i uses the same backplane as the MDS 9216A, but includes a fixed 14+2 supervisor module to provide 14 full capability target optimized Fibre Channel ports and 2 Gigabit ethernet interfaces. This model will accept any of the optional modules into its second slot.

The layout of the MDS 9216 is illustrated in Figure 11-4.

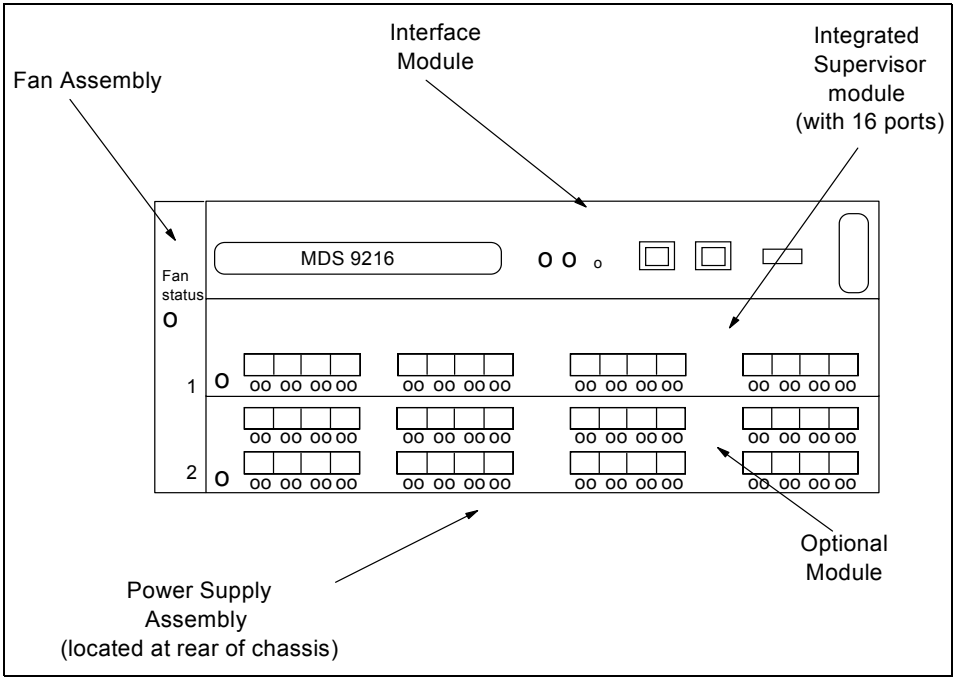


Figure 11-4 Cisco MDS 9216 Multilayer Fabric Switch layout

The interface module, which is located above slot 1, provides the following local and remote management interfaces for the supervisor module:

- **Console:** This has an RJ-45 connection that allows you to perform these tasks:
  - Configure the switch from the CLI

  - Monitor network statistics and errors
  - Configure SNMP agent parameters
  - Distribute software images residing in Flash memory to attached modules
- **10/100 MGMT:** This has a 10/100-Mbps ethernet interface with an RJ-45 connection that provides network management capabilities.

COM2: This connects to an external serial communication device, such as an uninterruptible power supply (UPS).

The MDS 9216 switch shares a common architecture with the Cisco MDS 9000 series of multilayer products, making it an intelligent and flexible fabric switch.

The MDS 9216 product family includes the following features to help provide the highest reliability.

- Provides power-on self testing
- Detects errors, isolates faults and performs parity checking
- Remote diagnostics using Call Home features
- LED displays that summarize the status for the switching modules, supervisor modules, power supply assembly and fan assembly

Physical dimensions

Listed in Table 11-1 are the physical specifications for the MDS 9216.

Table 11-1 Physical specifications for the MDS 9216

Cisco MDS 9216 Multilayer Director (IBM 2062-D01)	
Dimensions	13.3 cm H x 43.9 cm W x 57.6 cm D (5.25 in x 17.25 in x 22.7 in)
Rack Height	3U
Weight (fully configured chassis)	32 kg (170 lb)
Operating environment	
Temperature	0^{o} to 40^{o} C (32^{o} to 104^{o} C)
Relative Humidity	10% to 90%
Power Supplies	845 W AC
Input	D07 Model: 100 to 240 V AC 50-60 Hz nominal

11.1.3 MDS 9506 Multilayer Director

The Cisco MDS 9506 (IBM 2062-D04) is a fourteen RU (rack-unit) Fibre Channel director that can support from 32 to 128 shortwave or longwave SFP fiber optic transceivers. These ports fully support either 1 Gbps or 2 Gbps Fibre Channel and are auto-sensing.

The chassis has six slots, of which two are reserved for dual redundant supervisor modules. The dual supervisor modules provide the logic control for the director and also provide high availability and traffic load balancing capabilities across the director. Either supervisor module can control the whole director, with the standby supervisor module providing full redundancy in the event of an active supervisor failure.

The remaining four slots can contain a mixture of switching modules which provide either 16 or 32 ports per module, 4 or 8 port IP storage services modules and virtualization Caching Services Modules.

The director is configured with dual redundant power supplies either of which can supply power for the whole chassis. It also includes a hot-swappable fan tray that manages the cooling and airflow for the entire director.

Shown in Figure 11-5 is the MDS 9506 Multilayer Director.

Figure 11-5 MDS 9506 Multilayer Director (IBM 2062-D04)

The IBM 2062-T04 product is designed for the telecommunication industry and ships with -48 to -60V DC fed 1900W power supplies. This is the only difference when compared to the 2062-D04.

The Cisco MDS 9506 highlights director class management, serviceability, and availability, including hot swap support and redundancy of active hardware components, and non-disruptive software upgrade support on all active components. Each supervisor module is capable of providing the ability to

automatically restart failed processes. In the event that a supervisor module is reset because of hardware failure or service action, complete synchronization between the active and standby supervisor modules support graceful process failover with no traffic disruption.

The director also provides management and software based availability features such as:

- Non-disruptive software upgrade capability
- Failover of supervisor module code to the redundant supervisor module
- Protection against link failure through the use of PortChannels; this is discussed in “PortChannels” on page 514
- Non-disruptive restart of a failed process running on the same supervisor module

11.1.4 MDS 9509 Multilayer Director

The Cisco MDS 9509 Model D07 (IBM 2062-D07) is a fourteen-rack-unit (14 RU) Fibre Channel director that can support from 32 to 224 shortwave or longwave SFP fiber optic transceivers. These ports fully support either 1 Gbps or 2 Gbps Fibre Channel and are auto-sensing.

Shown in Figure 11-6 on page 495 is the MDS 9509 Multilayer Director.

.

Figure 11-6 MDS 9509 Multilayer Director (IBM 2062-D07)

The chassis has nine slots, of which two are reserved for dual redundant supervisor modules. The dual supervisor modules provide the logic control for

the director and also provide high availability and traffic load balancing capabilities across the director. Either supervisor module can control the whole director, with the standby supervisor module providing full redundancy in the event of an active supervisor failure.

The backplane of the 9509 provides the connectivity for two supervisor modules and up to seven switching modules. As well as the supervisor and switching modules the redundant power supplies and the redundant dual clock modules also plug directly into the backplane.

If one clock module fails, the remaining clock module will take over operation of the director; however this will cause a reset of the director. Although there are dual redundant clock modules, if one needs to be replaced, then an outage will be required, as these modules are not currently hot-pluggable.

The remaining seven slots can contain a mixture of switching modules which provide either 16 or 32 ports per module, 4 or 8 port IP storage services modules and virtualization Caching Services Modules.

The director is configured with dual redundant power supplies either of which can supply power for the whole chassis, the power supplies are hot-swappable and provide self-monitoring functions by reporting their status to the supervisor module. It also includes a hot-swappable fan tray with nine fans to manage the cooling and airflow for the entire director. If an individual fan (or fans) within the assembly fails, the Fan Status LED turns red. Individual fans cannot be replaced; however, the fan assembly can be replaced. It will continue to run if the fan assembly is removed, as long as preset temperature thresholds are not been exceeded. This allows you to swap out a fan assembly without having to bring systems down. The director is designed with a side-to-side airflow, which is commonly used in LAN switching environments. Sufficient space between racks is required to provide adequate cooling.

The IBM 2062-T07 product is designed for the Telecommunication industry and ships with -48 to -60V DC fed 2500W power supplies. This is the only difference when compared to the 2062-D07.

The Cisco MDS 9509 highlights director class management, serviceability, and availability, including hot swap support and redundancy of active hardware components, and non-disruptive software upgrade support on all active components. Each supervisor module is capable of providing the ability to automatically restart failed processes. In the event that a supervisor module is reset because of hardware failure or service action, complete synchronization between the active and standby supervisor modules support graceful process failover with no traffic disruption.

The director also provides management and software based availability features such as:

- Non-disruptive software upgrade capability
- Failover of supervisor module code to the redundant supervisor module
- Protection against link failure through the use of PortChannels; this is discussed in “PortChannels” on page 514
- Non-disruptive restart of a failed process running on the same supervisor module

The 9509 has a 9-slot chassis that is 14-RU high and can be installed inside a standard EIA rack providing from 32 to 224 ports. Of these 9-slots, 2 are reserved for the redundant supervisor modules with the remaining 7-slots available for optional modules (switching modules contain the actual FC ports or SFPs).

The layout of the 9509 is illustrated in Figure 11-7.

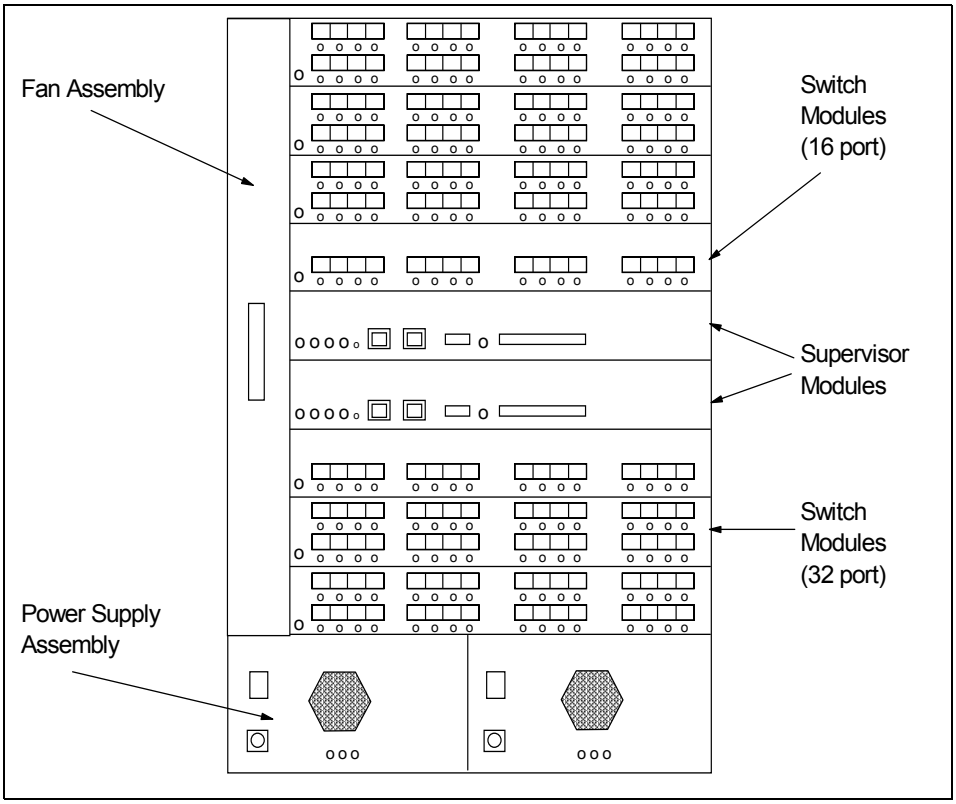


Figure 11-7 Cisco MDS 9509 Multilayer Director layout

Physical dimensions

Listed in Table 11-2 are the physical specifications for a fully configured MDS 9509.

Table 11-2 Physical specifications for the MDS 9509

Cisco MDS 9509 Multilayer Director (IBM 2062-D07)	
Dimensions	62.3 cm H x 43.9 cm W x 46.8 cm D (24.5 in x 17.25 in x 18.4 in)
Rack Height	14U
Depth	55.0 cm (21.6 in)
Weight (fully configured chassis)	78 kg (170 lb)
Operating environment	
Temperature	0^{o} to 40^{o} C (32^{o} to 104^{o} C)
Relative Humidity	10% to 90%
Power Supplies	D07 Model: 2500 W AC
Input	D07 Model: 100 to 240 V AC 50-60 Hz nominal
Output	D07 Model: 1300 W at 100 to 120 V AC 2500 W at 200 to 240 V AC

11.2 Features

In the section below we will discuss the common components of the Cisco MDS 9000 Multilayer products. Many of the components in this section are common to both the 9500 and 9200 series of switches and directors.

11.2.1 Supported attachments

The MDS 9000 family support the following attachment types:

- FICON
- FCP
- FC_AL (including public and private loop support)
- Cascading (attaching multiple switches or directors together)
- Interoperability (attachment to other vendors switches)

The unique architecture of the 9000 Multilayer products allows integration of new transport protocols for greater flexibility. For example, they are designed to support Fibre Channel (FC), Fiber Connection for zSeries channel (FICON), Internet SCSI (iSCSI), Fibre Channel over IP (FCIP) and Gigabit ethernet (GigE).

The IP Services Modules, described in 11.3.2, “Optional modules” on page 505, provide the Gigabit Ethernet (GigE) interfaces, to enable iSCSI and FCIP capabilities for the 9200, and 9500 families.

11.2.2 Port addressing and port modes

The Fibre Channel ports in the Cisco MDS9000 family are addressed with addresses in the form of *f/c<slot>/<port>*, where *<slot>* is the slot number of the line card (1-9), and *<port>* is the port number on the line card (1-32). For example, the first port of the line card in slot 1 is *f/c1/1*, and the seventh port of the line card in slot 3 is *f/c3/7*.

11.2.3 Fibre Channel IDs and Persistent FC_ID

Contrary to other switch manufacturers, there is no direct correlation between physical Fibre Channel ports and Fibre Channel IDs (FCIDs). This is necessary to allow intermixing line cards with different number of ports, while being able to utilize all port addresses, to allow both fabric and loop devices to co-exist, and also to allow switches larger than 256 ports in the future.

The following applies to the FC_ID assignment for any VSAN:

- When an N_Port or NL_Port logs into the switch it is assigned a FC_ID
- N_Ports receive the same FC_ID if disconnected and reconnected to any port within the same switch
- NL_Ports receive the same FC_ID only if reconnected to the same port within the same switch, where the port was originally connected

If the Persistent FC_ID feature is not enabled for a VSAN, the following apply:

- The WWN of the N_Port or NL_Port and the assigned FC_ID are stored in a volatile cache, and are not saved across switch reboots
- The switch preserves the binding of FC_ID to WWN on a best-effort basis
- The volatile cache has room for a maximum of 4000 entries and if the cache gets full, oldest entries are overwritten

If the Persistent FC_ID feature is enabled for a VSAN, the following apply:

- The FC_ID to WWN mapping of the WWNs currently in use is stored to a non-volatile database, and is saved across reboots
- The FC_ID to WWN mapping of any new device connected to the switch is automatically stored into the non-volatile database

- You can also manually configure the FC_ID to WWN mappings if necessary

Important: If you attach AIX or HP-UX hosts to a VSAN, you need to have persistent FCIDs enabled for that VSAN. This is because these operating systems use the FCIDs in device addressing. If the FC_ID of a device changes, the operating system considers it to be a new device, and gives it a new name.

11.2.4 Supported port types

The Fibre Channel ports on all models of the MDS 9000 family provide an auto-sensing 1 or 2 Gbps SFP that use LC connectors. The operating port modes supported are described in the following sections.

Auto Mode

Interfaces configured in the default auto mode are allowed to operate in one of the following modes: F_Port, FL_Port, E_Port, or TE port. The port mode is determined during interface initialization. For example, if the interface is connected to a node (server or disk), it operates in F_Port or FL_Port mode depending on the N_Port or NL_Port mode. If the interface is attached to a third-party switch, it operates in E_Port mode. If the interface is attached to another MDS 9000 switch, it may become operational in TE_Port mode. TL_Ports, SD_ports and ST_ports are not automatically determined during initialization and must be administratively configured.

E_Port

In expansion port (E_Port) mode, an interface functions as a fabric expansion port. This port may be connected to another E_Port to create an ISL between two switches. E_Ports carry frames between switches for configuration and fabric management. They serve as a conduit between switches for frames destined for remote N_Ports and NL_Ports. E_Ports support class 2, class 3, and class F service.

An E_Port connected to another switch may also be configured to form a PortChannel.

F_Port

In fabric port (F_Port) mode, an interface functions as a fabric port. This port may be connected to a node (server, disk or tape) operating as an N_Port. An F_Port can be attached to only one N_Port. F_Ports support class 2 and class 3 service.

FL_Port

In fabric loop port (FL_Port) mode, an interface functions as a fabric loop port. This port may be connected to one or more NL_Ports (including FL_Ports in other switches) to form a public arbitrated loop. If more than one FL_Port is detected on the arbitrated loop during initialization, only one FL_Port becomes operational and the other FL_Ports enter non-participating mode. FL_Ports support class 2 and class 3 service.

Fx_Port

Interfaces configured as Fx_Ports automatically negotiate operation in either F_Port or FL_Port mode. The mode is determined during interface initialization depending on the attached N_Port or NL_Port. This administrative configuration disallows interfaces to operate in other modes — for example, preventing an interface to connect to another switch.

TL_Port

In translative loop port (TL_Port) mode, an interface functions as a translative loop port. It may be connected to one or more private loop devices (NL_Ports). TL_Port mode is specific to Cisco MDS 9000 family switches and have similar properties as FL_Ports. TL_Ports enable communication between private loop devices and one of the following target devices:

- A device attached to any switch on the fabric
- A device on a public loop anywhere in the fabric
- A device on a different private loop anywhere in the fabric
- A device on the same private loop

TL_Ports support class 2 and class 3 services.

TE_Port

In trunking E_Port (TE_Port) mode, an interface functions as a trunking expansion port. It connects to another TE_Port to create an Extended ISL (EISL) between two switches. TE_Ports are specific to the Cisco MDS 9000 family. They expand the functionality of E_Ports to support these features:

- Multiple VSAN trunking
- Transport quality of service (QoS) parameters
- Fibre Channel trace (**`fctrace`**) feature

In TE_Port mode, all frames are transmitted in EISL frame format, which contains VSAN information. Interconnected switches use the VSAN ID to multiplex traffic from one or more VSANs across the same physical link. This feature is referred to as *trunking* in the Cisco MDS 9000 Family.

TE_Ports support class 2, class 3, and class F service.

SD_Port

In switch port analyzer (SPAN) destination port (SD_Port) mode, an interface functions as a SPAN. The SPAN feature is specific to switches in the Cisco MDS 9000 family. It monitors network traffic that passes though a Fibre Channel interface. This monitoring is done using a standard Fibre Channel analyzer (or a similar switch probe) that is attached to an SD_Port. SD_Ports do not receive frames, they merely transmit a copy of the source traffic. The SPAN feature is non-intrusive and does not affect switching of network traffic for any SPAN source ports.

ST_Port

Interfaces configured as ST ports serve as an entry point port in the source switch for a Fibre Channel tunnel. ST ports are specific to remote SPAN (RSPAN) ports and cannot be used for normal Fibre Channel traffic.

Shown in Figure 11-8 is an example of the port types available with the Cisco MDS 9000 family of products.

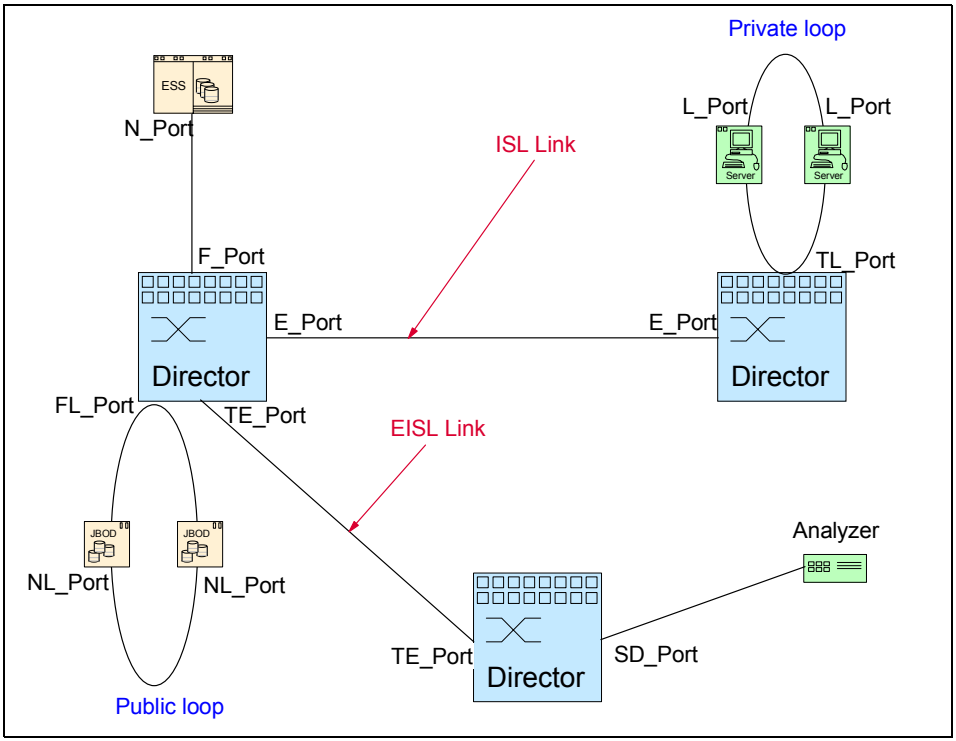


Figure 11-8 Cisco MDS 9000 family port types

SFP transceivers

The ports in the MDS 9000 series can be configured using a mixture of either shortwave (f/c 5230) or longwave (f/c 5240) SFP optic transceivers. Specific configuration options should be checked when ordering.

Listed in Table 11-3 on page 503 are the distance specifications for the SFPs.

Table 11-3 Distance specifications for SFP optics

Optics	Media	Supported distance
2 Gbps - SW, LC SFP	50/125 micron multi-mode	300 m
2 Gbps - SW, LC SFP	62.5/125 micron multi-mode	150 m
2 Gbps - LW, LC SFP	9/125 micron single-mode	10 km

Buffer credits

Each target optimized port supports up to 255 buffer credits and host optimized ports are fixed at 12 buffer credits per port.

11.3 Supervisor module

The supervisor module is the heart of the 9500 series directors as it provides the control and management functions for the director, as well as an integrated crossbar switching fabric. The crossbar fabric provides up to 720 Gbps full duplex switching capacity.

Note: The MDS 9216 uses a different supervisor module which integrates 16 target optimized ports. The function provided by the MDS 9216 supervisor is the same as that described in this section.

The MDS 9500s come standard with two supervisor modules for redundancy and availability. In the event of a supervisor module failing, the surviving module will become active taking over the operation of the director.

11.3.1 Control and management

The supervisor module provides the following control and management features:

- Multiple paths to avoid a single point of failure.
- A redundant central arbiter that provides traffic control and access fairness.

- Non-disruptive restart of a single failing process on the same supervisor. A kernel service running on the supervisor module keeps track of the high availability policy of each process and issues a restart when a process fails. The type of restart issued is based on the process' capability:
  - Warm or stateful (state is preserved)
  - Cold or stateless (state is not preserved)

  If the kernel service is unable to perform a warm restart of the process, it issues a cold restart.
- Non-disruptive switchover from the active supervisor to a redundant standby without loss of traffic.

If the supervisor module has to be restarted, then the secondary supervisor (which is continuously monitoring the primary) takes over. Switchover is non-revertive. Once a switchover has occurred and the failed supervisor has been replaced or restarted, operation does not switch back to the original primary supervisor, unless it is forced to switch back or unless another failure occurs.

Crossbar switching fabric

The MDS 9500 supervisor module provides a crossbar switching fabric that connects all the modules. A single crossbar provides 720 Gbps full duplex speed allowing 80 Gbps bandwidth per switching module.

Dual supervisor configurations provide 1.4 Tbps throughput with a 160 Gbps bandwidth per switching module. This architecture is designed for coping with emerging 10 Gbps interface technologies.

Figure 11-9 shows a picture of the 9500 series supervisor module.

Figure 11-9 MDS 9500 Series supervisor module

Interfaces

The supervisor module provides the following interfaces:

- **Console:** Has an RJ-45 connection that allows you to:

  - Configure the switch from the CLI
  - Monitor network statistics and errors
  - Configure SNMP agent parameters
- **10/100 Mgmt:** Has a 10/100-Mbps ethernet interface with an RJ-45 connection that provides network management capabilities.
- **COM1:** May be connected to a modem for remote CLI access.
- **CompactFlash slot:** Provides a slot for an optional CompactFlash card. The cards can be used for storing additional software images, and configuration, debugging, and syslog information.

11.3.2 Optional modules

The MDS 9200 and 9500 families allow for optional modules, as described in this section, to provide additional port connectivity, IP services or storage virtualization functionality into empty expansion slots.

The MDS 9216 can accept one optional module, while the MDS 9506 can accept four, and the MDS 9509 supports up to a maximum of seven optional modules.

With switching modules, network architects can allocate bandwidth to meet specific application requirements while decreasing the switching footprint which can help lower the overall storage network deployment cost. Optional switching modules are designed to be interchanged or shared between all Cisco MDS 9000 Directors and Fabric Switches.

16 port switching module

The 16 port switching module is designed to deliver high performance for demanding storage networking applications providing up to 64 Gbps of continuous aggregate bandwidth. Autosensing 1 Gbps and 2 Gbps target optimized ports deliver 200 MBps and 255 buffer credits per port.

Note: The 64 Gbps continuous aggregate bandwidth is based on 2 Gbps per port in full duplex mode.

That is: 16 ports @ 2 Gbps (or 200 MBps) in both directions = 64 Gbps

The 16 port module is designed for attaching high-performance servers and storage subsystems as well as for connecting to other switches via ISL connections.

Figure 11-10 on page 506 shows a picture of the 16 port switching module for the Cisco MDS 9000 family.

Figure 11-10 16 port switching module

32 port switching module

The 32 port switching module is designed to deliver an optimal balance of performance and port density. This module provides high line-card port density along with 64 Gbps of total bandwidth and 12 buffer to buffer credits per port. Bandwidth is allocated across eight 4-port groups, providing 4 Gbps (200 MBps) of sustained bandwidth per port-group. This module provides a low-cost means of attaching lower performance servers and storage subsystems to high-performance crossbar switches without requiring ISLs.

Note: The 64 Gbps continuous aggregate bandwidth is based on providing 4 Gbps bandwidth to each port-group (each port-group contains 4-ports).

That is: 8 port-groups @ 4 Gbps in both directions = 64 Gbps

4 Gbps to each port group does not mean that every port has only 1 Gbps bandwidth. The whole port-group has an aggregate bandwidth of 4 Gbps. One port could be transmitting at 2 Gbps with the other ports sharing the remaining bandwidth.

By combining 16 and 32 port switching modules in a single, modular chassis administrators can configure price/performance-optimized storage networks for a wide range of application environments.

Switching modules are designed to be interchanged or shared between all Cisco MDS 9200 Switches and 9500 directors.

Figure 11-11 on page 507 shows the 32 port switching module for the Cisco MDS 9000 family.

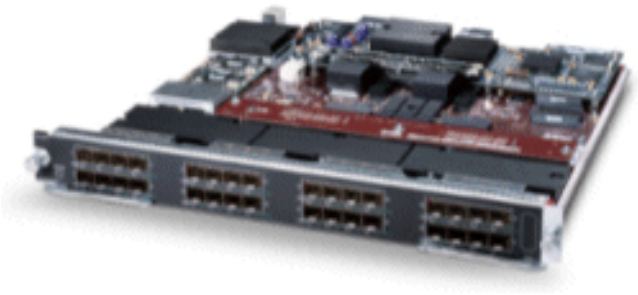


Figure 11-11 32 port switching module

Cisco MDS 9000 14+2 Multi-Protocol Services Module

The Cisco MDS 9000 14+2 Multi-Protocol Services Module is designed to provide 14 Fibre Channel and 2 IP storage interfaces. The 14 Fibre Channel ports are based around the same full rate target optimized ports as the 16 port module, providing all the same operating modes.

The 2 IP storage interfaces provide the same functionality as the IP Services Module architecture, providing hardware compression and security.

Shown in Figure 11-12 is the Cisco MDS 9000 14+2 Multi-Protocol Services Module.

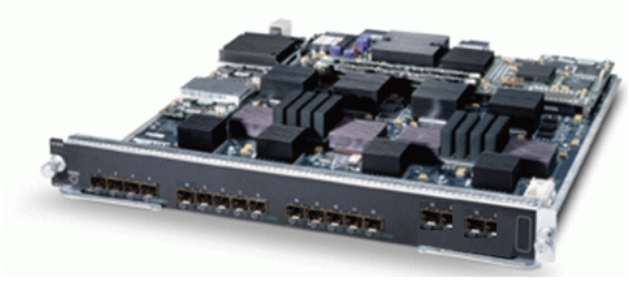


Figure 11-12 Cisco MDS 9000 14+2 Multi-Protocol Services Module

IP Services Module

The IP Services Module is available in two versions, IPS-4 or IPS-8, providing four or eight gigabit ethernet ports, that can support iSCSI and FCIP protocols simultaneously. Since the bit rate of gigabit ethernet is different from the bit rate of Fibre Channel, the card only supports the Tri-rate SFPs.

The 8-port IP Services module is shown in Figure 11-13.

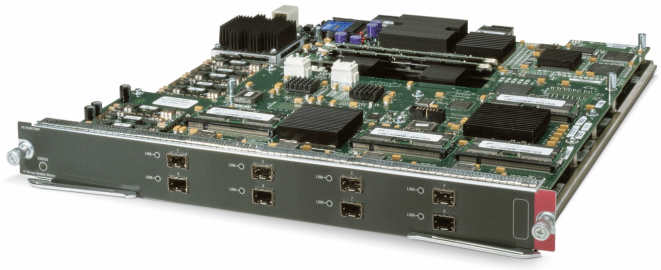


Figure 11-13 8-port IP Services Module

Ports configured to run FCIP

The ports configured for FCIP can support up to three virtual ISL connections (FCIP tunnels). This way you can transport the Fibre Channel traffic transparently over an IP network between two FCIP capable switches. Each virtual ISL connection acts as a normal Fibre Channel ISL or EISL.

To use FCIP, you need to purchase the “FCIP Activation for 8-port IP Services Line Card feature for every 8-port IP Line Card that needs to support FCIP. The feature codes are f/c 2209 for the MDS 9216, and f/c 2210 for the MDS 9506 and MDS9509.

Ports configured to run iSCSI

Ports configured to run iSCSI work as a gateway between iSCSI hosts and Fibre Channel attached targets. The module terminates iSCSI commands and issues new Fibre Channel commands to the targets.

The Cisco Fabric Manager is used to discover and display iSCSI hosts. These iSCSI hosts are bound to assigned WWNs creating a static relationship enabling:

- Zoning of iSCSI initiators
- Accounting against iSCSI initiators
- Topology mapping of iSCSI initiators

Caching Services Module

The Caching Services Module is used by IBM to provide storage virtualization within the SAN to provide a pool of storage from disparate storage devices. The imbedded IBM TotalStorage SAN Volume Controller software uses a cache based, clustered architecture to provide high availability, scalability, and the performance required for storage environments.

As this virtualization solution is within the switch, it does not use any of the Fibre Channel port interfaces to connect.

Two Caching Services Modules are required for the IBM TotalStorage SAN Volume Controller storage software for Cisco MDS 9000 to provide a fully redundant solution.

Further information on the IBM TotalStorage SAN Volume Controller storage software for Cisco MDS 9000 can be found in the *Implementing the IBM TotalStorage SAN Volume Controller Storage Software on the Cisco MDS 9000*, SG24-7059.

The Caching Services Module is shown in Figure 11-14.

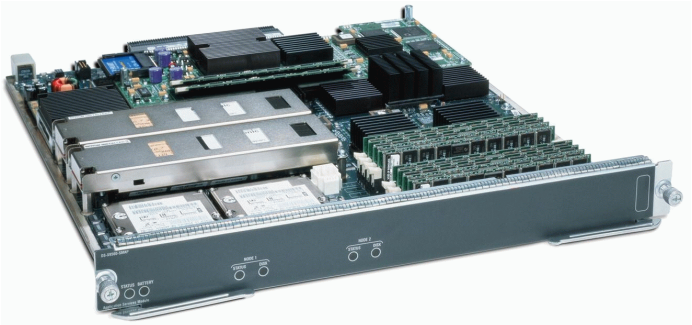


Figure 11-14 Caching Services Module

Cisco MDS 9000 Family Port Analyzer Adapter

The Cisco MDS 9000 Family Port Analyzer Adapter is a stand-alone product which enables advanced debugging and performance analysis of MDS 9000 fabrics.

The Cisco MDS 9000 Family Port Analyzer Adapter encapsulates the Fibre Channel frames coming from the MDS 9000 SPAN port and delivers them to standard ethernet frames over a 1000base-T ethernet interface. The frames can then be analyzed using the freeware *ethereal* network protocol analyzer software. This allows cost-effective monitoring of the Fibre Channel traffic.

The Cisco MDS 9000 Family Port Analyzer Adapter-2 is shown in Figure 11-15.

Figure 11-15 Cisco MDS9000 Port Analyzer Adapter -2

Since the actual data within the Fibre Channel frame is often not relevant for protocol analysis, the Cisco MDS 9000 Family Port Analyzer Adapter can also truncate the Fibre Channel frames. The truncate modes available are described in Table 11-4 on page 510. The modes can be configured with a four-position DIP switch located in the rear of the adapter.

Table 11-4 Cisco MDS9000 Port Analyzer Adapter truncate modes

Mode	Description
No truncate	Fibre Channel frames are passed without any modification to the payload
Ethernet truncate	Fibre Channel frames are truncated to 1496 bytes to fall within the maximum ethernet frame size
Shallow truncate	Fibre Channel frames are truncated, if the payload of the frame is more than 256 bytes
Deep truncate	Fibre Channel frames are truncated, if the payload of the frame is more than 64 bytes
Management	Only fixed 288 byte ethernet frames that contain internal debug information are transmitted

11.4 MDS 9000 SAN-OS

The Cisco SAN-OS is the operating system running within the MDS 9000 supervisor and modules to enable the multilayer functionality of the products. While the SAN-OS installable files are specific to each MDS 9000 platform i.e. 9100, 9200 and 9500; the standard features provided by the SAN-OS are common and include:

- FCP
- iSCSI
- Fabric Manager
- VSANs
- Zoning
- FCC
- Virtual Output Queuing
- Diagnostics (SPAN, RSPAN etc.)
- SNMPv3
- SSH
- SFTP
- RBAC
- Radius
- High Availability

- Port Channels
- RMON
- Call home
- TACACS+
- FDMI
- SMI-S (XML-CIM)
- iSNS Client
- iSNS Server
- WWN based
- IPS ACLs

11.4.1 Licensed feature packages

In addition to the standard features provided in the SAN-OS, there are also four optional licensed feature packages to address different enterprise requirements. The four optional licensed packages are:

- SAN Extension package
- Enterprise Package
- Fabric Manager Server
- Mainframe Package

These packages are licensed at either the switch level or module level, currently all packages except the SAN extension package are switch based, a SAN extension package license is required per IP Services Module.

SAN Extension Package

The SAN Extension Package optional license enables:

- FCIP Protocol
- FCIP Compression (S/W)
- FCIP Write Acceleration

Enterprise Package

The Enterprise Package optional license enables:

- LUN-Zoning
- Read-only Zones
- Port Security
- VSAN Roles
- Host/Switch Authentication
- QoS
- Inter-VSAN Routing

Fabric Manager Server

The Fabric Manager Server (FMS) package allows centralized discovery for those switches connected via ISLs, and roaming profiles, eliminating manual effort to maintain consistent topology maps and customized definitions across multiple client stations. It also adds advanced management and troubleshooting capabilities to the MDS 9000 fabric, including historical performance management.

The Fabric Manager Server package should be installed on a dedicated server that is always available. It will allow up to 16 clients workstations to connect simultaneously, and is capable of managing more than one fabric.

Note: SAN-OS 1.3(2a) or later is required for Fabric Manager Server.

Shown in Table 11-5 is the comparison between the Standard Fabric Manager (without FM Server License) and the additional features provided with the Fabric Manager Server optional license installed.

Table 11-5 Fabric Manager standard and Server comparison

Feature	Standard Fabric Manager	Fabric Manager Server
Switch embedded Java application	X	X
FC fabric virtualization	X	X
Fabric, Device and Summary views	X	X
Port, switch and fabric level configuration	X	X
MDS event and security management	X	X
Configuration wizards	X	X
MDS configuration analysis tools	X	X
Network diagnostic and troubleshooting tools	X	X
Real-time performance monitoring	X	X
Centralized management server	-	X
Continuous health and event monitoring	-	X
Historical performance monitoring and reporting	-	X
Cisco traffic Analyzer integration	-	X
Roaming user profiles	-	X

Mainframe Package

Applying the Mainframe Package optional license enables all FICON requirements with a single license key. The mainframe package optional license enables:

- FICON w/ CUP
- Intermixing-VSANs
- Switch Cascading
- Fabric Binding

Note: A license is required for each switch participating in a FICON cascaded fabric.

11.4.2 PortChannels

PortChannels allow users to aggregate up to sixteen physical ISLs into a single logical bundle, designed to optimize bandwidth utilization and resilience between

switches. The group of Fibre Channel ISLs designated to act as a PortChannel may consist of any port on any 16 Port switching module within the MDS 9000 chassis, allowing the overall PortChannel to remain active upon failure of one or more ports, or failure of one or more switching modules. These PortChannels:

- Increase the aggregate bandwidth on an ISL or EISL by distributing traffic among all functional links in the channel.
- Load balance across multiple links and maintains optimum bandwidth utilization. Load balancing is based on a source ID (SID), destination ID (DID), and optionally the originator exchange ID (OX ID) that identify the flow of the frame.
- Provide high availability on an ISL. If one link fails, traffic previously carried on this link is switched to the remaining links. If a link goes down in a PortChannel, the upper protocol is not aware of it. To the upper protocol, the link is still there, although the bandwidth is diminished. The routing tables are not affected by a link failure. PortChannels can contain up to 16 physical links and can span multiple modules for added high availability.

Shown in Figure 11-16 is a diagram representing ISLs and PortChanneling.

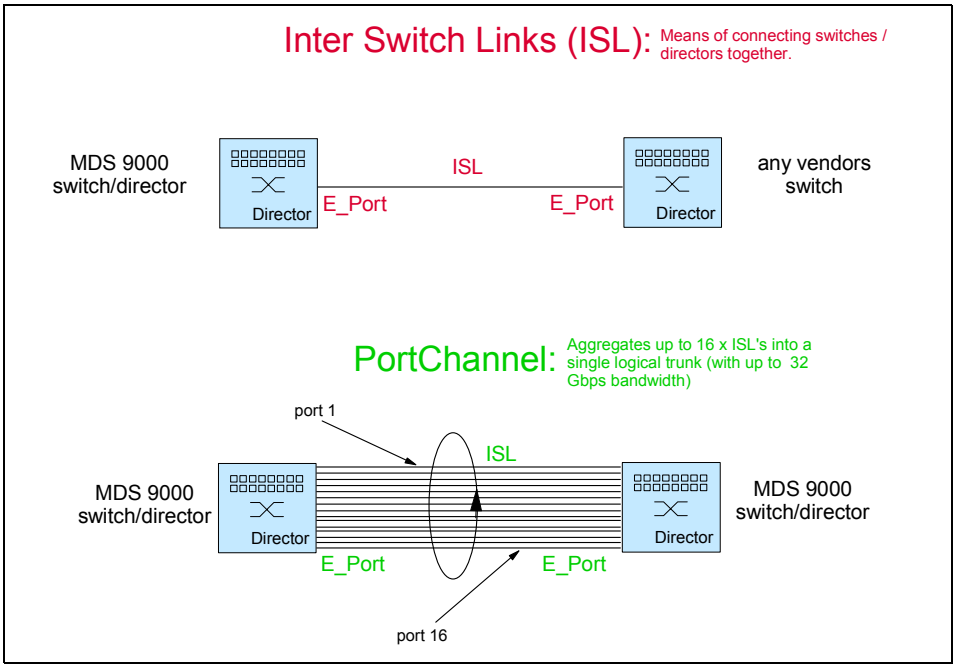


Figure 11-16 PortChannels and ISLs on the Cisco MDS 9000 family of switches

11.4.3 Trunking

The Cisco MDS 9000 family uses the term *trunking* to refer to an ISL link that carries one or more VSANs. Trunking ports receive and transmit EISL frames. EISL frames carry an EISL header containing the VSAN information. Once EISL is enabled on an E_Port, that port becomes a TE_Port.

Trunking is also referred to as VSAN trunking, as it only applies to a VSAN. If a trunking enabled E_Port is connected to another vendor's switch, the trunking protocol ensures that the port will operate as a standard E_Port.

Shown in Figure 11-17 is a diagram of trunking. This also demonstrates how a combination of PortChannels and trunking can be used to create an aggregate bandwidth of up to 32 Gbps between switches.

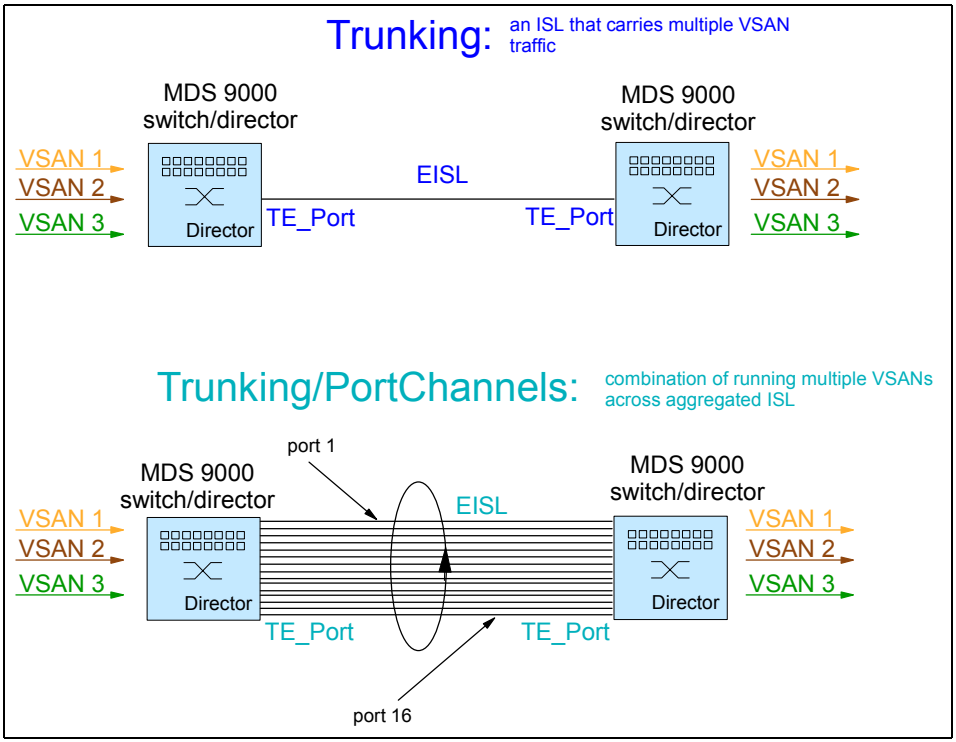


Figure 11-17 Trunking and PortChanneling

11.4.4 Quality of Service

The switches provide low and high priority quality of service (QoS) queues. While time-critical traffic is marked as high priority traffic, all other traffic is assigned to the default low priority queue.

Quality of Service offers the following advantages:

- Guarantees absolute and relative bandwidth choices
- Provides latency to reduce frame loss
- Prioritizes transactional traffic over bulk traffic
- Supports multiple VSANs on the same fabric by guaranteeing bandwidth and latency available to each VSAN

The Cisco MDS 9000 family supports QoS for internally and externally generated control traffic. Within a switch the control traffic is sourced to the supervisor module and is treated as a high priority frame. A high priority status provides absolute priority over all other traffic and is assigned in the following cases:

- Internally generated time-critical control traffic (generally Class F frames).
- Externally generated time-critical control traffic entering a switch in the MDS 9000 range from another vendor's switch. High priority frames originating from other vendor switches retain the priority as they enter a switch in the MDS 9000 family.

By default, the QoS feature for control traffic is enabled but can be disabled if required.

11.4.5 Congestion control

Forward Congestion Control (FCC) is a Cisco proprietary flow control mechanism that alleviates congestion on Fibre Channel networks. FCC reduces the congestion in the fabric without interfering with the standard Fibre Channel protocols. The FCC protocols increases the granularity and the scale of congestion control applied to any class of traffic.

Any switch in the network can detect congestion for an output port. The switches sample frames from the congested queue and generate messages about the congestion level upstream toward the source of the congestion. The switch closest to the source, with FCC enabled, performs one of two actions:

- Forwards the frames as other vendor switches do.
- Limits the flow of frames from the port causing the congestion.

This is illustrated in Figure 11-18.

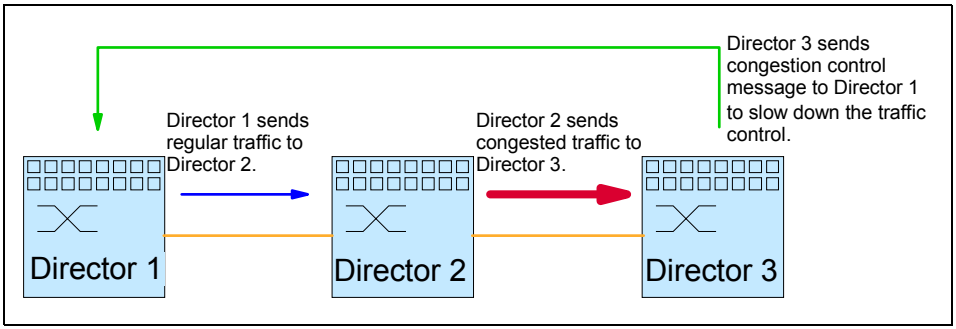


Figure 11-18 Forward Congestion Control

By default, the FCC protocol is disabled. You can enable the protocol globally for all the VSANs configured in the switch, or selectively enable or disable it for each VSAN.

Congestion control methods

With FCC enabled, these are the different congestion control methods:

- **Path quench control:** Reduces severe congestion temporarily by slowing the source to the whole path in the fabric.
- **Edge quench control:** Provides feedback to the source about the rate at which frames should be entered into the network (frame intervals).

FCC process

When a node in the network detects congestion for an output port, it generates an edge or a path quench message. These frames are identified by the Fibre Channel destination ID (DID) and the source ID (SID). A switch from other vendors simply forwards these frames.

Any receiving switch in the Cisco MDS 9000 family handles frames in one of these ways:

- It forwards the frame.
- It limits the rate of the frame flow in the congested port.

Behavior of the flow control mechanism differs, based on the Fibre Channel DID:

- If the Fibre Channel DID is directly connected to one of the switch ports, the input rate limit is applied to that port.
- If the destination of the edge quest frame is a Cisco domain or the next hop is a Cisco MDS 9000 family switch, the frame is forwarded.
- If neither of these conditions is true, then the frame is processed in the port going towards the FC DID.

All switches (including the edge switch) along the congested path process path quest frames. However, only the edge switch processes edge quest frames. The FCC protocol is implemented for each VSAN and can be enabled or disabled on a specified VSAN or for all VSANs at the same time.

11.4.6 IP services

By installing the IP Services Module into the Cisco MDS 9000 family it will provide support the following IP services:

- **FCIP:** Fibre Channel-over-IP — used to tunnel Fibre Channel frames over TCP/IP connections
- **iSCSI**: Internet SCSI — allows storage to be accessed over a TCP/IP network as though it was locally attached

FCIP allows for extending the fabric over an existing TCP/IP network, while iSCSI is used by hosts without Fibre Channel HBAs to talk directly to storage devices in the SAN. The host does require that an iSCSI driver be loaded to enable the protocol over its existing ethernet connection.

11.4.7 Call home

The Cisco MDS 9000 family includes a call home facility that detects any failure alerts and forwards them with the relevant information via e-mail.

11.4.8 Virtual SAN

The MDS 9000 SAN Fabric family introduces Cisco's Virtual SAN (VSAN) technology, offering the capability to overlay multiple hardware enforced virtual fabric environments within a single physical fabric infrastructure. Each VSAN contains separate (dedicated) fabric services designed for enhanced scalability, resilience, and independence among storage resource domains.

This is especially useful in segregating service operations and failover events between high availability resource domains allocated to different VSANs. Each VSAN contains its own complement of hardware-enforced zones, dedicated fabric services, and management capabilities, just as if the VSAN were configured as a separate physical fabric.

Therefore, VSANs are designed to allow more efficient SAN utilization and flexibility, because SAN resources may be allocated and shared among more users, while supporting secure segregation of traffic and retaining independent control of resource domains on a VSAN-by-VSAN basis.

VSANs offer the following features:

- **Traffic isolation:** Traffic is contained within VSAN boundaries and devices reside only in one VSAN thus ensuring absolute separation between user groups, if desired.
- **Scalability:** VSANs are overlaid on top of a single physical SAN. The ability to create several logical VSAN layers increases the scalability of the SAN.
- **Per-VSAN fabric services:** Replication of fabric services on a per-VSAN basis provides increased scalability and availability.
- **Redundancy:** Several VSANs created on the same physical SAN ensure redundancy. If one VSAN fails, redundant protection is provided by a configured backup path between the host and the switch.
- **Ease of configuration:** Users can be added, moved, or changed between VSANs without changing the physical structure of a SAN. Moving a device from one VSAN to another only requires configuration at the port level, not at a physical level.
- **Shared topology:** Multiple VSANs can share the same physical topology.
- **Same FCIDs:** The same FCIDs can be assigned to a host in another VSAN, thus increasing VSAN scalability.
- **Required protocols:** Every instance of a VSAN runs all required protocols such as FSPF, domain manager, and zoning.
- **Independence:** Fabric-related configurations in one VSAN do not affect the associated traffic in another VSAN.
- **Containment:** Events causing traffic disruptions in one VSAN are contained within that VSAN and are not propagated to other VSANs.
- **Isolation:** No communication is possible between VSANs.

Figure 11-19 represents a typical SAN environment that has a number of servers, each with multiple paths to the SAN. The SAN in this case consists of a Fibre Channel director attached to a disk and tape subsystem.

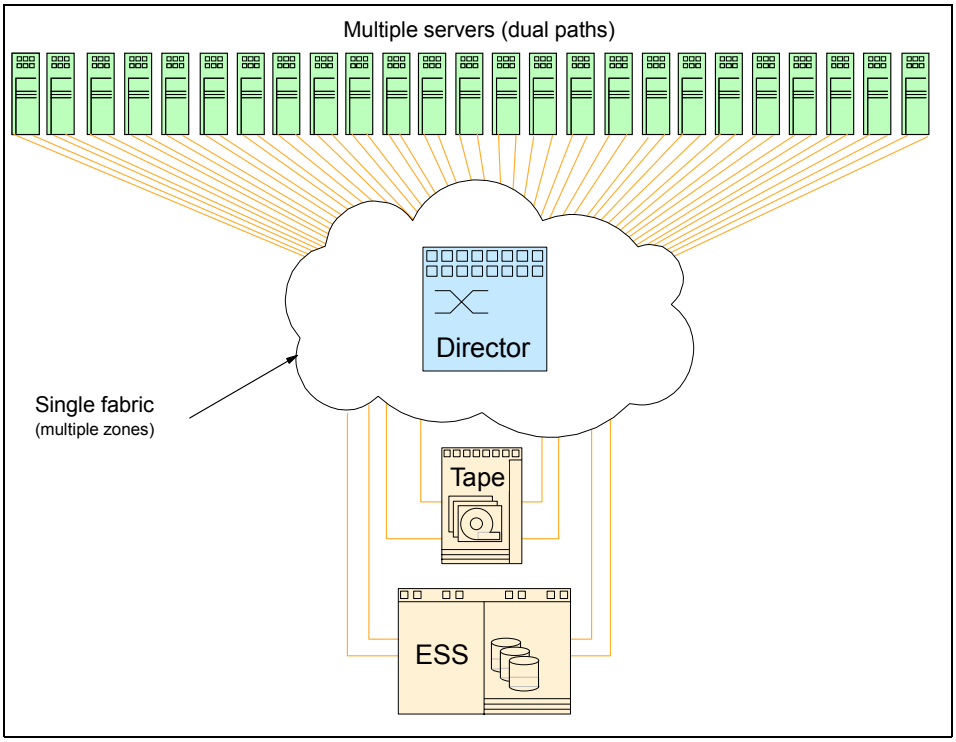


Figure 11-19 Traditional SAN

In Figure 11-20 we show how the same scenario is implemented using Cisco's Virtual SAN.

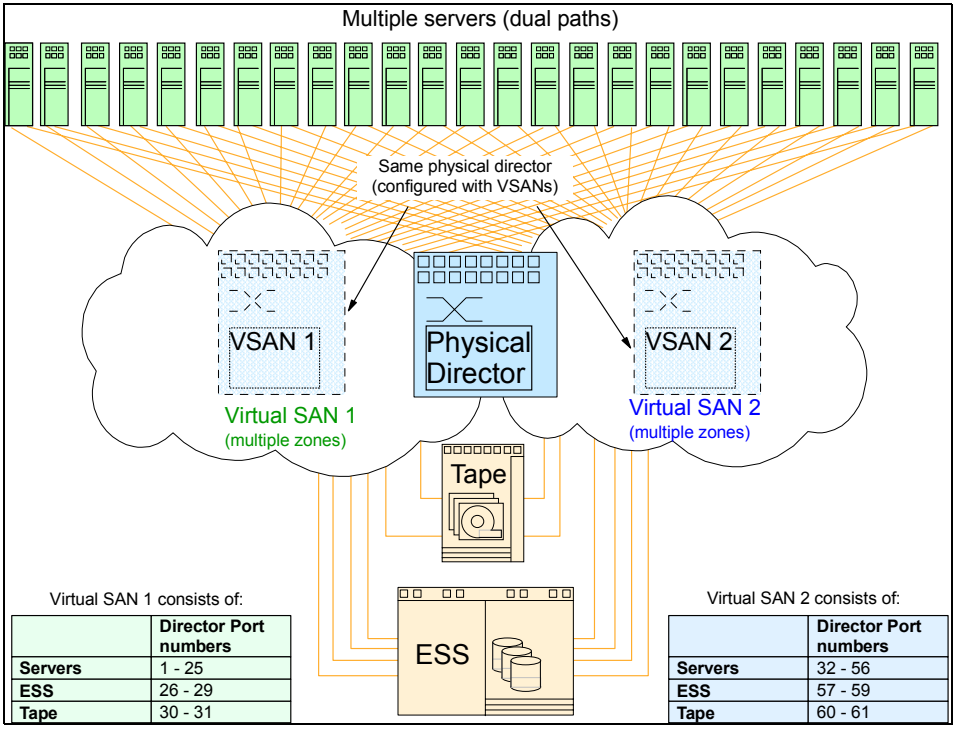


Figure 11-20 Virtual SAN

In this example, the servers are still connected to the SAN, but the SAN consists of a single 9509 which is attached to the same disk and tape subsystems. In this case we have configured the first 31 ports in the director into a Virtual SAN (called Virtual SAN 1) and the second 31 ports into another virtual SAN (called Virtual SAN 2). The servers have a connection to each virtual SAN thereby providing a solution that consists of multiple SAN fabrics.

The virtual SANs cannot communicate with each other, they appear to be totally separate SANs. They have their own FSPF tables, domain manager and zoning requirements. Any traffic disruption in one virtual SAN will have no impact on the other virtual SAN. A port cannot belong in multiple VSANs.

VSANs versus zoning

Shown in Table 11-6 are the main differences between a zone and a VSAN.

Table 11-6 VSAN versus zoning

VSANs	Zones
VSANs equal SANs with their own routing, naming and zoning protocols.	These protocols are not always available on a per-zone basis.
VSANs can contain multiple zones.	Zones are always contained within a VSAN - they cannot span a VSAN.
VSANs limit unicast, multicast and broadcast traffic.	Zones limit unicast traffic.
Membership is defined using VSAN ID to Fx ports.	Membership is typically defined using WWN or port number (Fx).
HBAs may only belong to a single VSAN - that is the VSAN associated with the Fx port.	HBAs can belong in multiple zones.
VSANs enforce membership at each E_Port, source port and destination port.	Zones enforce membership only at the source and destination ports.
VSANs encompass the entire fabric.	Zones typically contain a subset of the fabric.

Default and isolated VSANs

Up to 1024 VSANs can be configured on a physical SAN. Of these, one is the default VSAN (VSAN 1) and another is an isolated VSAN (VSAN 4094). User-specified VSAN IDs range from 2 to 4093.

Default VSAN

The factory settings for switches in the Cisco MDS 9000 family have only the default VSAN 1 enabled. If you do not need more than one VSAN for a switch, use this default VSAN as the implicit parameter during configuration. If no VSANs are configured, all devices in the fabric are considered part of the default VSAN. By default, all ports are assigned to the default VSAN.

Isolated VSANs

VSAN 4094 is an isolated VSAN. All non-trunking ports are transferred to this VSAN when the VSAN to which they belong is deleted. This avoids an implicit transfer of ports to the default VSAN or to another configured VSAN. All ports in the deleted VSAN are isolated (disabled).

VSAN membership

Port VSAN membership on the switch is assigned on a port-by-port basis. By default each port belongs to the default VSAN. Trunking ports have an associated list of VSANs that are part of an allowed list.

VSAN attributes

VSANs have the following attributes:

- **VSAN ID:** The VSAN ID identifies the VSAN as the default VSAN (VSAN 1), user-defined VSANs (VSAN 2 to 4093), and the isolated VSAN (VSAN 4094).
- **State:** The administrative state of a VSAN can be configured to an active (default) or suspended state. Once VSANs are created, they may exist in various conditions or states.
  - The active state of a VSAN indicates that the VSAN is configured and enabled. By enabling a VSAN, you activate the services for that VSAN.
  - The suspended state of a VSAN indicates that the VSAN is configured but not enabled. If a port is configured in this VSAN, it is disabled. Use this state to deactivate a VSAN without loosing the VSAN's configuration. All ports in a suspended VSAN are disabled. By suspending a VSAN, you can preconfigure all the VSAN parameters for the whole fabric and activate the VSAN immediately.
- **VSAN name:** This text string identifies the VSAN for management purposes. The name can be from 1 to 32 characters long and it must be unique across all VSANs. By default, the VSAN name is a concatenation of VSAN and a four-digit string representing the VSAN ID. For example, the default name for VSAN 3 is VSAN0003.
- **Load balancing attributes:** These attributes indicate the use of the source-destination ID (src-dst-id) or the originator exchange OX ID (src-dst-ox-id, the default) for load balancing path selection.

Operational state of a VSAN

A VSAN is in the operational state if the VSAN is active and at least one port is up. This state indicates that traffic can pass through this VSAN. This state cannot be configured.

Deleted VSAN

When an active VSAN is deleted, all of its attributes are removed from the running configuration.

VSAN related information is maintained by the system software:

- VSAN attributes and port membership details are maintained by VSAN manager. This feature is affected when you delete a VSAN from the

configuration. When a VSAN is deleted all the ports in that VSAN are made inactive and the ports are moved to the isolated VSAN. If the same VSAN is recreated, the ports do not automatically get assigned to that VSAN. You must explicitly reconfigure the port VSAN membership.

- VSAN-based runtime (name server), zoning and configuration (static route) information is removed when the VSAN is deleted.
- Configured VSAN interface information is removed when the VSAN is deleted.

Inter VSAN Routing

Inter VSAN Routing (IVR) is available when the Enterprise Package license has been applied to a switch running v1.3(2a) SAN-OS or above.

IVR can be used to allow data traffic to flow between VSANs while maintaining the VSAN segregation as no management data is passed. This proves useful, for example, when a host defined in one VSAN is required to have access to a tape drive defined in another VSAN. This feature reduces the amount of required hardware required to meet the needs for multiple systems.

An IVR is defined in a similar manner to normal zoning within a VSAN, although instead of working within a VSAN and performing the zoning definitions, we work from the IVR group, and create an IVR zoneset which can be activated or deactivated without affecting the VSANs.

In Figure 11-21 we show how the same scenario is implemented using Cisco's Inter VSAN routing.

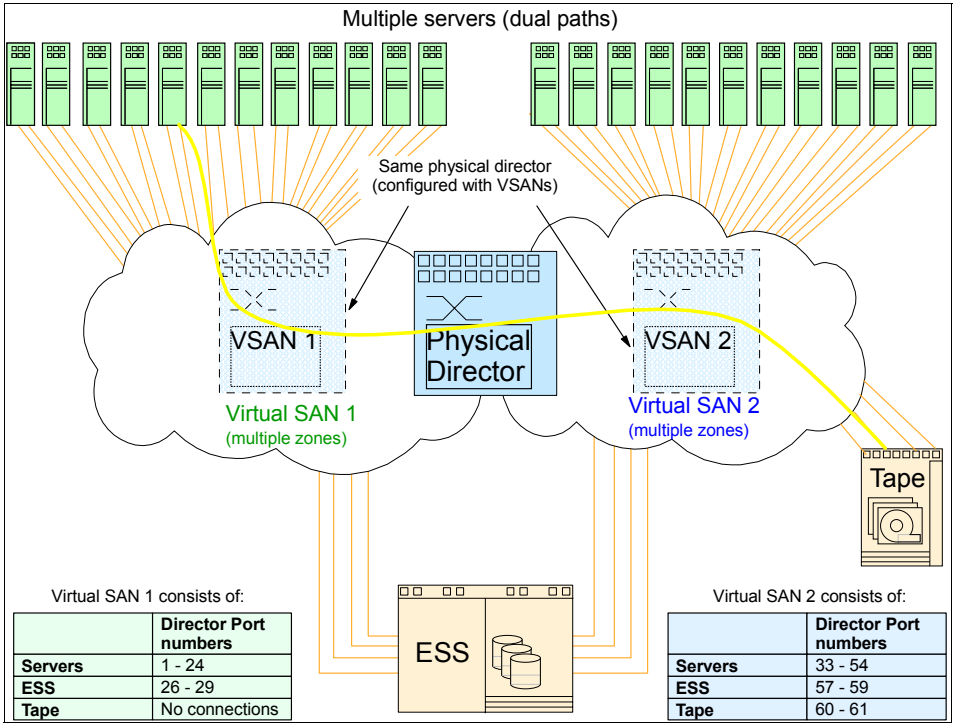


Figure 11-21 Inter VSAN Routing

In this example, there are two groups of servers connected to different VSANs, within a single MDS 9509 Director. The disk has paths defined to both VSAN 1 and VSAN 2, although the requirement is for all servers to access all tape drives. In this case we have configured the first 29 ports in the director into VSAN 1 and the second 31 ports into VSAN 2.

The VSANs cannot communicate with each other, and they appear to be totally separate SANs. Although by defining an IVR zone we can allow a data connection from a server in VSAN 1 through to a tape drive in VSAN 2. No management data is passed over this connection and any disruptions in one VSAN will still have no impact on the other VSAN.

11.5 Security management

The Cisco MDS 9000 family of switches offer strict and secure switch management options through switch access security, user authentication, and role-based access.

11.5.1 Switch access security

Each switch can be accessed through the CLI or SNMP.

- **Secure switch access:** Available when you explicitly enable Secure Shell (SSH) access to the switch. SSH access provides additional controlled security by encrypting data, user IDs, and passwords. By default, Telnet access is enabled on each switch.
- **SNMP access:** SNMPv3 provides built-in security for secure user authentication and data encryption.

11.5.2 User authentication

A strategy known as authentication, authorization, and accounting (AAA) is used to verify the identity of, grant access to, and track the actions of remote users. The RADIUS protocol provides AAA solutions.

Based on the user ID and password combination provided, switches perform local authentication using a local database or remote authentication using the RADIUS server(s). A global, preshared, secret key authenticates communication between the RADIUS client and server. This secret key can be configured for all RADIUS servers or for only a specific RADIUS server. This kind of authentication provides a central configuration management capability.

Role-based access

Role-based access assigns roles or groups to users and limits access to the switch. Access is assigned based on the permission level associated with each user ID. Your administrator can provide complete access to each user or restrict access to specific read and write levels for each command.

SNMP and CLI access rights are organized by roles. Each role is similar to a group. Each group of users has a specific role, and the access for that group can be enabled or disabled.

User authentication

Authentication is the process of verifying the identity of the person managing the switch. This identity verification is based on the user ID and password combination provided by the person trying to manage the switch. Cisco MDS

9000 family switches allow you to perform local authentication (using the lookup database) or remote authentication (using one or more RADIUS servers).

For each management path (console or Telnet and SSH), you can enable only one of three options — local, RADIUS, or none. The option can be different for each path.

Local authentication

The system maintains the user name and password locally and stores the password information in encrypted form. You are authenticated based on the locally stored information, as shown in Figure 11-22.

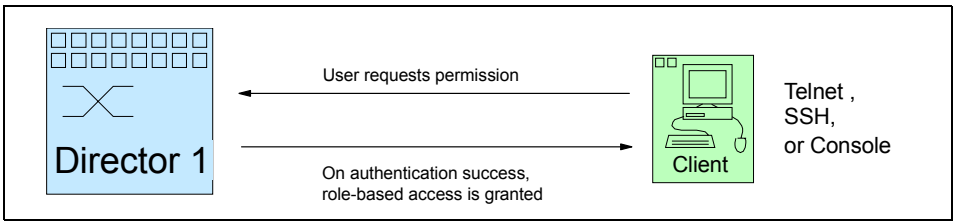


Figure 11-22 Security with local authentication

RADIUS authentication

Cisco MDS 9000 Family switches can provide remote authentication through RADIUS servers. You can also configure multiple RADIUS servers, and each server is tried in the order specified.

RADIUS protocols support one-time password (OTP) schemes that all switches can make use of for authentication purposes.

The use of RADIUS servers in the authentication process is shown in Figure 11-23.

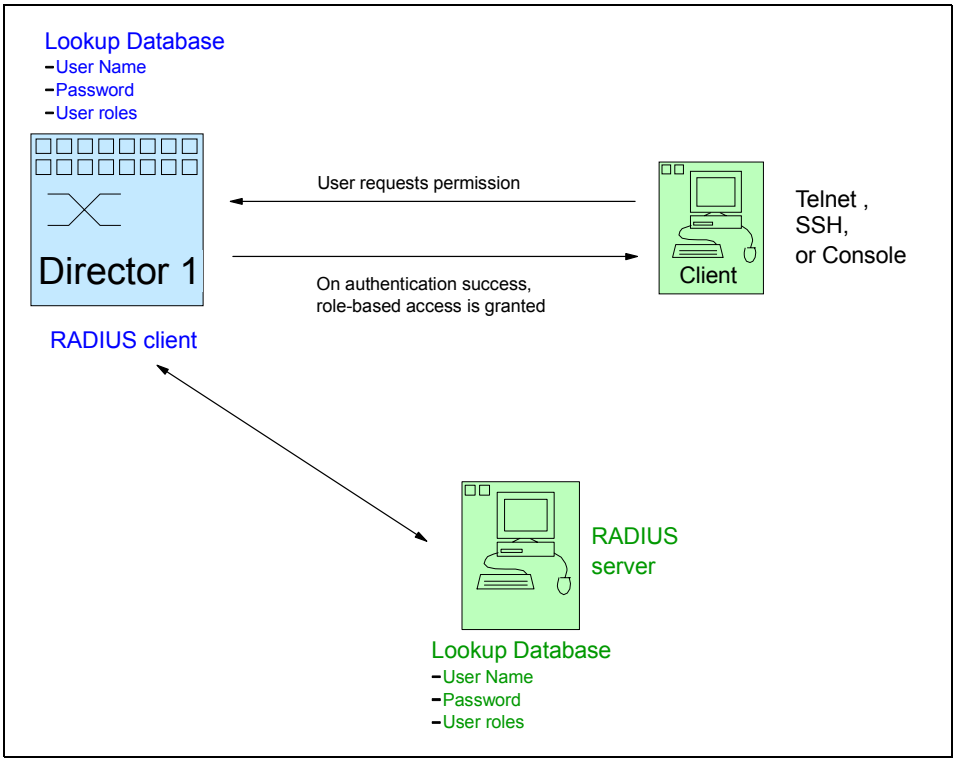


Figure 11-23 Security with RADIUS server

Role-based authorization

By default, two roles exist in all switches:

- **Network operator (network-operator):** Has permission to view the configuration only. The operator cannot make any configuration changes.
- **Network administrator (network-admin):** Has permission to execute all commands and make configuration changes. The administrator can also create and customize up to 64 additional roles.

The two default roles cannot be changed or deleted. Vendor-specific attributes (VSA) contain the user profile information used by the switch. To use this option, configure the VSANs on the RADIUS servers.

Accounting

Accounting refers to the log that is kept for each management session in a switch. This information may be used to generate reports for troubleshooting

purposes and user accountability. Accounting can be implemented locally and remotely (using RADIUS).

11.6 Troubleshooting features

The Cisco MDS 9000 family of switches and directors provide a comprehensive range of diagnostic and troubleshooting features. The components within the chassis have status LEDs that can be used to determine the status of the various components including fans, power supplies, switching modules and supervisor modules. As well as the LEDs, the MDS 9000 family provide extensive software tools that assist with the troubleshooting process some of these are discussed in the following section. For more detailed information, refer to:

- *Cisco MDS 9000 Family Fabric Manager Switch Configuration Guide*, 78-16493-01

11.6.1 Troubleshooting with Fabric Manager

In this section we describe some of the tools provided by the Fabric View and Device View that can be used to verify and troubleshoot fabric connectivity and switch configuration issues.

Switch health option

Allows you to determine the status of the components of a specific switch. The Switch Health Analysis window will display any problems affecting the selected switches.

Analyzing end-to-end connectivity

This option can be used to check connectivity among devices within the SAN fabric. The tool checks to see that every pair of end devices in an active zone can talk to each other.

This is achieved using the **`FC Ping`** and **`Traceroute`** commands which have been modified to handle Fibre Channel networks.

The End-to-End Connectivity Analysis window displays the selected end points of the switch and to which end point each is attached, and the source and target ports used to connect it.

The output shows all the requests which have failed. Descriptions possible are:

- **Ignoring empty zone:** No requests are issued for this zone.
- **Ignoring zone with single member:** No requests are issued for this zone.

- **Source/Target are unknown:** No name server entries exist for the ports or the ports have not been discovered during discovery.
- **Both devices are on the same switch:** The devices are not redundantly connected.
- **No paths exist:** (self-explanatory).
- **Only one unique path exists:** (self-explanatory).
- **VSAN does not have an active zone set:** (self-explanatory).
- **Average time (in microseconds):** The latency value was more than the threshold supplied.

Analyze switch fabric configuration

This option enables you to analyze the configuration of a switch by comparing the current configuration to a specific switch or to a policy file. You can save a switch configuration to a file and then compare all switches against the configuration in this file.

Analyzing the results of a zone merge

You can use the Zone Merge option on the Fabric View Troubleshooting menu to determine if two connected switches have compatible zone configurations.

Other troubleshooting tools

Other troubleshooting tools include:

- **Traceroute command:** Used to verify the connectivity between two end devices that are currently selected on the Map pane.
- **Device Manager:** Launch the Fabric Device Manager for the switch selected on the Map pane.
- **Command Line Interface:** Open a Telnet or SSH session for the switch selected on the Map pane.

11.6.2 Monitoring network traffic using SPAN

The Cisco MDS 9000 family provides a feature called the switch port analyzer (SPAN). As mentioned in 11.2.4, “Supported port types” on page 500, the SPAN or SD_Ports allow us to monitor network traffic through the Fibre Channel interface.

Traffic through any Fibre Channel interface can be replicated to a special port called the SPAN destination port. Any Fibre Channel port in a switch can be configured as an SD_Port. Once an interface is in SD_Port mode, it cannot be

used for normal data traffic. You can attach a Fibre Channel analyzer to the SD_Port to monitor SPAN traffic.

SD_Ports do not receive frames, they only transmit a copy of the SPAN source traffic. The SPAN feature is non-intrusive and does not affect switching of network traffic for any SPAN source port.

Illustrated in Figure 11-24 is an overview of the SPAN port.

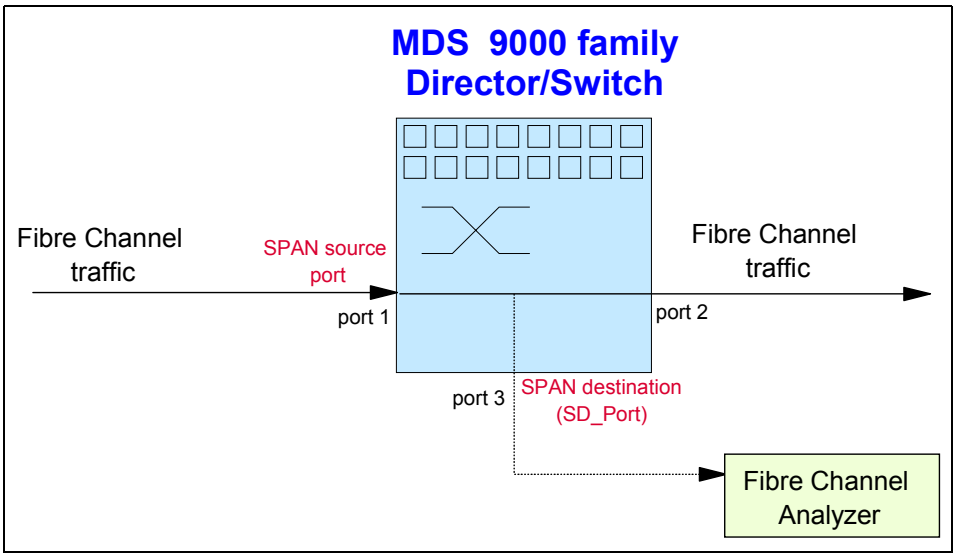


Figure 11-24 SPAN destination ports

SPAN sources

A SPAN source is the interface from which traffic can be monitored. You can also specify a VSAN as a SPAN source, in which case, all supported interfaces in the specified VSAN are included as SPAN sources. You can choose the SPAN traffic in the ingress direction, the egress direction, or both directions, for any source interface.

- **Ingress source (rx):** Traffic entering the switch fabric through this source is spanned or copied to the SD_Port, as shown in Figure 11-25.

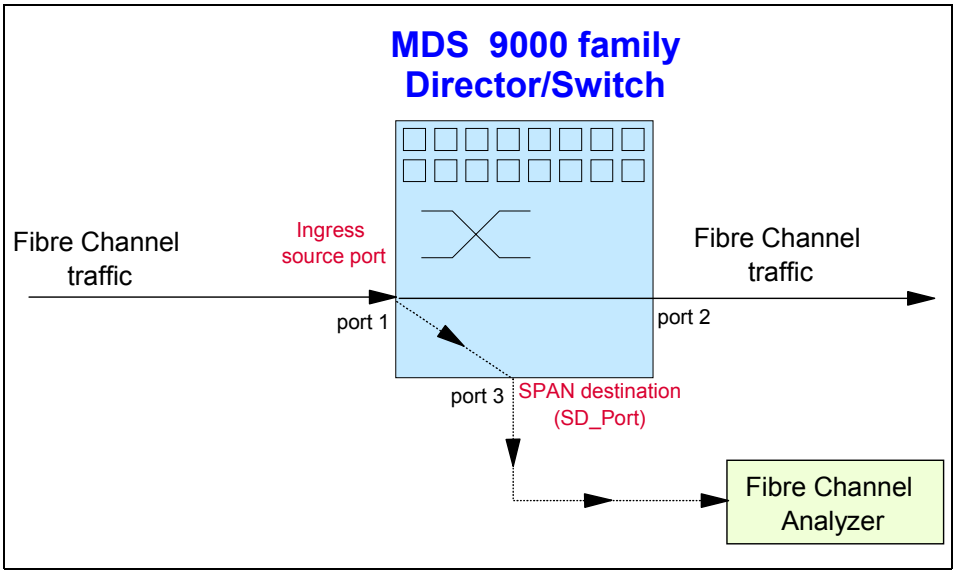


Figure 11-25 SD_Port for incoming traffic (ingress direction)

- **Egress source (tx):** Traffic exiting the switch fabric through this source interface is spanned or copied to the SD_Port, as shown in Figure 11-26.

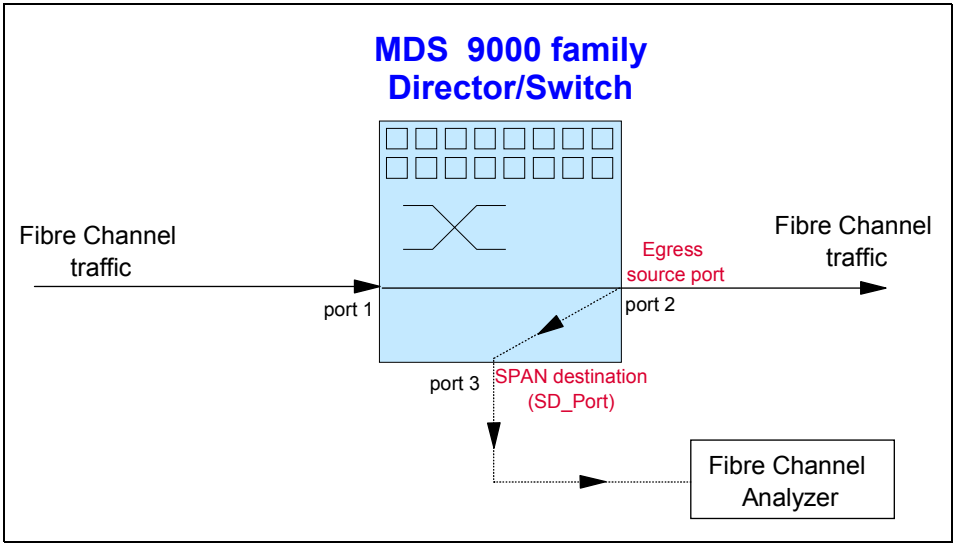


Figure 11-26 SD_Port for outgoing traffic (egress direction)

Allowed source interface types

The SPAN feature is available for the following interface types:

- Physical ports:
  - F_Ports, FL_Ports, TE_Ports, E_Ports, and TL_Ports.
- Interface sup-fc0 (traffic to and from the supervisor):
  - The Fibre Channel traffic from the supervisor module to the switch fabric, through the sup-fc0 interface, is called ingress traffic. It is spanned when sup-fc0 is chosen as an ingress source port.
  - The Fibre Channel traffic from the switch fabric to the supervisor module, through the sup-fc0 interface, is called egress traffic. It is spanned when sup-fc0 is chosen as an egress source port.
- PortChannels:
  - All ports in the PortChannel are included and spanned as sources.
  - You cannot specify individual ports in a PortChannel as SPAN sources. Previously-configured SPAN-specific interface information is discarded.

VSAN as a SPAN source

When a VSAN as a source is specified, then all physical ports and PortChannels in that VSAN are included as SPAN sources. A TE_Port is included only when the port VSAN of the TE_Port matches the source VSAN. A TE_Port is excluded even if the configured allowed VSAN list may have the source VSAN, but the port VSAN is different.

Guidelines for configuring VSANs as a source

The following guidelines apply when configuring VSANs as a source:

- Traffic on all interfaces included in a source VSAN is spanned only in the ingress direction.
- When a VSAN is specified as a source, you will not be able to perform interface-level configuration on the interfaces that are included in the VSAN. Previously-configured SPAN-specific interface information is discarded.
- If an interface in a VSAN is configured as a SPAN source, you will not be able to configure that VSAN as a source. You must first remove the existing SPAN configurations on such interfaces before configuring VSAN as a source.
- Interfaces are only included as sources when the port VSAN matches the source VSAN.

SPAN sessions

Each SPAN session represents an association of one destination with a set of source(s) along with various other parameters that you specify to monitor the network traffic. One destination can be used by one or more SPAN sessions. You can configure up to 16 SPAN sessions in a switch. Each session can have several source ports and one destination port.

To activate a SPAN session, at least one source and the SD_Port must be up and functioning. Otherwise, traffic will not be directed to the SD_Port.

To temporarily deactivate (suspend) a SPAN session use the **`suspend`** command in the SPAN submode. The traffic monitoring is stopped during this time. You can reactivate the SPAN session using the **`no suspend`** command.

Specifying filters

You can perform VSAN-based filtering to selectively monitor network traffic on specified VSANs. You can apply this VSAN filter to the selected source or to all sources in a session. Only traffic in the selected VSANs is spanned when you configure VSAN filters. You can specify two types of VSAN filters:

- **Interface level filters:** You can apply VSAN filters for a specified TE_Port or trunking PortChannel to filter traffic using one of three options — the ingress direction, the egress direction, or both directions.
- **Session filters:** This option filters all sources in the specified session. These filters are bi-directional and apply to all sources configured in the session.

Guidelines for specifying filters

The following guidelines apply to SPAN filters:

- Specify filters in either the ingress direction, or in the egress direction, or in both directions.
- PortChannel filters are applied to all ports in the PortChannel.
- If no filters are specified, the traffic from all active VSANs for that interface is spanned.
- The effective filter on a port is the intersection (filters common to both) of interface filters and session filters.
- While you can specify any arbitrary VSAN filters in an interface, traffic can only be monitored on the port VSAN or on allowed-active VSANs in that interface.
- When you configure VSAN as a source, that VSAN is implicitly applied as an interface filter to all sources included in the specified VSAN.

SD_Port characteristics

An SD_Port has the following characteristics:

- Ignores buffer-to-buffer credits.
- Allows data traffic only in the egress (tx) direction.
- Does not require a device or an analyzer to be physically connected.
- Supports only 1 Gbps or 2 Gbps speeds. The auto speed option is not allowed.
- Multiple sessions can share the same destination ports.
- If the SD_Port is shut down, all shared sessions stop generating SPAN traffic.
- The port mode can not be changed if it is being used for a SPAN session.
- The outgoing frames can be encapsulated in EISL format.
- The SD_Port does not have a port VSAN.

The following guidelines apply for a SPAN configuration:

- You can configure up to 16 SPAN sessions with multiple ingress (rx) sources.
- You can configure a maximum of three SPAN sessions with one egress (tx) port.
- In a 32 port switching module, you must configure the same session in all four ports in one port group. If you wish, you can also configure only two or three ports in this unit
- SPAN frames are dropped if the sum of the bandwidth of the sources exceeds the speed of the destination port.
- Frames dropped by a source port are not spanned.

11.6.3 Monitoring traffic using Fibre Channel analyzers

You can use SPAN to monitor traffic on an interface without any disruption to traffic. This feature is very useful in troubleshooting scenarios when traffic disruption can change the problem environment, and it can also make it difficult to reproduce the problem.

Without SPAN

You can monitor traffic, as shown in Figure 11-27, using interface 1 in a Cisco MDS 9000 family switch that is connected to another device on the SAN. You would need to physically connect a Fibre Channel analyzer between the switch and the storage device to analyze the traffic through interface 1.

This type of connection has the following limitations:

- It requires you to insert the Fibre Channel analyzer between the devices, which disrupts traffic.
- The Fibre Channel analyzer captures data only on the rx links in both port A and port B. Port A captures traffic exiting interface 1 and port B captures ingress traffic into interface 1.

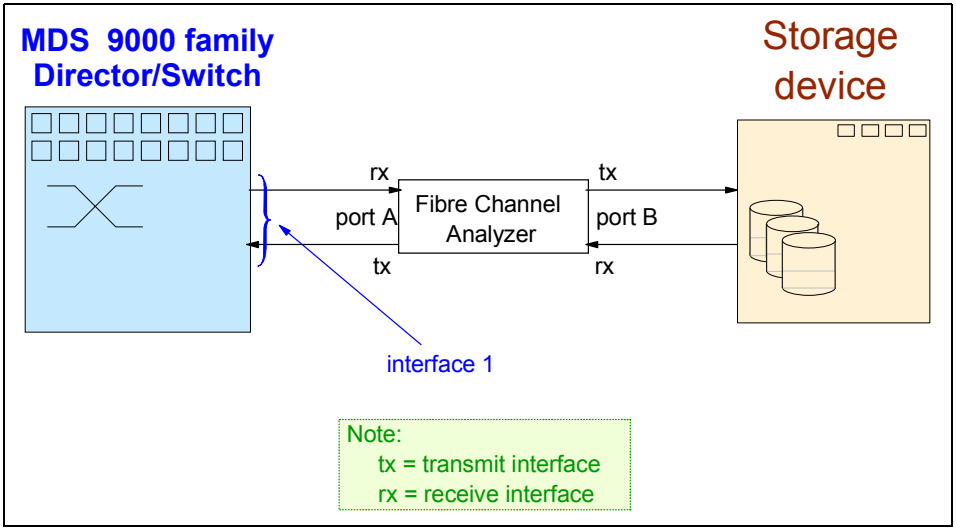


Figure 11-27 Fibre Channel analyzer without SPAN

Using SPAN

Using SPAN you can capture the same traffic scenario shown in Figure 11-27 without any traffic disruption. The Fibre Channel analyzer uses the ingress (rx) link at port A to capture all the frames going out from port 1. It uses the ingress link at port B, to capture all the frames coming in on port 1. This is illustrated in Figure 11-28.

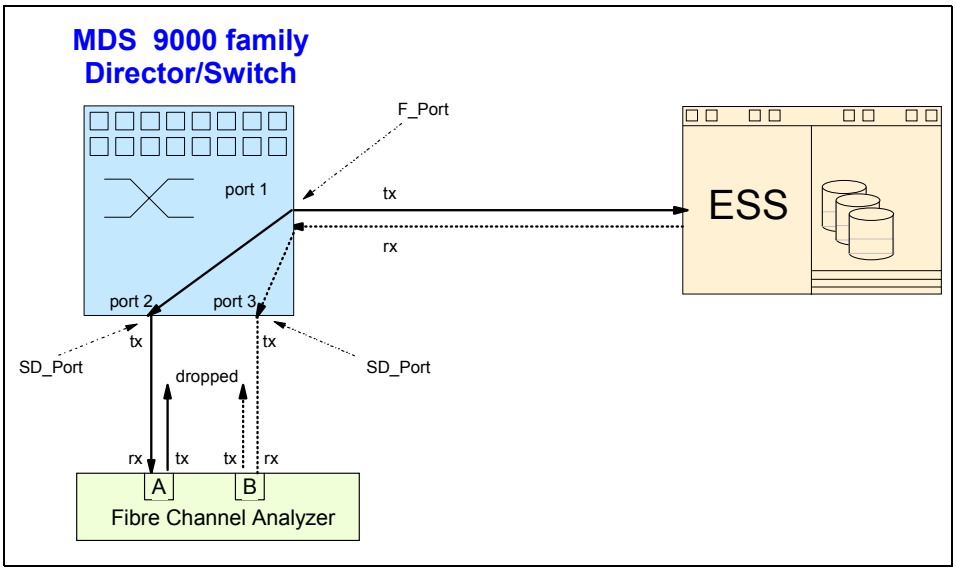


Figure 11-28 Fibre Channel analyzer using SPAN

Using a single SD_Port to monitor traffic

You do not need to use two SD_Ports to monitor bi-directional traffic on any interface, as shown in Figure 11-28. You can use one SD_Port and one FC analyzer port by monitoring traffic on the interface at the same SD_Port (Port 2).

Figure 11-29 shows a SPAN setup where one session with a destination of Port 2 and a source interface of Port 1 is used to capture traffic in both ingress and egress directions. This setup is more advantageous and cost-effective than the setup shown in Figure 11-28 — it uses one SD_Port and one port on the analyzer, instead of using a full, two-port analyzer.

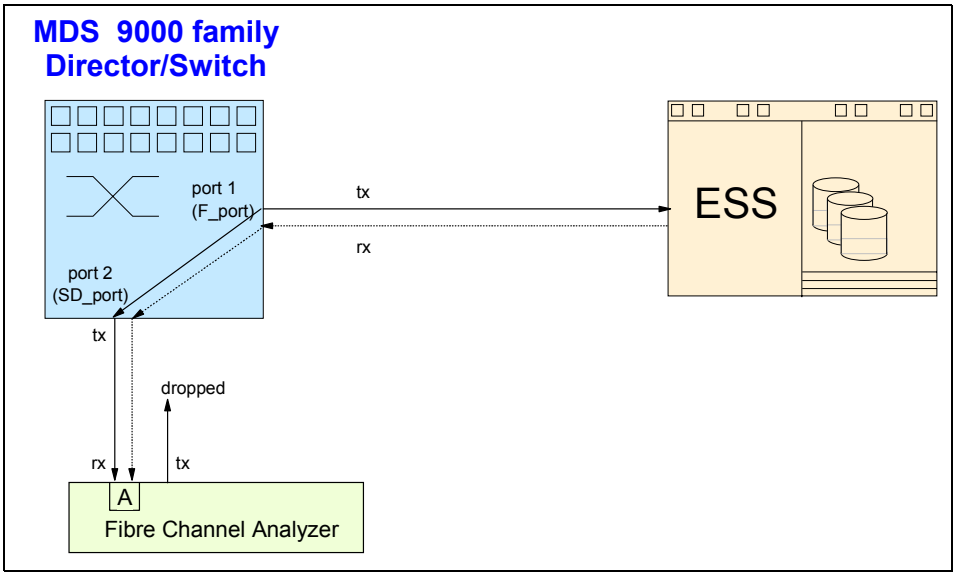


Figure 11-29 Using a single SD_Port to monitor traffic

11.7 FICON

IBM qualification for FICON covers the Cisco MDS 9216 Multilayer Fabric Switch, Cisco MDS 9506 Multilayer Director and Cisco MDS 9509 Multilayer Director as these switches are fully compliant with FC-SB-2 and FC-SB-3 standards. The Cisco Mainframe Package optional license is required as discussed in 11.4.1, “Licensed feature packages” on page 511.

Cisco VSAN technology can be used to provide greater security for a FICON traffic in an intermix environment. FICON may also be encapsulated over FCIP using the IP Services Module to provide an alternative to using channel extender devices.

The Cisco FICON environment may be managed from z/OS using the CUP function, from Cisco Fabric and Device Managers or by using the Cisco SAN-OS CLI.

11.8 Fabric management

The Cisco MDS 9000 family provides three principal modes of management:

- **The MDS 9000 family command line interface (CLI):** This presents the user with a consistent, logical CLI, which adheres to the syntax of the widely known Cisco IOS CLI. This is an easy-to-use command interface which has broad functionality.
- **Cisco Fabric Manager:** This is a responsive, easy-to-use Java application that simplifies management across multiple switches and fabrics. It enables administrators to perform vital tasks such as topology discovery, fabric configuration and verification, provisioning, monitoring, and fault resolution. All functions are available through an interface that enables remote management from any location.
- **Integration with third-party storage management tools:** This may be used independently or in conjunction with third party management applications. Cisco provides an extensive application programming interface (API) for integration with third-party and user developed management tools

11.8.1 Cisco MDS 9000 Fabric Manager

The Cisco Fabric Manager is included with the Cisco MDS 9000 family of switches and is a Java and SNMP-based network fabric and device management tool. It provides a GUI that displays real-time views of your SAN fabric and installed devices. The Cisco Fabric Manager provides three views for managing your network fabric:

- The Device View displays a continuously updated physical picture of device configuration and performance conditions for a single switch.
- The Fabric View displays a view of your network fabric, including multiple switches.
- The Summary View presents a summary view of switches, hosts, storage subsystems, and VSANs.
- The Cisco Fabric Manager provides an alternative to the CLI for most switch configuration commands.

The Cisco Fabric Manager is included with each switch in the Cisco MDS 9000 family.

11.8.2 In-band management and out-of-band management

The Cisco Fabric Manager requires an out-of-band (ethernet) connection to at least one Cisco MDS 9000 family switch to enable it to discover and manage the fabric.

The interface used for an out-of-band management connection is a 10/100 Mbps ethernet interface on the supervisor module, labeled mgmt0. The mgmt0 connection can be connected to a management network to access the switch through IP over ethernet.

ethernet connectivity is required to at least one Cisco MDS 9000 family switch. This connection is then used to manage the other switches using in-band (Fibre Channel) connectivity. Otherwise, you need to connect the mgmt0 port on each switch to your ethernet network.

Each supervisor module has its own ethernet connection; however, the two ethernet connections in a redundant supervisor system operate in active or standby mode. The active supervisor module also hosts the active mgmt0 connection. When a failover event occurs to the standby supervisor module, the IP address and media access control (MAC) address of the active ethernet connection are moved to the standby ethernet connection. This eliminates any need for the management stations to relearn the location of the switch.

An example of an out-of-band management solution is shown in Figure 11-30.

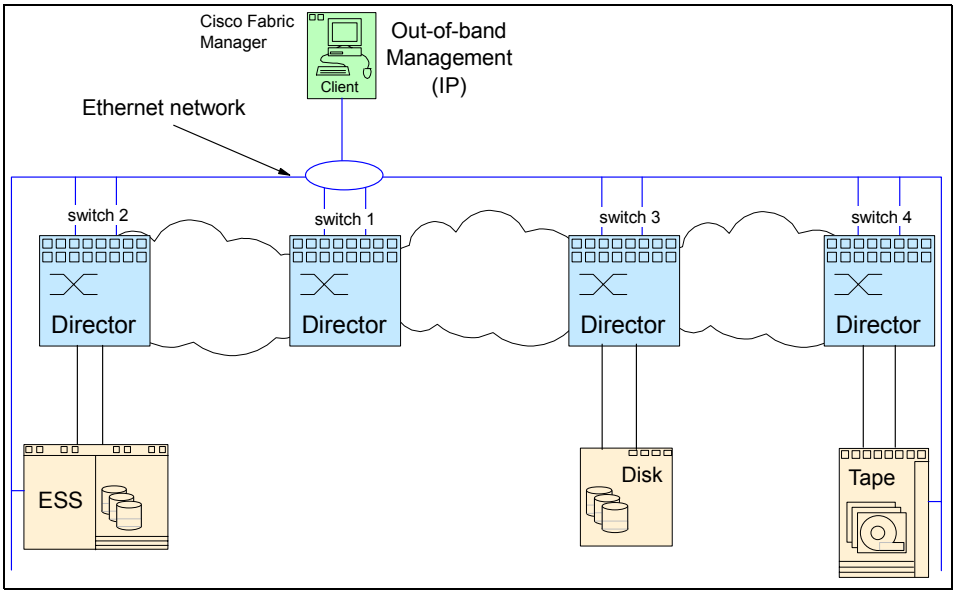


Figure 11-30 Out-of-band management connection

You can also manage switches on a Fibre Channel network using an in-band connection to the supervisor module. This in-band connection supports either management protocols over Fibre Channel or IP embedded within Fibre Channel. The Cisco MDS 9000 family supports RFC 2625 IP over Fibre Channel (IPFC), which allows IP to be transported between Fibre Channel devices over the Fibre Channel protocol, as shown in Figure 11-31.

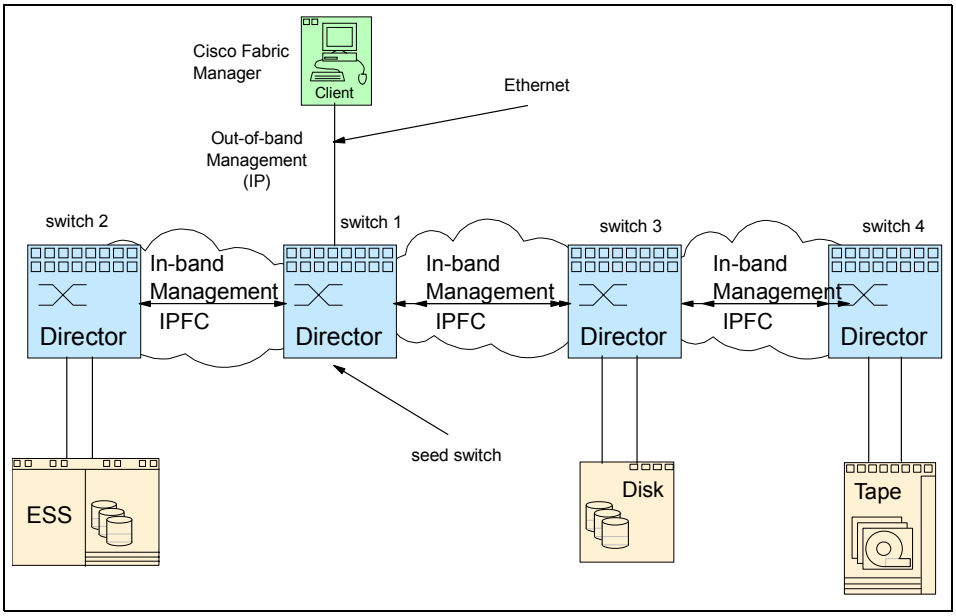


Figure 11-31 In-band management connection

IPFC encapsulates IP packets into Fibre Channel frames so that management information can cross the Fibre Channel network without requiring a dedicated ethernet connection to each switch. IP addresses are resolved to the Fibre Channel address through Address Resolution Protocol (ARP). With host bus adapters (HBAs) that support IP drivers, this capability allows for a completely in-band management network. The switch also uses the in-band interface to discover its own environment, including directly connected and fabric-wide elements.

11.8.3 Using the setup routine

When you connect to a Cisco MDS 9000 family switch using the local console and start the switch for the first time, the system displays a setup routine that helps you perform the basic configuration required to manage and connect the switch to end nodes or other switches. The setup routine must be completed

before you can connect to the switch or manage it using the Cisco Fabric Manager.

The setup routine prompts for the following configuration values:

- Default switch port status (enabled or disabled)
- Trunking (enabled or disabled)
- Zone policy (deny or permit access)
- IP address and subnet mask for mgmt0 (user supplied)
- Telnet/SSH policy (enabled or disabled)
- SNMPv3 user name and password (admin/admin_Pass)

In addition to these settings, each Cisco MDS 9000 family switch is configured with the following default values:

- VSAN membership: All ports are in VSAN 1
- Switch port speed and type: Autosense

11.8.4 Controlling administrator access with users and roles

The Cisco MDS 9000 family switches support role-based management access whether using the CLI or the Cisco Fabric Manager. This lets you assign specific management privileges to particular roles and then assign one or more users to each role.

When using the CLI, you can configure a local database of users using the CLI, or establish this database using a RADIUS server and then configure the CLI to verify console access with the RADIUS server. You can also use the Cisco Fabric Manager to configure the RADIUS server for authenticating CLI access.

Cisco Fabric Manager uses SNMPv3 to establish role-based management access. After completing the setup routine, a single role, user name, and password are established. If the defaults are accepted during the setup routine, the user name is admin and the default password is admin_Pass. Changing the password is highly recommended. The role assigned to this user allows the highest level of privileges, which includes creating new users and roles. Use the Cisco Fabric Manager to create roles, and users, and to assign passwords as required for secure management access in your network.

11.8.5 Accessing Cisco Fabric Manager

Before you can access the Cisco Fabric Manager, you must complete the following tasks:

- A supervisor module must be installed on each switch that you want to manage

- The supervisor module must be configured with the following values using the setup routine or the CLI:
  - IP address assigned to the mgmt0 interface
  - SNMPv3 user name and password

11.8.6 Connecting to a supervisor module

The Cisco Fabric Manager software executables reside on each supervisor module of each Cisco MDS 9000 Family switch in your network. The supervisor module provides an HTTP server that responds to browser requests and distributes the software to Windows or UNIX network management stations.

To install the software for the first time, or if you want to update or reinstall the software, access the supervisor module with a Web browser. When you click the install buttons on the Web page that is displayed, the software running on your workstation is verified to make sure you are running the most current version of the software. If it is not current, the most recent version is downloaded and installed on your workstation.

11.9 Zoning

Zoning is a mechanism for protecting resources within the SAN or a VSAN by grouping devices together that require common access. Zoning allows us to enable or disable certain ports within the SAN.

For example, in Figure 11-32 we show a typical scenario where a customer has a SAN with an NT server and a UNIX server. In this scenario we want to restrict the NT server to only see the IBM TotalStorage DS4400 disk subsystem and nothing else in the SAN. We also want to enable the UNIX server to see the same IBM TotalStorage DS4400 disk subsystem and the 3584 tape library. To do this, we define a zone (Zone_1) that contains the NT server and the IBM TotalStorage DS4400. We also create another zone (Zone_2) to handle allow the UNIX server to see the IBM TotalStorage DS4400 disk subsystem and the 3584 tape subsystem.

In this example we choose to do this to stop NT from acquiring all the devices attached to the SAN. In this example, if the SAN were not zoned, then there would be nothing to stop the NT server from acquiring the UNIX systems disk and tape devices.

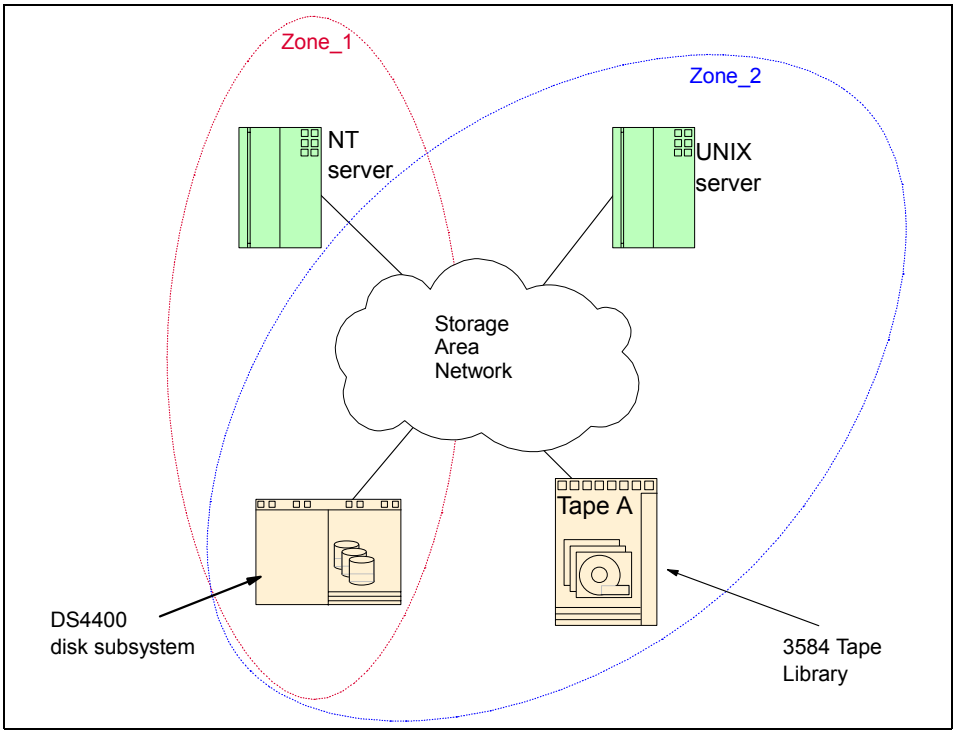


Figure 11-32 Zoning overview

11.9.1 Zone features

Zoning has the following features:

- A zone consists of multiple zone members:
  - Members in a zone can only access each other.
  - If zoning is not activated, all devices are members of the default zone.
  - If zoning is activated, any device that is not in an active zone (a zone that is part of an active zone set) is a member of the default zone.
  - Zones can vary in size.
  - Devices can belong to more than one zone.
- A zone set consists of one or more zones:
  - A zone set can be activated or deactivated as a single entity across all switches in the fabric.
  - Only one zone set can be activated at any time.
  - A zone can be a member of more than one zone set.

- Zoning can be administered from any switch in the fabric:
  - Because zoning information is distributed to all switches in the fabric, zoning changes made on one switch are available in all switches.
  - If a new switch is added to an existing fabric, zone sets are acquired by the new switch.
- Zone changes can be configured non-disruptively:
  - New zones and zone sets can be configured without interrupting traffic on unaffected ports or devices.
- The Default Zone includes all ports or WWNs that do not have a specific membership association.
  - Access between default zone members is controlled by the default zone policy.

11.9.2 Zone membership

The Cisco MDS 9000 family offer a number of methods of zoning. Zones can be set up using:

- **Port World Wide Node Name (pWWN):** This specifies the pWWN of an N_Port attached to the switch
- **Fabric pWWN:** This specifies the WWN of the fabric port (or the switch port's WWN). This is also referred to as port-based zoning.
- **Fibre Channel ID:** This specifies the Fibre Channel ID of an N_Port.

11.9.3 Configuring a zone

A zone can be configured using one of the following types to assign members:

- **pWWN:** The WWN of the N or NL port in hexadecimal format (for example, 10:00:00:23:45:67:89:ab).
- **Fabric port WWN:** The WWN of the fabric port name in hexadecimal format (for example, 10:00:00:23:45:67:89:ab).
- **FC_ID:** The N port ID in 0xhhhhhh format (for example, 0xce00d1).
- **FC alias:** The alias name is in alphabetic characters (for example, *Payroll*) and denotes a port ID or WWN. The alias can also include multiple members.

11.9.4 Zone enforcement

Zoning can be enforced in two ways: soft and hard.

Each end device (N_Port or NL_Port) discovers other devices in the fabric by querying the name server. When a device logs in to the name server, the name server returns the list of other devices that can be accessed by the querying device. If an Nx port does not know about the FCIDs of other devices outside its zone, it cannot access those devices.

In soft zoning, zoning restrictions are applied only during interaction between the name server and the end device. If an end device somehow knows the FC_ID of a device outside its zone, it can access that device.

Hard zoning is enforced by the hardware on each frame sent by an Nx port. As frames enter the switch, source-destination IDs are compared with permitted combinations to allow the frame at wirespeed.

The Cisco MDS 9000 family of switches supports both of these methods.

11.9.5 Zone sets

While zones provide access control to devices, a zone set is a group of zones to enforce access control across the whole fabric. Multiple zone sets can be created but only a single zone set can be activated and active at once. Zone sets contain the names of member zones.

If one zone set is currently active and another zone set is activated then the current zone set is de-activated and the new zone set is activated.

11.9.6 Default zone

Each member of a fabric can belong to any zone. If a member is not part of any active zone, it is considered to be part of a default zone. Therefore, if no zone set is active in the fabric, all devices are considered to be in the default zone. Even though a member can belong to multiple zones, a member that is part of the default zone cannot be part of any other zone.

The switch determines whether a port is a member of the default zone when the attached port comes up. Traffic can be permitted or denied to members of the default zone.

This information is not distributed to all switches; it must be performed for each switch. When the switch is initialized for the first time, no zones are configured and all members are considered to be part of the default zone. Members are not permitted to talk to each other.

This ensures that devices do not gain access to each other before zoning is activated.

11.10 Switch interoperability mode

Switch interoperability mode enables other vendors' switches to connect to the MDS family. This mode is required because vendors' often implement features on their switches that are not compatible with other manufacturers, so this mode provides a common implementation that all vendors' switches can apply.

Switch interoperability mode will disable a number of advanced or proprietary features, so it is worth understanding what these may affect before proceeding down this path.

Table 11-7 lists the changes required should interoperability mode be enabled on a Cisco MDS 9000 family switch or director.

Table 11-7 Interoperability mode changes

Switch feature	Changes if Interoperability Mode is enabled
Domain IDs	Some vendors do not use the full range of 239 available domain IDs within a fabric. Interop Domain IDs are restricted to the range 97-127 to accommodate all vendors implementations.
Timers	All Fibre Channel timers must be set to the same value on all switches as these values are exchanged by E_Ports when establishing an ISL. The Time-Out Value timers are described in the following rows. (this is always required for connection of two switches)
F_S_TOV	Verify that the Fabric Stability Time-Out Value timers match exactly.
D_S_TOV	Verify that the Distributed Services Time-Out Value timers match exactly.
E_D_TOV	Verify that the Error Detect Time-Out Value timers match exactly.
R_A_TOV	Verify that the Resource Allocation Time-Out Value timers match exactly.
Trunking	Trunking is not supported between two different vendors switches. This feature may be disabled on a per port basis.
Default Zone	The default zone behavior of permit (all nodes can see other nodes) or deny (all nodes are isolated when not explicitly placed in a zone), may change.
Zoning attributes	Zones may be limited to the WWPN and other proprietary zoning methods (physical port number), may be eliminated.

Switch feature	Changes if Interoperability Mode is enabled
Zone propagation	Some vendors do not pass the full zone configuration to other switches, only the active zoneset gets passed.
VSAN	Will only affect the specified VSAN.
TE_Ports and PortChannels	TE_Ports and PortChannels only apply when connecting from one MDS 9000 to another. Only E_Ports can be used to connect to non-MDS switches. TE_Ports and PortChannels can be used to connect to other MDS 9000 switches when Interoperability Mode is enabled.
Domain reconfiguration disruptive	This can require the entire switch to be restarted when changing Domain IDs.
Domain configuration nondisruptive	This will only impact the affected VSAN. Only the domain manager for the affected VSAN is restarted but other VSANs will be unaffected.
Name Server	Need to verify that all vendors have the correct values in their respective Name Server tables.

Interoperability mode in the Cisco MDS 9000 family can be enabled non-disruptively, but the default is to have this mode turned off.

It is still important to check with the OEM vendors involved as to the specific steps that must be taken.

11.10.1 Interoperability matrix

To understand the device and host attachment IBM supports for connection to the Cisco MDS family of switches the following URL should be viewed. This list is being constantly updated and should be consulted regularly:

http://www.storage.ibm.com/ibmsan/products/cisco/library.html

Selecting the Interoperability Matrix will display a complete list of servers, disk and tape systems that are supported with the MDS 9509 Multilayer Director. This list also contains supported operating system versions and links to other Web sites that document the required HBA levels.

11.11 IBM Feature Codes

Machine type/model 2062-D01

- **FC 2218:** 4 Port IP Services Module

- **FC 2219:** FCIP Act 4 Port IPS Module
- **FC 5311:** 1470 nm CWDM FC SFP
- **FC 5312:** 1490 nm CWDM FC SFP
- **FC 5313:** 1510 nm CWDM FC SFP
- **FC 5314:** 1530 nm CWDM FC SFP
- **FC 5315:** 1550 nm CWDM FC SFP
- **FC 5316:** 1570 nm CWDM FC SFP
- **FC 5317:** 1590 nm CWDM FC SFP
- **FC 5318:** 1610 nm CWDM FC SFP
- **FC 6111:** Cisco Warranty 1yr 4 Port IPS
- **FC 6112:** Cisco Warranty 2yr 4 Port IPS
- **FC 7015:** MDS 9216 FMS Pkg Bundle
- **FC 7030:** MDS 9216 Mainframe Pkg Bundle
- **FC 9120:** Cabinet Jumper Power Cord

Machine type/model 2062-D04

- **FC 2218:** 4 Port IP Services Module
- **FC 2220:** FCIP Act 4 Port IPS Module
- **FC 5311:** 1470 nm CWDM FC SFP
- **FC 5312:** 1490 nm CWDM FC SFP
- **FC 5313:** 1510 nm CWDM FC SFP
- **FC 5314:** 1530 nm CWDM FC SFP
- **FC 5315:** 1550 nm CWDM FC SFP
- **FC 5316:** 1570 nm CWDM FC SFP
- **FC 5317:** 1590 nm CWDM FC SFP
- **FC 5318:** 1610 nm CWDM FC SFP
- **FC 6111:** Cisco Warranty 1yr 4 Port IPS
- **FC 6112:** Cisco Warranty 2yr 4 Port IPS
- **FC 7025:** MDS 9500 FMS Package Bundle
- **FC 7035:** MDS 9500 Mainframe Package Bundle
- **FC 9225:** PwrC, 250 V ac 16A, So Africa
- **FC 9226:** PwrC, 250 V ac 16A 23G Switz
- **FC 9227:** PwrC, 250 V ac 13A, U.K.
- **FC 9229:** Cabinet Jumper Power Cord

Machine type/model 2062-T04

- **FC 2218:** 4 Port IP Services Module
- **FC 2220:** FCIP Act 4 Port IPS Module
- **FC 5311:** 1470 nm CWDM FC SFP
- **FC 5312:** 1490 nm CWDM FC SFP
- **FC 5313:** 1510 nm CWDM FC SFP
- **FC 5314:** 1530 nm CWDM FC SFP
- **FC 5315:** 1550 nm CWDM FC SFP
- **FC 5316:** 1570 nm CWDM FC SFP

- **FC 5317:** 1590 nm CWDM FC SFP
- **FC 5318:** 1610 nm CWDM FC SFP
- **FC 6111:** Cisco Warranty 1yr 4 Port IPS
- **FC 6112:** Cisco Warranty 2yr 4 Port IPS
- **FC 7025:** MDS 9500 FMS Package Bundle
- **FC 7035:** MDS 9500 Mainframe Package Bundle

Machine type/model 2062-D07

- **FC 2218:** 4 Port IP Services Module
- **FC 2220:** FCIP Act 4 Port IPS Module
- **FC 5311:** 1470 nm CWDM FC SFP
- **FC 5312:** 1490 nm CWDM FC SFP
- **FC 5313:** 1510 nm CWDM FC SFP
- **FC 5314:** 1530 nm CWDM FC SFP
- **FC 5315:** 1550 nm CWDM FC SFP
- **FC 5316:** 1570 nm CWDM FC SFP
- **FC 5317:** 1590 nm CWDM FC SFP
- **FC 5318:** 1610 nm CWDM FC SFP
- **FC 6111:** Cisco Warranty 1yr 4 Port IPS
- **FC 6112:** Cisco Warranty 2yr 4 Port IPS
- **FC 7025:** MDS 9500 FMS Package Bundle
- **FC 7035:** MDS 9500 Mainframe Package Bundle
- **FC 9214:** PwrC, 250 V ac 16A, So Africa
- **FC 9215:** PwrC, 250 V ac 16A 23G Switz
- **FC 9229:** Cabinet Jumper Power Cord

Machine type/model 2062-T07

- **FC 2218:** 4 Port IP Services Module
- **FC 2220:** FCIP Act 4 Port IPS Module
- **FC 5311:** 1470 nm CWDM FC SFP
- **FC 5312:** 1490 nm CWDM FC SFP
- **FC 5313:** 1510 nm CWDM FC SFP
- **FC 5314:** 1530 nm CWDM FC SFP
- **FC 5315:** 1550 nm CWDM FC SFP
- **FC 5316:** 1570 nm CWDM FC SFP
- **FC 5317:** 1590 nm CWDM FC SFP
- **FC 5318:** 1610 nm CWDM FC SFP
- **FC 6111:** Cisco Warr 1yr 4 Port IPS
- **FC 6112:** Cisco Warr 2yr 4 Port IPS
- **FC 7025:** MDS 9500 FMS Package Bundle
- **FC 7035:** MDS 9500 Mainframe Package Bundle

12

SAN event data gathering tips

The purpose of this chapter is to detail all the information that needs to be collected to assist in resolving SAN related problems. It may seem daunting to have to go through several steps to gather all the data requested, but the most common cause of delays in problem resolution is a lack of data. It should never be assumed that the root cause of any problem is in the most obvious place.

By gathering logs from all parts of the SAN we give ourselves the greatest chance of getting a fast and effective resolution to the problem being experienced.

The second most common cause of delays in problem resolution is providing data which has been collected some hours or even days after the problem occurred, often in this case there is no longer evidence of the original problem.

Timely and complete data collection will aid in problems being resolved quickly.

© Copyright IBM Corp. 2004. All rights reserved.

12.1 Overview

Collection of timely and detailed information for various host platforms, SAN switches and storage devices is outlined in the following sections.

The collection of log information is critical to understanding the cause of an event in a SAN environment was, and to aid the support people analyzing the collected logs it is very useful to also provide the individual equipments time offsets. As some hardware may have never had its real time clock set to the local time, it can become very difficult to match events from one piece of equipment to the other.

Another important piece of information to aid in timely error analysis is a physical diagram of the SAN topology. This diagram should be kept up to date, and include all hosts, switches, directors and storage devices within the SAN. This document can save many hours of recontructing *the big picture* from piecing log information together.

12.2 Hosts

The following sections details some of the useful information to be collected at a host level.

12.2.1 AIX

Time difference

Use the **date** command to display the system date and time.

Log collection

Collect both **errpt** and **errpt -a** (each piped to a file).

Hardware configuration collection

Take a snap (the **errpt** is found in a snap, but it is good to have a separate copy).

The preferred snap for IBM TotalStorage DS Family problems is: **snap -gfiLc** where:

- **g** - Gathers the output of the **lslpp -hBc** command, which collects the exact operating system environment
- **f** - Gathers file system information
- **i** - Gathers installation debug vital product data (VPD) information
- **L** - Gathers LVM information

- **c** - Creates a compressed pax image (**snap.pax.Z** file)

Multi-pathing data collection

SDD (all versions of AIX)

Issue the following commands and capture the output. This data is not found in a snap. Preferably provide the output of these commands during the failure.

- **datapath query adapter**
- **datapath query device**
- **lsvpcfg**

MPIO (available on AIX 5.2 and above)

Issue the following commands and capture the output. This data is not included in a snap.

- **pcmpath query adapter**
- **pcmpath query device**
- **pcmpath query essmap**

12.2.2 HP-UX

Time difference

Use the **date** command to get the system date and time.

Log collection

Collect the contents of the */var/adm/syslog/syslog.log* file.

Hardware configuration collection

Provide the following server details for each server involved with the SAN:

- Manufacturer
- Machine Type/Model Number
- Feature details, for example, number of CPUs, amount of memory

HBA details

For SAN problems we always need the following details about the FC HBAs:

- Hardware manufacturer, brand and model
- BIOS (firmware) level - BIOS settings if QLogic
- Driver level

Software configuration collection

Capture the ouput of the **uname -a** command

Multi-pathing data collection

SDD

Issue the following commands and capture the output. Preferably provide the output of these commands during the failure.

- **datapath query adapter**
- **datapath query device**

12.2.3 Linux

Time difference

Use the **date** command to get the system date and time.

Log collection

Captute the contents of **/var/log/messages**

Capture the output of the **dmesg** command

Hardware configuration collection

For IBM xSeries hardware, the best way to collect configuration data is by using the e-gatherer tool. Make sure you also supply the HBA details however.

e-gatherer can be downloaded from:

http://www-306.ibm.com/pc/support/site.wss/MIGR-4R5VKC.html

Software configuration collection

Capture the ouput of the command: **uname -a**

If you are running Redhat, install and run **sysreport** and send the output.

Multi-pathing Data Collection

SDD

Issue the following commands and capture the output. Preferably provide the output of these commands during the failure.

- **datapath query adapter**
- **datapath query device**

12.2.4 Microsoft Windows

Time difference

Display the system date and time using the clock in the bottom right hand corner or by issuing the **time** and **date** commands at a command prompt.

Log collection

Always save the system logs and the application logs as soon as possible after the event.

Do not export the logs and providing the logs in EVT format is not helpful.

To find the system logs either right click **My computer** then **Manage** or click:

Start —> Programs —> Administrative Tools —> Computer Management

When opened go to:

System Tools —> Event Viewer —> System

And then:

Action —> Save Log file as changing the save as type to CSV

Repeat for **Application** logs.

Hardware configuration collection

For IBM xSeries hardware, the best way to collect configuration data is by using the e-gatherer tool. Make sure you also supply the HBA details however.

e-gatherer can be downloaded from:

http://www-306.ibm.com/pc/support/site.wss/MIGR-4R5VKC.html

Software configuration data collection

If you cannot take the e-gatherer data provide:

- Operating System
- Service Pack Level

Multi-pathing data collection

SDD

If you are running SDD, issue the following commands and capture the output.

- **datapath query adapter**

- **datapath query device**

12.2.5 Novell Netware

Time difference

Display the system date and time.

Log collection

CONLOG.EXE is a utility which writes all system console messages to a .LOG file.

More details can be found at url:

http://www.novell.com/documentation/lg/nw42/index.html?utlrfenu/data/hq1lykxx.html

Hardware configuration collection

There is no e-gatherer for Netware.

Software configuration data collection

Provide:

- Operating System Level
- State whether this is a clustered system

Multi-pathing data collection

SDD

If you are running SDD, issue the following commands and capture the output.

- **datapath query adapter**
- **datapath query device**

12.2.6 SUN Solaris

Time difference

Use the **date** command to get the system date and time.

Log collection

Save the **/var/adm/messages** file. Previous days messages are normally available as **/var/adm/messagesx** where **x** is the number of days since the logs rolled.

Hardware configuration collection

There is no e-gather or **snap** command to collect these details, so a good description of the hardware uncluding the following is required:

Software configuration data collection

Provide :

- Operating System details.
- A copy of your **sd.conf** file
- output from **iostat -El**

Depending on the HBA there will be a **/kernel/drv/*.conf** where the * could be QLogic or JNI.

Multi-pathing data collection

SDD

Issue the following commands and capture the output.

- **datapath query adapter**
- **datapath query device**

Veritas Volume Manager DMP

Provide the output from the following commands:

- **ls -lL /dev/rdsk/***
- **ls -la /dev/vx/dmp/***
- **format**

12.3 Switches

The following topics show the necessary documentation to gather.

12.3.1 SAN Switch 2031/2032 (McDATA)

Time difference

From the hardware view of the switch in EFCM, go to the Configure menu and select the 'Date and Time' menu, to get the switch date and time.

Log collection

From EFCM element manager window. Select **Maintenance —> Data Collection**. This procedure will create a zip file on your EFC Client machine, and we suggest saving the file to floppy disk.

If the switch(s) are not managed by EFCM and SanPilot is used instead, then collect the output of the following commands in a telnet session to the switch.

- **show switch**
- **show system**
- **show zoning**
- **show eventLog**
- **show features**
- **show loginServer**
- **show nameServerExt**
- **show port config**
- **show port info**
- **show port status**
- **show port technology**
- **show security fabricBinding**
- **show security portBinding**
- **show security switchBinding**

Tip: just entering **show** without a parameter, will put you into *show* mode which saves repeatedly typing **show xxxxxx**.

Hardware configuration collection

This will be included in the EFCM Data Collection procedure. However a configuration diagram showing how your SAN is connected is of tremendous value.

12.3.2 SAN Switch 2062 (Cisco)

Time difference

From the CLI use the **sh clock** command to display the switch date and time.

Log collection

CLI

Using a telnet (or SSH) client, log in to the switch and turn on *session logging*.

Issue the command: **term len 0** (this will allow the output to continuously scroll)

Issue the command: **show tech-support details** (or **sh tech det** for short)

The term length can be set back to default by issuing **term len 25** or it will be reset by logging off the CLI session.

This should be performed for every switch in the fabric.

Fabric Manager GUI

The **show tech** may also be collected from the Fabric Manager menuby selecting the **Tools —> collect techSupport** menu.

When collecting a **show tech** from the Fabric Manager interface a JPG will be placed in the zip file that is the current fabric map view. To make the JPG meaningful it is best to ensure the entire fabric is displayed in the map view and that it has been arranged for best viewing.

Hardware configuration collection

A diagram showing how your SAN configuration is connected is of tremendous value.

12.3.3 SAN Switch 2109 (Brocade)

Time difference

Use the **date** command to display the switch date and time.

Correcting the time

This is a concurrent act and should be done if the displayed time is not accurate.

The syntax is:

Date "mmddHHMMyy" (month,day,hour,minute,year)

Example 12-1 setting the time to 15:31 on Feb 27 2005

```
sw5:admin> date "0227153105"
         Thu Feb 27 15:31:00 2005
```

Log collection

Using a telnet (or Secure telnet) client, log in to the switch and turn on *session logging*.

At the command prompt issue the command: **supportShow**

This needs to be performed for every switch in the fabric.

After the **supportshow** completes, issue the commands:

- **portLogClear**This will clear the port activity logs
- **portStatsClear x** where x is the port you are interested in)

The **portstatsclear** will clear the port statistics on every port attached to one ASIC, (normally a total of four ports). To clear the statistics for every port in the switch or director, the command must be specified for each quad within the switch.

Hardware configuration collection

A diagram showing how the SAN configuration is connected is of tremendous value.

12.3.4 SAN Switch 2042 (CNT)

Time difference

Login to the FCM debug port using a null-modem cable. The current machine time is displayed at the top of the screen.

Log collection

All of the logs are gathered by executing the **DebugBackup.bat** from the **INRANGE\FC9000** directory. The output file created is **DebugBackup.jar** or **debugBackup.zip**.

Important: All EM clients and the EM Server must be stopped before executing **DebugBackup.bat**. Do not perform from an EM client PC, this must be collected at the IN-VSN Server PC.

12.4 Storage

The following sections describe the documentation to gather at a storage level.

12.4.1 IBM TotalStorage DS Family disk subsystem

Time difference

Time is usually synchronized with the Storage Manager PC. To display and alter the time from the Storage Manager GUI go to **Storage Subsystem —> Set Controller Clock.**

The time displayed is the current IBM TotalStorage DS Family (formerly FAStT) controller time. Take a note of that time delta before altering it.

Note: Storage Manager 8.4 has enhancements to better manage the Controller Clock.

Log collection

This procedure will save what is known as the MEL logs.

Start the Storage Manager software and connect to the IBM TotalStorage DS Family (formerly FAStT) controller.

From the **Subsystem Management** window of each IBM TotalStorage DS Family (formerly FAStT) involved, click on the second icon from the left that looks like an open book as shown in Figure 12-1.

Figure 12-1 Storage Manager View Event Log icon

This event log view may also be accessed by clicking the **View —> View Event Log** in the Subsystem Management window.

From the Event Log window select the **Save All sections** button, navigate to the appropriate path, type filename and then **Save**. Save as many events as possible.

Hardware configuration collection

To provide support an understanding of the physical layout of the IBM TotalStorage DS Family (formerly FAStT) we want to collect what is known as a *Profile*.

Open the **FAStT Storage Manager Client**.

Double-click the target storage subsystem in the right pane to launch the "Storage subsystem" window.

Select **View —> Storage Subsystem Profile**

Select the **Save As** button to open the dialog box for a location and filename to save the profile to a file.

Attention: If the choice to **save All sections** is offered — do so; navigate to the appropriate path, type a filename and then **Save**.

12.4.2 IBM TotalStorage Enterprise Storage Server

When an IBM TotalStorage Enterprise Storage Server is involved in the SAN configuration that experienced problems, a service call to IBM should be placed to allow remote technical support, or an onsite service representitive to collect the relevant logs and statesaves as soon as possible after the event occurred.

12.4.3 3583 Tape Library and SDGM

Time difference

3583 Library

The libraries date and time may be viewed and set from the front panel. Refer to the 3583 Operators Guide for the procedure.

3583 SDGM

If your 3583 tape library has an embedded SAN Data Gateway Module (SDGM) installed, connect to the SDGM serial port and issue the **date** command to display the date and time.

Most SDGMs have never been set to the local time as it does not affect operation. We suggest, while connected, to follow this procedure to set it correctly:

Use the **rtcDateSet** command to manually set the Real Time Clock (RTC) using the following Syntax :

```
rtcDateSet
[year],[month],[dayofmonth],[dayofweek],[hour],[minute],[second]
```

In Example 12-2 we show how to set the date and time to Friday 26th of January 2004 at 9:30 AM

Example 12-2 Setting the date and time on the 3583 SDGM

```
SN60023 > rtcDateSet 2004,1,26,5,9,30,00
```

Note: The time is in 24 hour format.

Then as shown in Example 12-3 we use the **dateSetFromRTC** command to set the gateways RTC as the source of date display:

Example 12-3

```
SN60023 > dateSetFromRTC
```

Then we issue the **date** command as shown in Example 12-4 to confirm our setting.

Example 12-4

```
SN60023 > date
SN60023 > FRI JAN 26 9:30:49 2004
```

Log Collection

3583 Library

The *IBM TotalStorage Ultrium Scalable Tape Library 3583 Maintenance Information* - SA37-0425-02 manual contains a section on *Methods of Capturing Logs*. Follow the procedures and ensure all the boot logs are collected.

3583 SGDM

Connect via the serial port to the SDGM and enable session logging on your terminal emulator.

Issue the command **`supportDump`**.

Attention: This command will briefly disrupt data flow through the SDGM, we recommend to wait for no tape activity before issuing.

13

General solutions

In this chapter we discuss some of the building blocks and general concepts for building reliable and powerful SANs. Included are requirements for servers and storage and their software, as well as fabric devices.

We present various implementations and uses of a SAN environment, starting with a simple fabric and building it up into a complex design.

© Copyright IBM Corp. 2004. All rights reserved.

13.1 Objectives of SAN implementation

To ensure the highest level of system uptime, utilization, and security, companies are implementing reliable storage networks capable of boosting the availability of data for all the users and applications that need it. These companies typically represent the industries that demand the highest levels of system and data availability — the utilities and telecommunications sector, brokerages and financial service institutions, and a wide variety of service providers.

By reducing or eliminating single points of failure in the enterprise environment, SANs can help to improve overall availability of business applications. By utilizing highly available components and secure solutions as well as a fault-tolerant design, enterprises can achieve the availability needed to support 24x7 uptime requirements.

In vital networks such as SANs, with their associated hosts, fabric, and storage components, as well as software applications, downtime can occur even if parts of the system are highly available or fault tolerant. To improve business continuance under a variety of circumstances, SANs can incorporate redundant components, connections, software, and configurations to minimize or eliminate single points of failure.

Implementing multiple levels of redundancy throughout a SAN environment can reduce down-time by orders of magnitude. For instance, hardware components, servers, storage devices, network connections, and even the storage network itself can be completely redundant. A fundamental rule for improving fault tolerance is to ensure multiple paths through separate components regardless of a vendor's assurances of high availability. This is especially true when physical location and disaster tolerance are concerns, or when a complex device can become a single point of failure.

13.2 Servers and host bus adapters

To ensure availability, hosts should include redundant hardware components with dual power supplies, dual network connections, and mirrored system disks typically used in enterprise environments. Hosts should also have multiple connections — two independent connections is a minimum — to alternate storage devices through Fibre Channel switches. In most cases, servers should feature dual-active or hot-standby configurations with automatic failover capabilities.

13.2.1 Path and dual-redundant HBA

The next single point of failure to consider after the server is the path between the server and the storage. Potential points of failure on this path might include HBA failures, cable issues, fabric issues, or storage connection problems. The HBA is the Fibre Channel interconnect between the host and the SAN (replacing the traditional SCSI card for storage connectivity). Using a dual-redundant HBA configuration helps ensure that a path is always available. In addition to providing redundancy, this configuration will enable overall higher performance due to the additional SAN connectivity.

13.2.2 Multiple paths

To achieve fault tolerance, multiple paths are connected to alternate locations within the SAN or even to a completely redundant SAN. Server-based software for path failover enables the use of multiple HBAs, and typically allows a dual-active configuration that can divide workload between multiple HBAs — improving performance. The software monitors the “health” of available storage, servers, and physical paths and automatically reroutes data traffic to an alternate path if a failure occurs.

Path failover

In the event of an HBA or link failure, the host software detects that the data path is no longer available and transfers the failed HBAs workload to an active one. The remaining HBA then assumes the workload until the failed HBA is replaced or the link is repaired. After identifying failed paths or failed-over storage devices and resolving the problem, the software automatically initiates fail back and restores the dual path without impacting applications. If desired, an administrator can manually perform the fail back to verify the process.

The software that performs this failover is typically provided by system vendors, storage vendors, or value-added software developers. Software solutions, such as IBM Subsystem Device Driver (SDD), help ensure that data traffic can continue despite a path failure. These types of software products effectively remove connections, components, and devices as single points of failure in the SAN to improve availability of enterprise applications.

To help eliminate unnecessary failover, the software distinguishes between actual solid failures and temporary network outages that might appear to be solid failures. By recognizing false failures, the software can help prevent unnecessary failover/fallback effects caused by marginal or intermittent conditions. After detecting an actual failure, the software typically waits to determine whether the event is an actual failure.

The typical delay in the failover process can range from an instant failover (when a loss of signal light is detected) up to a minute (if the light signal is still available and the path failure is in another part of the network). These delays are typically adjustable to allow for a variety of configurations and to allow other, more rapid recovery mechanisms such as path rerouting in the SAN.

13.3 Software

One of the keys to improving availability is shifting the focus from server availability and recovery to application availability and recovery. Mission-critical applications should be supported on clustered or highly available servers and storage devices to ensure the applications' ability to access data when they need it — even in the midst of a failure. Sophisticated software applications can enable application or host failover, in which a secondary server assumes the workload if a failure occurs on the primary server. Other types of software, such as many database applications, enable workload sharing by multiple servers — adding to continuous data availability where any one of several servers can assume the tasks of a failed server.

In addition, many server vendors and value-added software providers offer clustering technology to keep server-based applications highly available, regardless of individual component failures. The clustering software is designed to transfer workload among active servers without disrupting data flow. As a result, clustering helps companies guard against equipment failures, keep critical systems online, and meet increased data access expectations.

Some clustering software, such as VERITAS Cluster Server, enables application failover on an application by application basis. This capability enables administrators to prioritize the order of application failover. Fibre Channel SANs facilitate high-availability clustering by simplifying storage and server connectivity. Moreover, SANs can provide one of the most reliable infrastructures for server clustering, particularly when clustered servers are distributed throughout the enterprise to achieve higher levels of disaster tolerance, a practice known as “stretched clusters.”

13.4 Storage

To improve performance and fault tolerance, many of today's storage devices feature multiple connections to the SAN. Multiple connections help guard against failures that might result from a damaged cable, failed controller, or failed SAN component, such as an SFP optical module. The failover process for storage connections typically follows one of the following methods.

Transparent failover

One method is transparent failover, in which a secondary standby connection comes online if the primary connection fails. Because the new connection has the same address as the original failed connection, failover is transparent to the server connection, and application performance is not affected. After the primary connection is repaired, it assumes the workload.

Active connections

Another method is to use dual or multiple active connections with each connection dedicated to certain logical volumes within a given storage system. If one connection fails, the other active connections automatically assume its logical volume workload until it comes back online. During this time, the alternate connections support all logical volumes, so there might be a slight performance impact depending on workload and traffic patterns.

Load balancing connections

A third method used for storage path failover also utilizes dual or multiple active connections. In this case, however, both connections can simultaneously access the logical volumes. This design can improve performance through load balancing, but typically requires host-based software.

During a storage connection failure, the alternate active connection continues to access the logical volumes. After the failed connection is repaired, the other path becomes active and load balancing resumes.

All of these failover methods are designed to ensure the availability of the enterprise applications that use them. In addition, failover generally is coordinated with server software to ensure an active path to data, transparent to the application.

Mirroring

Another effective way to achieve high availability in a SAN environment is by mirroring storage subsystems. SANs enable efficient mirroring of data on a peer-to-peer basis across the fabric.

These mirroring functions contribute tremendous fault tolerance and availability characteristics to SAN-based data. Combining the mirroring functions with switch-based routing algorithms (which enable traffic to be routed around path breaks within the SAN fabric) creates a resilient, self-healing environment to support the most demanding enterprise storage requirements. The mirrored subsystems can provide an alternate access point to data regardless of path conditions.

A common use of mirroring involves the deployment of remote sites within the enterprise. Implementing SANs through Fibre Channel switches enables the distribution of storage and servers throughout a campus, metropolitan area, and beyond. Fibre Channel overcomes many of the distance limitations of traditional SCSI connections, enabling devices to be extended over much longer distances for remote mirroring, tape backup, and disaster recovery operations.

13.5 Fabric

The switched fabric, as the central part of the SAN, is the focus of any discussion about performance and availability. The fabric design should provide a high performing environment for all storage-related enterprise applications and ensure connectivity even during partial outages. By implementing redundancy, the fabric design helps to prevent isolated failures from causing widespread outages and minimizes disruption to system operations.

13.5.1 The fabric-is-a-switch approach

Typically, when one thinks of a director, the assumption is that all the fabric redundancy is consolidated into one box. Theoretically, a director is supposed to provide full internal redundancy. Numbers of critical field replaceable units (FRUs) installed in a director will failover automatically should a component malfunction.

High availability is provided through the hardware and options, such as these:

- Redundancy of all active components
- All active components providing support for automatic failover
- Redundant power and cooling
- Hot swapping of all FRUs
- Hot swapping of spare ports
- Automatic fault detection and isolation
- Non-disruptive firmware updates

An example of this approach is shown in Figure 13-1.

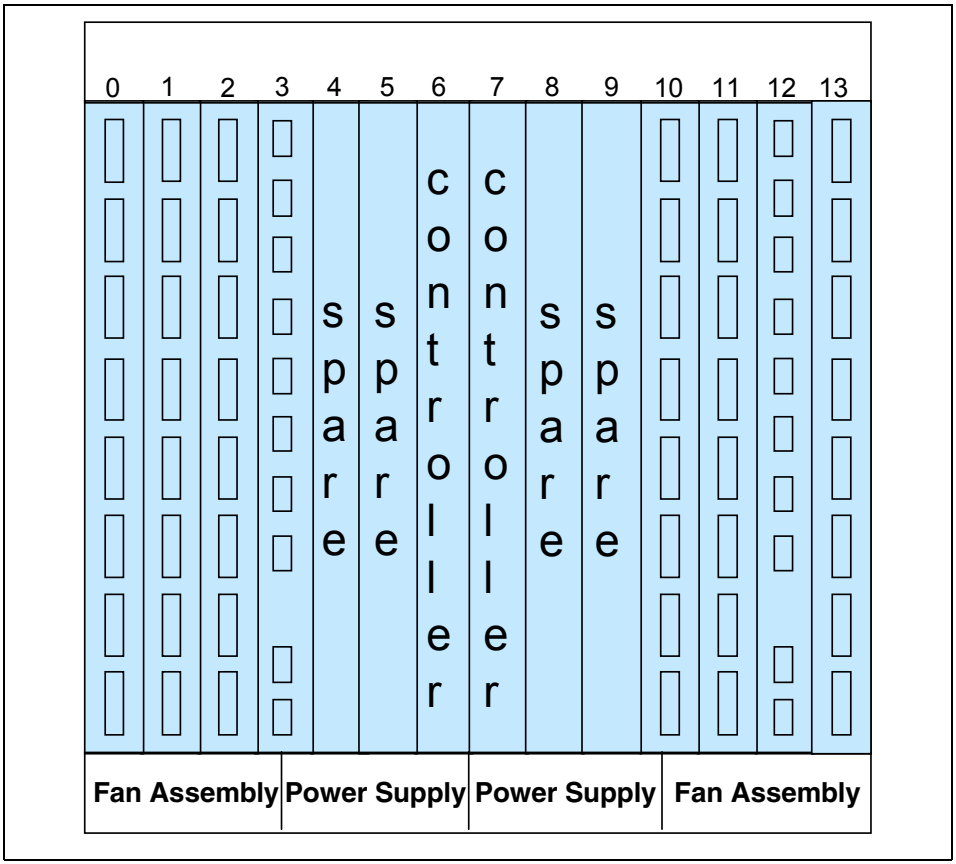


Figure 13-1 Fabric in a director

A director also has a high port density in a single footprint and can usually scale up to an even higher port count. You may build your fabric by implementing one director and have a highly performing and highly available fabric. From a security point of view, a single director is easier to handle and to protect than a widespread fabric, but there is still one single point of failure left, which is the fabric (director) itself. Intentionally or by user error, a fabric can be taken down and therefore the fabric or director should have a backup of its own: a dual fabric.

13.5.2 The fabric-is-a-network approach

Redundancy in the SAN fabric can be built through a network of switches to provide a robust mission-critical SAN solution. With its connected servers, switches, and storage ensuring high availability, the meshed fabric provides a most resilient infrastructure. With an infrastructure of switches, SAN administrators will scale their network to guarantee performance, availability and security by building it into the network rather than relying on a single footprint.

SAN infrastructures require high availability and a high port aggregation to solve problems such as backup and storage consolidation. Since ISLs can be utilized most efficiently nowadays, a network of smaller switches may enable the SAN to support the appropriate level of bandwidth by increasing the number of switches and can be considered as an opposite strategy to the pure director-based SANs.

However, there is no design that is without a limit. To provide every server in the SAN with the appropriate bandwidth to exchange data with its storage, all at the same time, is not a feasible concept. It would mean having to provide ISL bandwidth for the cumulative server port bandwidth, or in other words to give each server its own dedicated ISL — but that is not in the spirit of networking. Ideally, a network should be oversubscribed to the maximum point possible, while maintaining the minimum acceptable performance.

This approach ensures that the fewest resources can support the greatest numbers of users or applications. A typical oversubscription ratio will be 7:1 or higher to start with. During operation you will observe port performance and decide whether to implement more ISLs or more device ports. By taking advantage of the scalability of the SAN switches and ISL trunking features, the switched fabric can be tailored very easily at the first and subsequent implementations.

13.6 High level fabric design

You can configure scalable solutions that help address your needs for high performance and availability for environments ranging from small workgroups to very large, integrated enterprise SANs.

If you start with a single switch, you will find that when your fabric grows, you will need to connect new switches to it. The first step may be a cascaded design. We show two possible options in Figure 13-2.

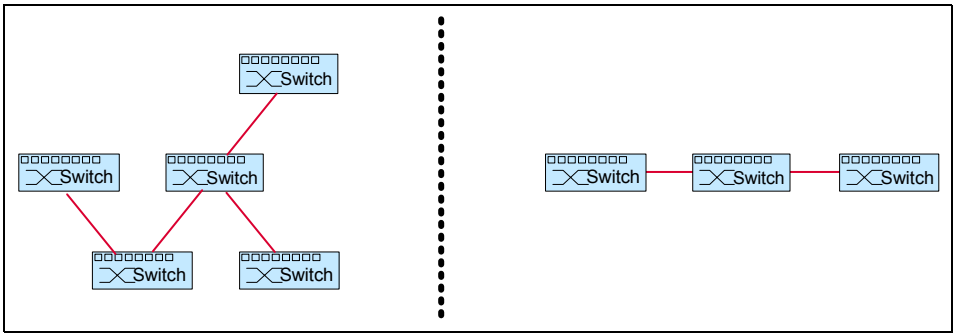


Figure 13-2 Two examples of switch cascading

When cascading switches you will need *(n-1)*-ISLs to connect *n*-switches. It raises your port rate compared to using one switch. With a cascaded fabric it is possible to introduce a bottleneck when traffic has to travel down the ISLs. For this reason there are many ways to ensure that you do not introduce a bottleneck into the SAN.

A next step towards higher performance and higher availability is a ring design, as shown in Figure 13-3.

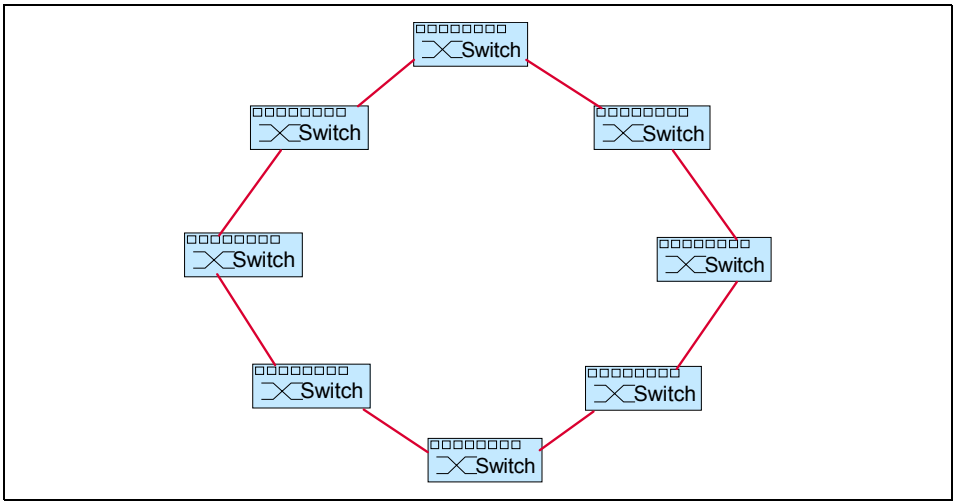


Figure 13-3 Ring design

Whenever one ISL fails, there is still connectivity throughout the whole fabric. However, if an ISL fails, then it may take more hops for the initiator to reach the target. To connect *n*-switches, you'll need *n*-ISLs.

However, these SAN designs do not show very much thought or structure to support the traffic flow. We could dedicate switches to be connected to storage or host only, but all the traffic would have to pass through the ISLs and this may be counter-productive. Also, simply structuring the SAN by dedicating switches to diverse departments will not increase performance nor availability.

The way to increase performance and availability in a fabric is to build a network of switches in a meshed network topology, as shown in Figure 13-4.

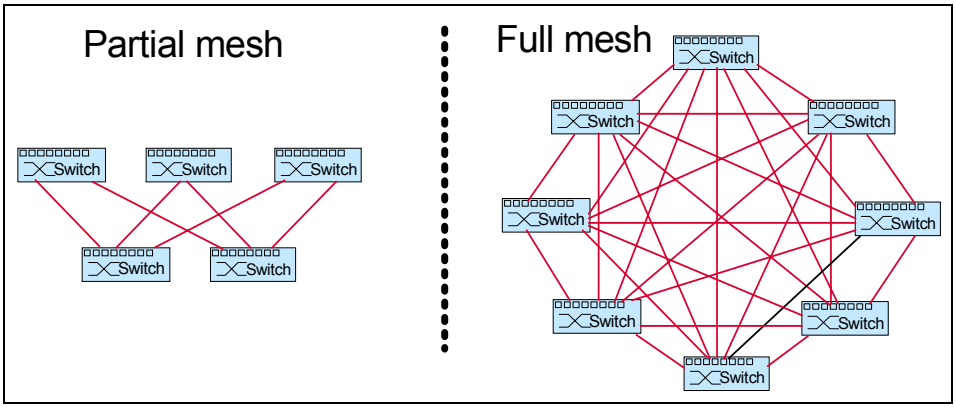


Figure 13-4 Meshed network design

Figure 13-4 shows a partial mesh and full mesh fabric design. For a partial mesh fabric, you will need at least *n*-ISLs to connect *n*-switches. For a full mesh fabric, you will need *m*-ISLs to connect *n*-switches, as follows:

$$m = \frac{n^2 - n}{2}$$

The number of ISLs increases very quickly when the count of switches increases. A full mesh network offers the highest performance and availability.

With either type of meshed topology, it is easy to structure the fabric for the sake of easier maintenance and administration, fault isolation, and higher traffic flow.

A common structure is a tier-layer design with a dedicated layer of switches for hosts and a layer for storage devices, as shown in Figure 13-5.

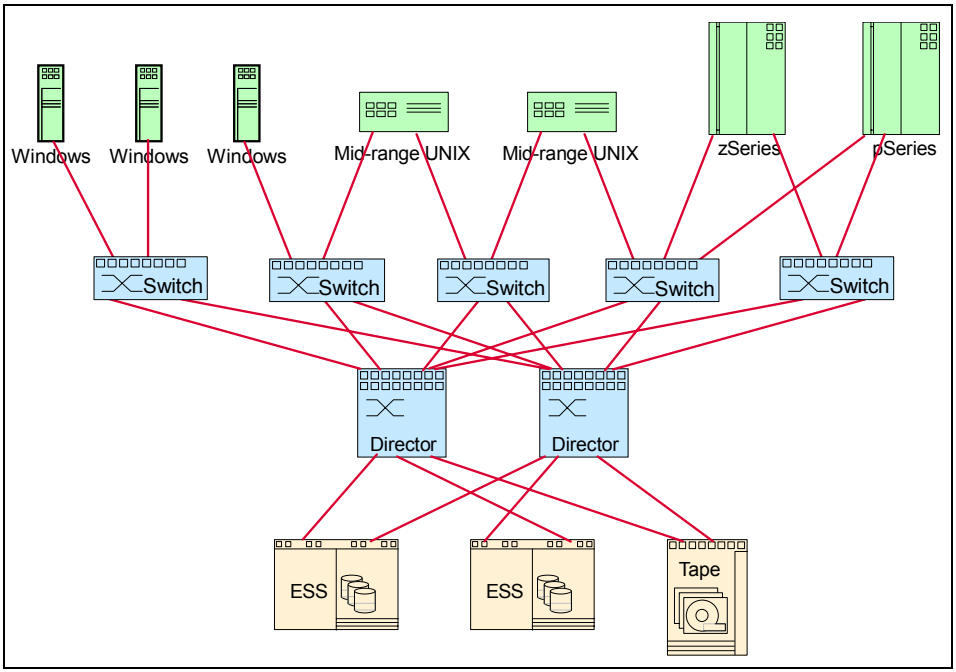


Figure 13-5 Host-tier and storage-tier

It is a partial mesh design with all hosts connected to the upper tier, and all storage devices to the lower one. Every data transfer from host to storage will cross the ISLs and we have to keep that in mind when provisioning the ISLs.

If the SAN was to grow bigger and there was eight switches in each tier, in order to connect every switch in the host-tier to every switch in the storage tier, it would cost 128 ISL ports, as shown in Figure .

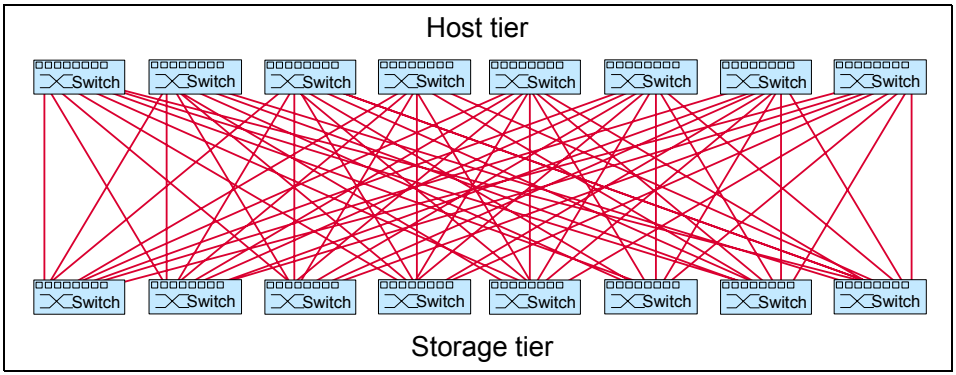


Figure 13-6 Tier to tier

This is just done to single-connect the switches from one tier to the other. The fabric does not gain any higher availability by such a design.

A better way to connect the tiers would be to introduce a “focal point” for all ISLs between the tiers, called core switches; the switches at the host and storage tier are called edge switches. We now have a core-edge design like that shown in Figure 13-7.

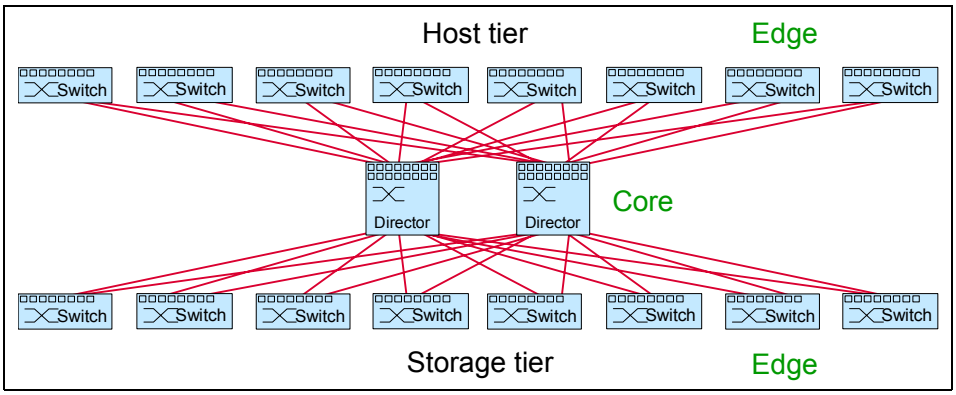


Figure 13-7 Core-edge design

With the core-edge design, you will have any-to-any edge-switch connectivity without having to connect any-to-any switch. We do have less cumulative ISL bandwidth in the SAN now, when compared to the design in Figure .

As the core is the focal point, you will want to deploy your core switches or directors that have redundancy inherent in their design. A storage-tier usually needs less ports than a host-tier. So, in some cases, when your storage devices are pooled locally, you don’t build up a separate edge-tier for storage, as shown in Figure 13-7. Rather, the ports of the core switches connect to the storage devices directly, as shown in Figure 13-5.

13.7 Definitions

In our examples in this chapter and in the following chapters, we will use the following terms:

- **Oversubscription:** This term means the ratio of the number of input devices weighted by their individual bandwidth to the number of output devices also weighted by their individual bandwidth. That may be the amount of hosts connecting to a storage device. For example, a storage device can handle up to 100 MB/s on one port, connecting four servers which will do 60 MB/s each will give us an oversubscription of 2.4:1:

$$oversubscription = \frac{\sum port_{input} \times bandwidth}{\sum port_{output} \times bandwidth} = \frac{4 \times 60\text{MB/sec}}{100\text{MB/sec}} = \frac{24}{10}$$

- **ISL-oversubscription:** This special case of oversubscription takes the ratio of host ports to the possible ISLs carrying that traffic - again we take bandwidth of the individual ports into account. ISL-oversubscription is of interest in a meshed fabric. The higher the ratio is, the more devices will share the same ISL and the more is it likely we will suffer from congestion. Adding a 2 Gb/s ISL (that is 200 MB/s) to our previous example will give us a value of 1.2:1:

$$\text{ISL-oversubsription} = \frac{\sum port_{input} \times bandwidth}{\sum ISL \times bandwidth} = \frac{4 \times 60\text{MB/sec}}{200\text{MB/sec}} = \frac{240}{200}$$

Note: If all ports on the switches are operating with the same speed, it is a simple division to calculate ISL-oversubscription. In cases where ports with 2 Gb/s and 1 Gb/s are intermixed, each host port and each ISL has to be multiplied by its bandwidth before computing the ratio.

- **Fan-out:** This is the ratio of server ports to a connected storage port. Fan-out differs from oversubscription in that it represents a ratio based on the number of connections regardless of the throughput. Our previous example would result in a fan-out of 4:1:

$$\text{Fan-out} = \frac{\sum port_{host}}{\sum port_{storage}} = \frac{4}{1}$$

- **Topology:** This is a synonym for design in networking. Fibre Channel historically supports only three topologies: point-to-point, arbitrated loop, and switched fabric, but this is *not* what is meant here. Topology in our solutions means the design itself of the switched fabric and whether it is a cascaded, meshed, tiered, or a core-edge design.

13.7.1 Port formulas

As we have stated already, oversubscription is an accepted bottleneck. Depending on the load profile of your Fibre Channel devices, you will accept these bottlenecks in such a way that they will never allow the SAN to end up in gridlock.

Typically the first implementation of a SAN is based on assumptions. By using this port formula, you may estimate how many host ports you will get out of a two-tier fabric with a given number of switch ports, and estimated values for host-to-storage oversubscription ($over_{hs}$) and ISL-oversubscription ($over_{ISL}$):

$$\sum port_{host} = \frac{\sum port_{fabric} - \sum port_{spare}}{1 + \frac{1}{over_{hs}} + \frac{2}{over_{ISL}}}$$

As an example, we will use the diagram Figure 13-5 on page 577. We will take seven 32-port switches and assume that we want to use only 80% of the ports in our first implementation, saving the other 20% for future expansions. We will assume a host-to-storage oversubscription of 6:1 and an ISL-oversubscription of 10:1. This is shown in Example 13-1.

Example 13-1 Host-to-storage oversubscription

5 switches with 32 ports each: $port_{fabric}$ = **160**
20% spare ports: $port_{spare}$ = **32**
oversubscription host-to-storage: $over_{hs}$ = **6:1**
ISL-oversubscription: $over_{ISL}$ = **10:1**

$$\sum port_{host} = \frac{\sum port_{fabric} - \sum port_{spare}}{1 + \frac{1}{over_{hs}} + \frac{2}{over_{ISL}}} = \frac{160 - 32}{1 + \frac{1}{6} + \frac{2}{10}} = 93$$

So:
The host ports come to: $port_{host}$ = **93**
The storage ports will be $port_{storage}$ (93 / 6) = **15**
The ISLs will be (93 / 10) = **10,** which is **20** ISL ports

The assumptions about oversubscription in the previous example will delegate the used ports of the fabric to 73% as host ports, 12% as storage ports, and 15% as ISL ports.

You can use that formula for a core-edge design too (see Figure 13-7 on page 578). If you ignore the core ports and just count the ports of your edge-switches to get $\Sigma port_{fabric}$, simply calculate the other port counts and assume that your core ports are the same value as the ISL ports. So, referring back to Example 13-1 on page 580, your core would consist of another 20 ISL ports.

13.8 Our solutions

In the following chapters we categorize and discuss the relevant items of SAN design in various implementations with different switches and directors. We have categorized the solutions according to:

- Performance
- Availability
- Clustering
- Secure

We have also added any vendor specific and unique solutions. Although we categorize, for example, a solution as a *performance* solution, this does not mean that it is only a performance solution. It will typically contain elements of all the other categories as well, but we have just chosen to focus on one aspect for clarity. A *Checklist* and *"What If" failure scenario* complements each solution.

14

IBM TotalStorage SAN b-type family solutions

In this chapter we illustrate and describe solutions that can be implemented using the IBM TotalStorage switches in diverse environments.

We will base our solutions on the availability characteristics of each product described in Chapter 8, “IBM TotalStorage SAN b-type family” on page 287. The solutions are categorized as follows:

- Performance solutions
- Availability solutions
- Clustering solutions
- Secure solutions
- Loop solutions

© Copyright IBM Corp. 2004. All rights reserved.

14.1 Performance solutions

When designing a SAN careful thought must be given on how to design the SAN so that performance does not suffer. One concept that needs to be evaluated is oversubscription.

One type of oversubscription would be the number of servers associated with a storage port. It is very difficult to work out the ratio of server ports to storage ports. The solution we show in Figure 14-1 illustrates how a general high performance profile could be applied to a SAN design using a director and a single IBM TotalStorage Enterprise Storage Server.

If we do not have accurate performance data from the servers we need to employ a high level methodology to help come up with a baseline. This methodology should only be used to generate a high level design. Final designs must be based on performance data collected from the servers.

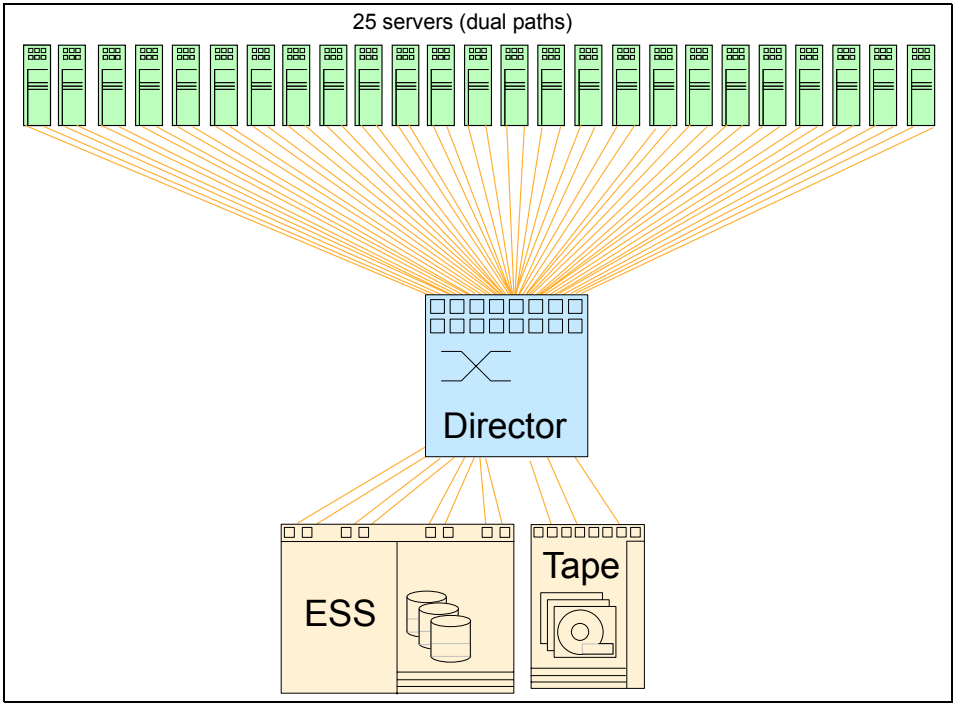


Figure 14-1 High performance design

Fibre Channel will operate at up to 200 MB/s. Of course this depends on many factors such as the type of data access, whether it is read or write intensive, the blocksize of the data, etc. For this example we will assume that the average is

130 MB/s. If we configured eight connections from the director to ESS, we would have a maximum SAN peak bandwidth capability of 1040 MB/s (8 x 130 MB/s).

If we connected 25 dual attach servers to a 2109-M14, and all servers were processing at the same time, we would potentially have a maximum SAN peak bandwidth of 41.6 MB/s per server (1040 MB/s / 25).

This throughput assumes that all 25 servers are able to generate this level of I/O at the same time. This could be categorized as a *high performance profile*.

Based on this theory, for a high performance profile, we have a server connection to ESS port ratio of 6.25 which we round down to 6. So our ratio in this case is 6:1.

Note: The high performance profile is calculated by determining the ratio between the number of server ports (or HBAs) and ESS Fibre Channel ports.

In our example above: 25 servers with dual paths = 50 server ports / 8 ESS ports = ratio of 6.25:1

For *low performance profiles*, such as file and print servers, we will use a rule-of-thumb of 12 server connections to one ESS port. In this case we would use a ratio of 12:1.

Tape device functions such as serverless backup and/or the servers the tape device is connected to must be taken into consideration as well. For best performance, separate HBAs should be used for tape access.

These profile ratios are recommended as a starting point when there are no server performance details available. These rules are very generic and should only be applied at the initial design stage. Prior to any final design a detailed performance profile should be conducted using open systems performance measuring tools such as IOMETER and IBM's Disk Magic.

In our solution we will connect 25 dual attach high performance profile servers to a single ESS.

Components

- SAN fabric:
  - IBM TotalStorage SAN Switch M14 configured with 64 ports
- Servers:
  - 25 servers each configured with dual FC HBAs.
- Storage:

  - One IBM TotalStorage Enterprise Storage Server Model 800 configured with 8 x Fibre Channel adapters
  - 3584 Automated Tape Library configured with 6 x FC drives
- Software:
  - SDD installed on servers
  - IBM StorWatch Enterprise Storage Server Specialist for configuring the ESS
  - 3584 with StorWatch Specialist
  - IOMETER and Disk Magic for performance modelling

Checklist

We checked the following items:

- Ports within ‘octets’ for optimum server to storage performance
- Spread dual connected ports across port blades to minimize the affect of a card failure within the director
- Spread ESS connections across different port blades
- Consider the impact of losing a port blade and balance the server groups to minimize impact
- Leave some ports spare for contingency
- Monitor the performance using Tivoli
- Collect MIB information to determine busy ports
- Conduct a detailed server performance profile
- All storage devices, server HBAs, switches, or directors are configured with the latest IBM supported versions of drivers/firmware levels

Performance

As detailed in our solution description a detailed server performance profile needs to be undertaken to be confident that there is no under- or over-utilization of the SAN’s bandwidth. Due to the performance of the director any SAN performance bottlenecks will likely be at the ISLs (if configured), or more likely at the HBAs of the storage device.

Based on this theory, the performance of the SAN will be determined on how much traffic will be moved through the E_Port or HBA. With detailed server profiles it is possible to balance this accordingly.

Scalability

Based on our performance profiling we could expand our solution and connect two directors together using dual E_Ports, as shown in Figure 14-2. Each director now has four connections to the ESS and 3 connections to the tape library. We have now created a higher availability SAN that could support 100 device connections (assuming 50 servers with dual HBAs) attached. This design provides protection against any possible complete failure of a director.

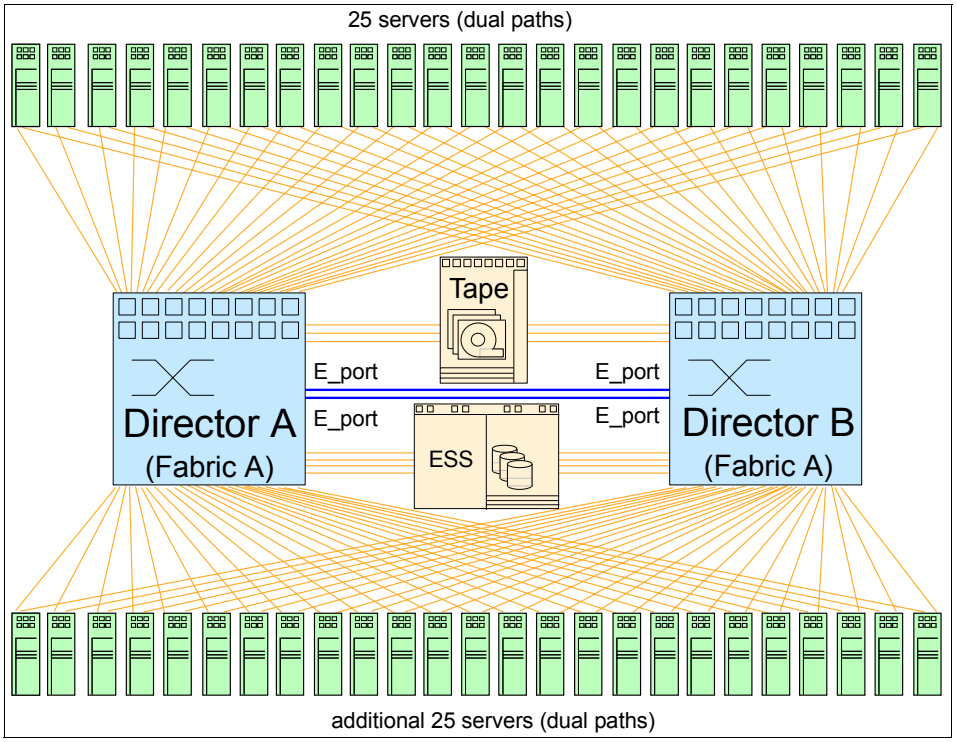


Figure 14-2 Expanding the SAN fabric via E_Ports

As you can see we have doubled the number of servers without changing the number of storage ports. This has increased the server port to storage port ratio to 12:1, and reduced the maximum SAN server bandwidth to 20.8 MB/s per server. This design is a much more cost effective solution.

Availability

While this design provides a higher-availability design than for the single director model a failure in the SAN fabric could result in all hosts losing access to the devices. For example, if an invalid zoning change was made to the fabric or the fabric was corrupted this would affect all devices in the SAN.

Security

We have not considered any security issues with this solution. These will be addressed in some of the following solutions.

“What If” failure scenarios

Here we consider the following scenarios:

- **Server HBA:** If one of the HBAs fails, IBM SDD software will automatically failover the workload to the alternate HBA. The server will lose up to 50% of the server SAN bandwidth. Once you have replaced the HBA, you will need to redefine the WWN for the HBA.
- **Cable:** If a cable between a server and the switch fails, IBM SDD software will automatically failover workload to the alternate path. The server SAN performance will be degraded by up to 50%.
- **Cable:** If a cable between the switch and the disk storage fails, the alternate route will be used. The overall SAN performance will degrade by up to 6.25%. On the servers connected to the affected switch, the SAN performance impact will be up to 50%. Servers connected to other storage will not be affected.
- **Switch port:** If one of the ports fails, you may replace it using a hot-pluggable SFP.
- **Switch power supply:** Another redundant power supply may be added to the switch, and should one fail, the other will take over automatically.
- **Switch:** If a CP module fails, there would be no effect as the spare CP module will automatically take over.
- **Switch:** If the backplane was damaged, we would lose connectivity to all servers at that site. The solution in Figure 14-2 provides protection against a complete director failure.
- **Storage:**If an ESS HBA is unavailable we have multiple other connections that will automatically be used. We would lose 12.5% of the available SAN bandwidth per HBA.
- **Storage:**If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used. We would lose 25% of the available bandwidth per bay. The ESS has 4 adapter bays each holding up to 4 interface cards. In our example we have configured the ESS with 2 x FC adapter cards per bay.
- **Storage:** If the ESS fails, the servers will not be able to access the storage. A redundant ESS may be added to mirror data from the primary ESS, or data may be restore from a tape subsystem device.
- **Fabric:** For both Figure 14-1 and Figure 14-2, a failure in the SAN fabric itself will cause a loss of connectivity for all devices.

14.2 Availability solutions

The focus of this topic is on the availability aspect of solutions.

14.2.1 Single fabric

In Figure 14-3 we show a single fabric, one 2109-M14 core withtwo 2109-F16 edge switches.

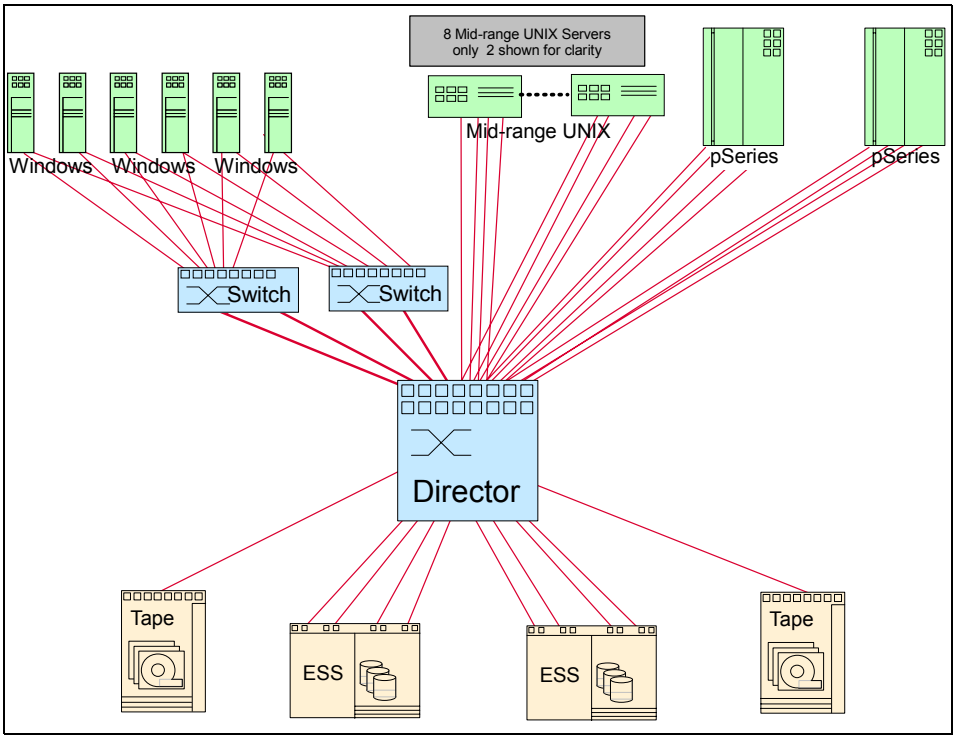


Figure 14-3 2109-M14 core-edge solution

Components

- SAN fabric:
  - One 64-port 2109-M14 director
  - Two 16-port 2109-F16 switches
- Servers:
  - Six xSeries servers each configured with dual FC HBAs
  - Eight UNIX servers each configured with four FC HBAs
  - Two pSeries servers each configured with four FC HBAs
- Storage:
  - Two ESS 2105-800 with four FC Adapters each
  - Two IBM 3590 Tape Drives with FC Adapter
- Software:
  - IBM Subsystem Device Driver (SDD)

Checklist

- Install and configure switches
- Install Fibre Channel HBAs
- Configure ESS
- Attach ESS to switch
- Attach servers to switch
- Validate failover/fail back operations
- All storage devices, server HBAs, switches, or directors are configured with the latest IBM supported versions of driver/firmware levels.

Performance

Just taking the disk storage into account, the server to storage oversubscription of this design is 6.5:1 (52 / 8). Windows-tier and core is connected by 4 ISLs. That gives us an ISL oversubscription of 3:1 (12 / 4), and we will check the utilization over a set time frame to decide whether to add more ISLs. Assuming the ESS FC adapter will operate at up to 130 MB/s. With the eight connections from the fabric to the storage, we get a maximum SAN peak bandwidth capability of 1040 MB/s (8 x 130 MB/s). This design can support hundreds of end-node connections with high throughput.

Scalability

Using the 2109-F16 switches at the edge of the fabric in our example, allows for reduced cabling at the director and lower cost ports for the lower utilized severs. This also allows for further edge switches to be added in the same manner, with a minimum useage of director ports. As we have only used 50 ports on the 2109-M14, we still have room for adding ISLs or high end server or storage ports.

This design is highly scalable, the 2109-M14 provides a very high port density and additional blades can be added. The 2109-F16s may be replaced by the 2109-F32 to increase the port count at the edge.

Availability

This design is appropriate when the server to storage connections need to be highly available; a single switch can fail or be taken off-line for maintenance such as firmware upgrade, and the fabric will still support all the connected devices, although there may be a reduced performance.

The 2109-M14 supports a concurrent firmware upgrade since it has CP failover due to its two redundant CPs. Failed hardware can also be replaced concurrently and extra ports can be added.

Security

The ESS performs LUN masking by default so all devices with LUNs defined have two levels of security; LUN masking and zoning.

If further security within the fabric is required, install the optional license “Advanced Security” on page 307.

“What If” failure scenarios

Here we consider the following scenarios:

- **Server HBA:** If one of the HBA fails, IBM SDD software will automatically failover the workload to the alternate HBA. A Windows server will lose up to 50% of the server SAN bandwidth. An UNIX/AIX server will lose up to 25% of the server SAN bandwidth. Once you have replaced the HBA, you will need to redefine the WWN for the HBA.
- **Cable:** If a cable between a server and the switch fails, IBM SDD software will automatically failover workload to the alternate path. A Windows server will lose up to 50% of the server SAN bandwidth. An UNIX/AIX server will lose up to 25% of the server SAN bandwidth.
- **Cable:** If a cable between the switch and the disk storage fails, the alternate route will be used. The overall SAN performance will degrade by up to 6.5%. On the servers connected to the affected switch, the SAN performance impact will be up to 50% (up to 25% for the UNIX/AIX servers). Servers connected to other storage will not be affected.
- **Cable:** If one of the ISLs breaks, an alternate route will be used based on FSPF. The SAN performance will degrade by up to 25%.
- **Switch port:** If one of the ports fails, you may replace it using a hot-pluggable SFP.
- **Switch power supply:** Redundant power supplies are already provided by the 2109-M14 and may be added to the 2109-F16. Should one power supply fail, another will take over automatically.
- **Switch:** If a switch at the Windows-tier fails, the server will use the alternate switch to connect to the disk storage. The performance of the servers connected to the switch will be affected by up to 50%.
- **Director:** If the 2109-M14 chassis fails, we lose all connectivity from the servers to storage and can be considered a fabric failure.
- **Storage:** If one ESS fails, the servers connected will not be able to access the storage. The other ESS may be used to mirror data from the primary ESS, or data has to be restored from a tape subsystem device.

14.2.2 Dual fabric

In Figure 14-4 we show a two-tier highly available enterprise SAN design with redundancy and failover, and a multi-stage switch interconnect to allow for many-to-many connectivity.

Typically, a two-tier design has a host-tier and a storage-tier. All hosts are connected to the host-tier and all storage is connected to the storage-tier. A path from a host to its storage is always just a single hop away.

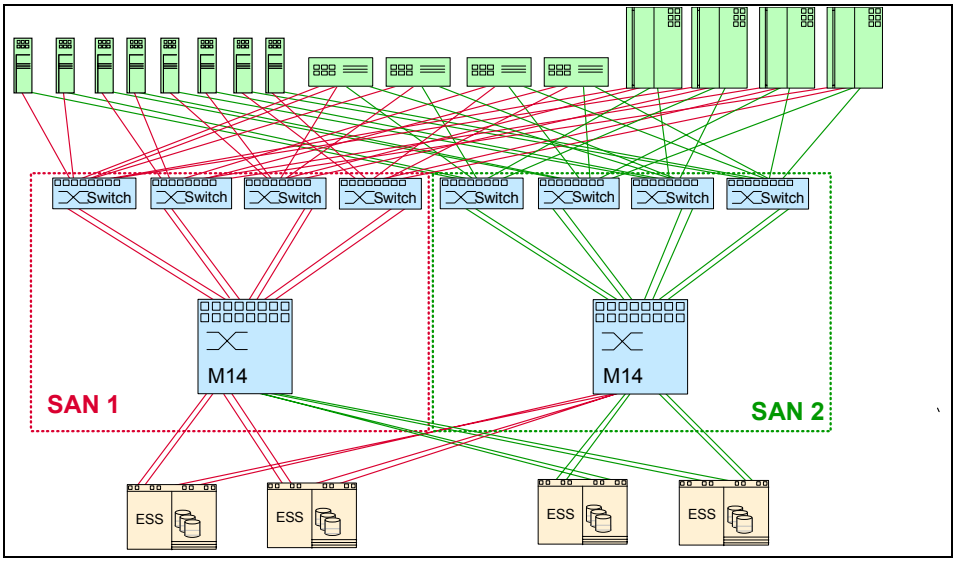


Figure 14-4 High availability dual enterprise SAN fabric

As mentioned earlier, a single SAN is still a single point of failure: it could be affected by a number of events including the following:

- Incorrect zoning change
- Site outages
- Firmware failures

By implementing a solution based on dual fabrics, we can avoid the impact of a SAN fabric failure. Two separate fabrics have been implemented, each one with a 2109-M14 director as the storage-tier and multiple 2109-F32s for host connection tier.

Components

- SAN fabric:
  - Two 128-port switches 2109-M14
  - Eight 32-port switches 2109-F32
- Servers:
  - Eight Windows 2000 servers each configured with dual FC HBAs
  - Four UNIX servers each configured with four FC HBAs
  - Four pSeries servers each configured with four FC HBAs
- Storage:
  - Four ESS 2105-800s with four FC adapters each
- Software:
  - IBM Subsystem Device Driver (SDD)

Checklist

- Install and configure switches
- Install Fibre Channel HBAs
- Configure ESS
- Attach ESS to switch
- Attach servers to switch
- Install IBM Subsystem Device Driver (SDD)
- Validate failover/fail back operations
- All storage devices, server HBAs, switches, or directors are configured with the latest IBM supported versions of driver/firmware levels

Performance

The server to storage port oversubscription of this design is 3:1 (48 / 16), which is well below the recommended 6:1. The host-tier and storage-tier is connected by 16 ISLs and gives us an oversubscription of 48:16, that is 3:1 and no bottleneck to the end-nodes. Assuming the ESS FC adapters operate at up to 130 MB/s, with the 16 connections from the fabric to the storage, we get a maximum SAN peak bandwidth capability of 2080 MB/s (16 x 130 MB/s). This design can support hundreds of end-node connections with high throughput.

Scalability

A dual SAN fabric is a topology where you have two independent SANs that connect the same hosts and storage devices. All hosts and storage must be connected to both switches to achieve high availability. The high port density of the switches makes the dual fabrics highly scalable.

We connected 48 host ports to the fabric and 16 storage ports. Each tier provides 256 ports. If we allow for the ISL connections to the 8 switches to be implemented as trunks of 4 ISLs each, this would leave us with 176 available ports at the host-tier, and to 192 available ports for the storage-tier.

Availability

This design is appropriate when the fabric itself needs to be highly available; a single switch can fail or be taken off-line for maintenance, and the fabric will still support all the connected devices (devices do require one redundant entry point to each fabric). Firmware upgrades on the 2109 products used in these fabrics may be activated concurrently, providing a very high availability solution.

"What If" failure scenarios

In addition to the previous solution:

- **Director:** If the 2109-M14 chassis fails in a fabric, the overall SAN performance will degrade by up to 50%.

14.3 Clustering solutions

These topics focus on clustering solutions.

14.3.1 Two-node clustering

In Figure 14-5 we show a typical basic high availability SAN design for a two-node clustering and redundant fabric. This design is typically for a small fabric with two to four hosts.

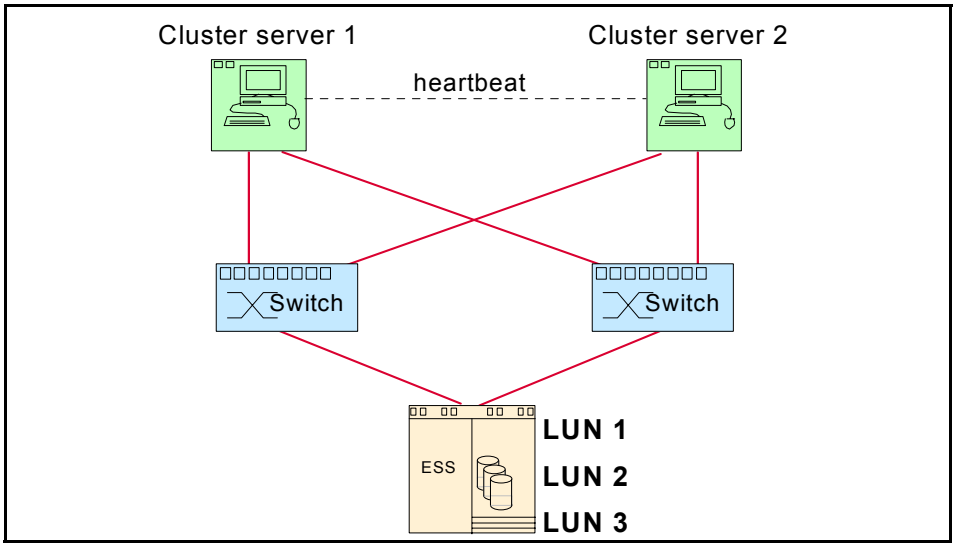


Figure 14-5 Microsoft cluster with dual switch with redundant fabric

Components

- SAN fabric:
  - Two 16-port IBM SAN Fibre Channel Switch Model 2109-F16
- Servers:
  - Two IBM xSeries servers each configured with dual FC HBAs
- Storage:
  - One ESS 2105-800 with two native FC Adapters,
- Software:
  - IBM Subsystem Device Driver (SDD)

Checklist

- Install and configure switches
- Install Fibre Channel HBAs
- Configure ESS
- Attach ESS to switch
- Attach servers to switch
- Install IBM Subsystem Device Driver (SDD)
- Validate failover/failback operations
- All storage devices, server HBAs, switches, or directors are configured with the latest IBM supported versions of driver/firmware levels

Performance

Typically, for a low performance server, the recommended server to storage oversubscription is 12:1, and for a high performance server, the server to storage oversubscription is 6:1. With a 2:1 ratio (4/2), the above configuration is within ratio provided based on four server connections to two storage connections.

To increase the performance of the SAN, multiple connections may be added from the hosts to the switches and from the switches to the storage devices.

Scalability

A dual SAN fabric is a topology where you have two independent SANs that connect the same hosts and storage devices. This design is not one of the most highly scalable as all hosts and storage must be connected to both switches to achieve high availability.

Availability

In addition to the server high availability clustering, SAN high availability is provided with this dual switch, dual fabric design. Dual HBAs are installed in each host and the storage device must have at least two ports. Fail over for a failed path or even a failed switch is dependent on host failover software, namely, the IBM Subsystem Device Driver (SDD). The switches do not reroute traffic for a failed link as there is no fabric or meshed network with this type of design. Each switch is a single-switch fabric.

Security

The ESS performs LUN masking by default so all devices with LUNs defined have two levels of security, LUN masking and zoning.

If further security within the fabric is required, installing the optional license “Advanced Security” on page 307.

“What If” failure scenarios

Here we consider the following scenarios:

- **Server:** The clustering solution will failover to the other server dynamically.
- **Server HBA:** If one of the HBA fails, IBM SDD software will automatically failover the workload to the alternate HBA. The active server of the cluster will lose up to 50% bandwidth.
- **Cable:** If a cable between a server and the switch fails, IBM SDD software will automatically failover workload to the alternate path.The active server will lose up to 50% bandwidth.

- **Cable:** If a cable between the switch and the disk storage fails, an alternate route will be used. The active server will lose up to 50% bandwidth.
- **Switch port:** If one of the ports fails, you may replace it using a hot-pluggable SFP.
- **Switch power supply:** Another redundant power supply may be added to the switches, and should one fail, the other will take over automatically.
- **Switch:** If a switch fails, the server will use the alternate switch to connect to the storage. The active server will lose up to 50% bandwidth. Once the switch is replaced, all zoning information will be automatically propagated from the other switch.
- **Storage:** If the ESS fails, the servers will not be able to access the storage. A redundant ESS may be added to mirror data from the primary ESS.

14.3.2 Multi-node clustering

In Figure 14-6 we extend the previous environment in shown in Figure 14-5. It is extended to increase the number of Server nodes to up to eight in a single cluster using Datacenter Server 2000 for Microsoft Cluster Server and connects to two 2109-F16 switches with redundant fabric, which allows access to the two ESSs.

Microsoft 2000 clustering: The Windows Server 2003 family supports server clusters for up to eight nodes. Refer to http://www.microsoft.com for more details.

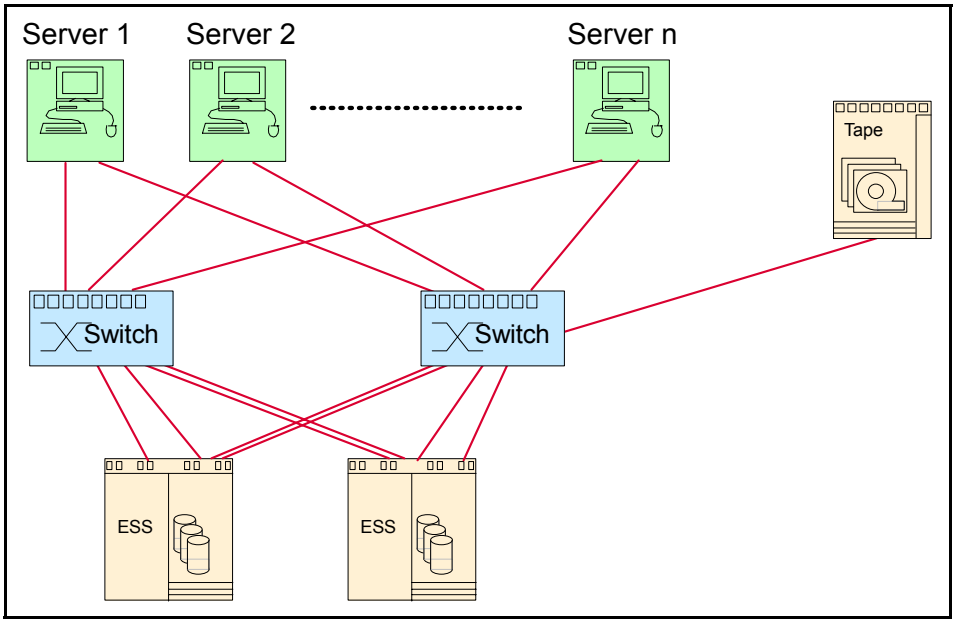


Figure 14-6 Datacenter Server 2000 for MSCS

Apart from server failover, this design provides failover for HBAs and switches. Dual HBAs are installed in each host node and each storage device must have at least two ports. Fail over for a failed path or even a failed switch is dependent on the host failover software, namely the IBM Subsystem Device Driver (SDD).

Components

- SAN fabric:
  - Two 2109-F16s
- Servers:
  - Eight IBM xSeries servers each configured with dual FC HBAs
- Storage:
  - Two ESS 2105-800 with four FC adapters
  - IBM 3590 Tape Subsystem with FC Adapters
- Software:
  - IBM Subsystem Device Driver (SDD)

Checklist

- Install and configure switches
- Install Fibre Channel HBAs
- Install and certify Datacenter Server 2000

- Install IBM Subsystem Device Driver (SDD)
- Configure tape subsystem
- Configure ESS
- Attach storage to switch
- Attach servers to switch
- Validate failover/failback operations
- All storage devices, server HBAs, switches, or directors are configured with the latest IBM supported versions of driver/firmware levels

Performance

This solution is configured based on the ratio of eight servers to two storage devices with a redundant fabric. Hence, the effective oversubscription will be 2:1 and is within our predefined 6:1 ratio.

Scalability

A dual SAN fabric is a topology where you have two independent SANs that connect the same hosts and storage devices. This design is not the highest scalable as all hosts and storage must be connected to both switches to achieve high availability.

This design accommodates eight hosts and two storage devices with redundancy with two F16s. The additional ports may be used to connect to other storage devices.

Availability

In addition to the server high availability clustering, SAN high availability is provided with this dual switch, dual fabric design. Dual HBAs are installed in each host node and the storage device must have a minimum of two ports. Fail over for a failed path or even a failed switch is dependent on host failover software, namely, the IBM Subsystem Device Driver (SDD). The switches do not reroute traffic for a failed link as there is no fabric or meshed network with this type of design. Each switch in this example is a single-switch fabric.

Security

The ESS performs LUN masking by default so all devices with LUNs defined have two levels of security, LUN masking and zoning.

If further security within the fabric is required, installing the optional license “Advanced Security” on page 307.

“What If” failure scenarios

Here we consider the following scenarios:

- **Server node:** The clustering solution will failover to another server within the cluster dynamically
- **Server HBA:** If one of the HBA fails, IBM SDD software will automatically failover the workload to the alternate HBA. The active server of the cluster will lose up to 50% bandwidth
- **Cable:** If a cable between a server and the switch fails, IBM SDD software will automatically failover workload to the alternate path.The active server will lose up to 50% bandwidth
- **Cable:** If a cable between the switch and the disk storage fails, an alternate route will be used. The active server will lose up to 50% bandwidth
- **Switch port:** If one of the ports fails, you may replace it using a hot-pluggable SFP
- **Switch power supply:** Another redundant power supply may be added to the switch, and should one fail, the other will take over automatically
- **Switch:** If a switch fails, the server will use the alternate switch to connect to the storage. The active server will lose up to 50% bandwidth. Once the switch is replaced, all zoning information will be automatically propagated from the other switch.
- **Storage:** If one ESS fails, the servers will not be able to access that storage. The other ESS may be used to mirror data from the primary ESS.

14.4 Secure solutions

The following example uses our previous solution “Single fabric” on page 589 and implemented here as a secure solution.

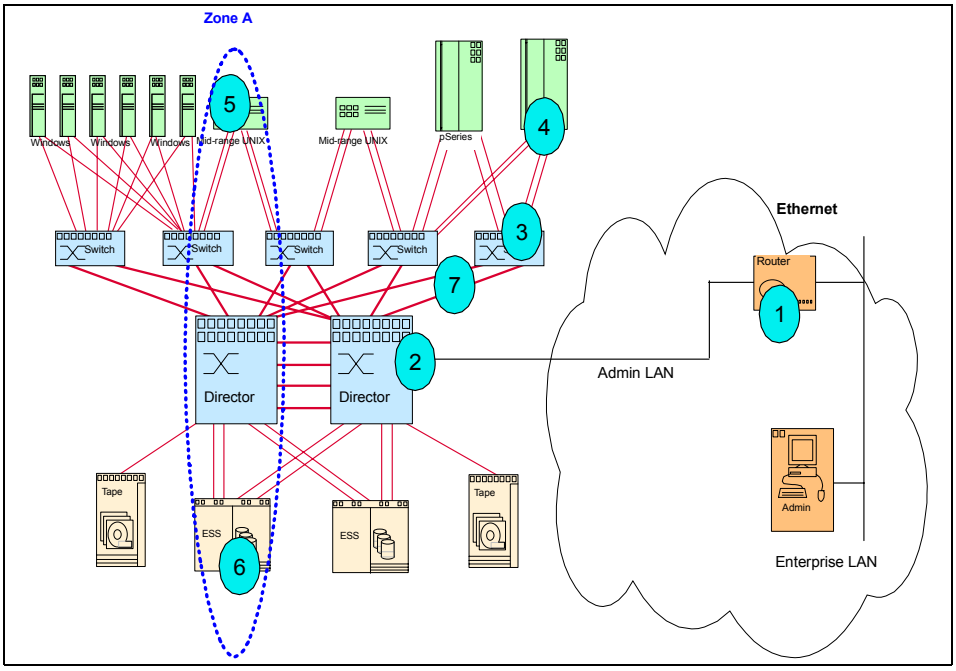


Figure 14-7 Secure SAN

Checklist

- Install and configure switches with Secure Fabric OS
- Install Secure telnet client and firewall
- Validate security functions
- All storage devices, server HBAs, switches, or directors are configured with the latest IBM supported versions of driver/firmware levels

"What If" violation scenarios

Here we consider the following scenarios:

- **Ethernet:** If unauthorized IP sessions are attempted to be established, firewall (1) will protect the admin LAN. The SAN administrator will connect from the enterprise LAN to the SAN director on the Admin LAN via the firewall. The Secure telnet session fully encrypts the data stream, including passwords, between the source and destination, so it cannot be directly read with a LAN sniffer. The administrator can now access the 2109-M14 (2). Denial of Service attacks and broadcast storms in the enterprise SAN are blocked at the firewall boundary.
- **Configuration changes:** If any change is to made, the 2109-M14 (2), as a *trusted switch*, acts as the Fabric Configuration Server and is responsible for

managing the zoning configuration and security settings of all other switches in the fabric. This prevents unauthorised switches from being connected to the fabric and in turn make unauthorized changes.

- **Device connection:** With Access Control Lists (ACLs) individual device ports can be bound to a set of one or more switch ports (3). Any device not specified in the ACL will not be able to log into the fabric (4).
- **Data traffic:** Any FC device (5) trying to attach to another device (6) in Zone A will be checked by hardware. If not authorized, access will be denied.
- **Switch connection:** When a new switch is connected to a switch that is already part of the fabric (7), both switches must be authenticated. This makes sure that only authorized switches may join the fabric.

The details for components, performance, scalability and availability remain the same as in 14.2.1, “Single fabric” on page 589.

14.5 Loop solutions

QuickLoop is used to connect a *non-fabric* aware host to a SAN. In Figure 14-8, we configured three ports on the switch with QuickLoop to support private FC_AL. In this case a tape drive attached to the switch is configured as a private FC_AL device, and Server 1 and Server 2 only support private loop. Server 3 is a fabric FC device.

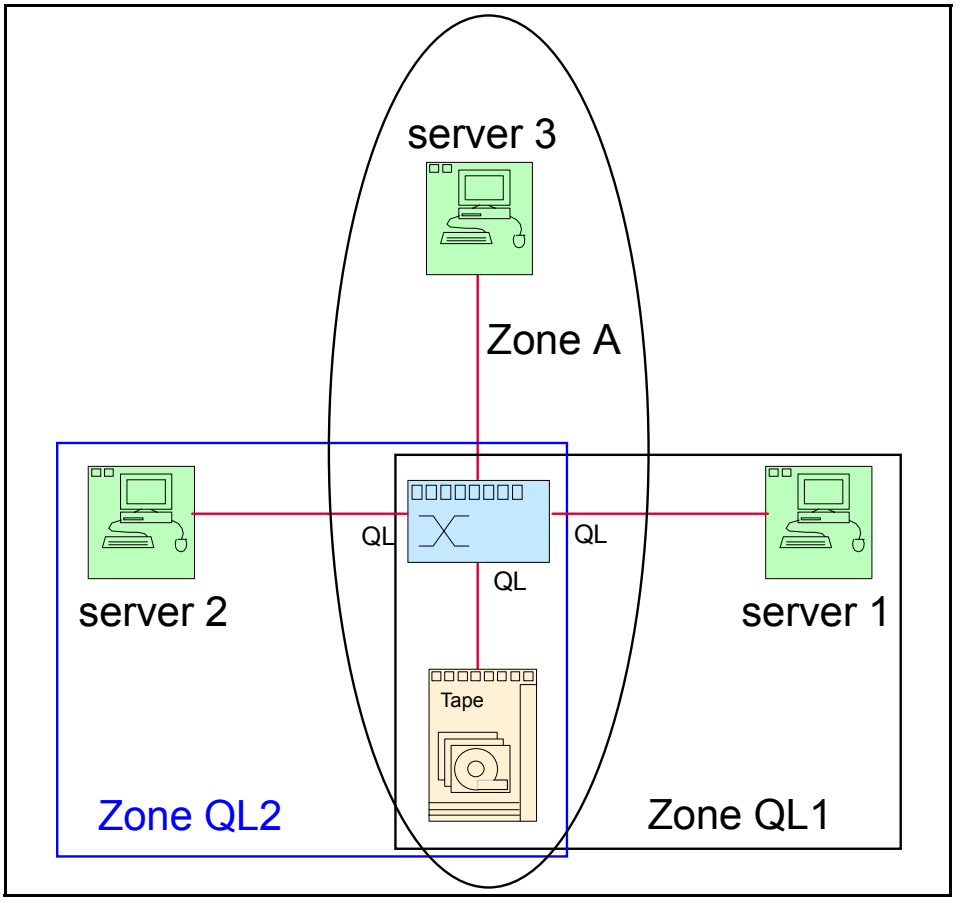


Figure 14-8 QuickLoop

The private hosts (Server 1 and Server 2) may connect to the private tape drive. The public host (Server 3) may communicate with the tape drive as well due to the translative function that QuickLoop provides.

Components

- IBM 2005-H08
- Servers:
  - Three HP UNIX servers each configured with one private FC_AL HBA
- Storage:
  - IBM 3590 Tape Subsystem with native FC adapter

Checklist

- Install and configure switch
- Attach storage to switch
- Attach servers to switch
- All storage devices, server HBAs, switches, or directors are configured with the latest IBM supported versions of driver/firmware levels

Performance

Each QuickLoop zone supports a 200 MB/s transfer rate shared throughout all devices within the loop. As more devices are added, the individual device performance will decrease.

Scalability

You may have up to 126 devices (127 including the initiator) in a FC_AL loop. a QuickLoop cannot span over more than two switches, and the 2005-H08 has eight ports only.

Availability

It is not a high availability solution as a failure of the fabric will stop any server to storage connectivity. A possible fabric failure could be an invalid zoning change, a corrupted configuration, a switch reboot, or similar outage.

Security

By establishing zones in a QuickLoop, you may isolate certain devices from being disrupted from LIPs in the QuickLoop.

If a server in Zone QL1 fails, LIP will take place and will affect the storage, but not the server in Zone QL2, and vice versa.

“What If” failure scenarios

Here we consider the following scenarios:

- **Server HBA:** If one of the HBA fails, the server will lose its storage access. Once you have replaced the HBA, you will need to redefine the WWN for the HBA.
- **Cable:** If a cable between a server and the switch fails, the server will lose its storage access.
- **Cable:** If a cable between the switch and the tape drive fails, each server will lose access to the drive.
- **Switch port:** If one of the ports fails, you may replace it using a hot-pluggable SFP.

- **Switch power supply:** Another redundant power supply may be added to the switch, and should one fail, the other will take over automatically.
- **Switch:** If the switch fails, each server will lose access to the drive. It is important to back up the zoning information periodically, and as and when changes are made. If the faulty switch is replaced you may restore information from the backup.
- **Storage:** If the tape drive fails, each server will lose access to it.

15

IBM TotalStorage SAN m-type family solutions

In this chapter we illustrate and describe solutions that can be implemented using the McDATA products included in the IBM portfolio.

We will base our solutions on the availability characteristics of each product described in Chapter 10, “IBM TotalStorage SAN m-type family” on page 429. The solutions are categorized as follows:

- Performance solutions
- Availability solutions
- Clustering solutions
- Secure solutions
- Loop solutions

© Copyright IBM Corp. 2004. All rights reserved.

15.1 Performance solutions

When designing a SAN careful thought must be given on how to design the SAN so that performance does not suffer. One concept that needs to be evaluated is oversubscription.

One type of oversubscription would be the number of servers associated with a storage port. It is very difficult to work out the ratio of server ports to storage ports. The solution we show in Figure 15-1 illustrates how a general high performance profile could be applied to a SAN design using a director and a single IBM TotalStorage Enterprise Storage Server.

If we do not have accurate performance data from the servers we need to employ a high level methodology to help come up with a baseline. This methodology should only be used to generate a high level design. Final designs must be based on performance data collected from the servers.

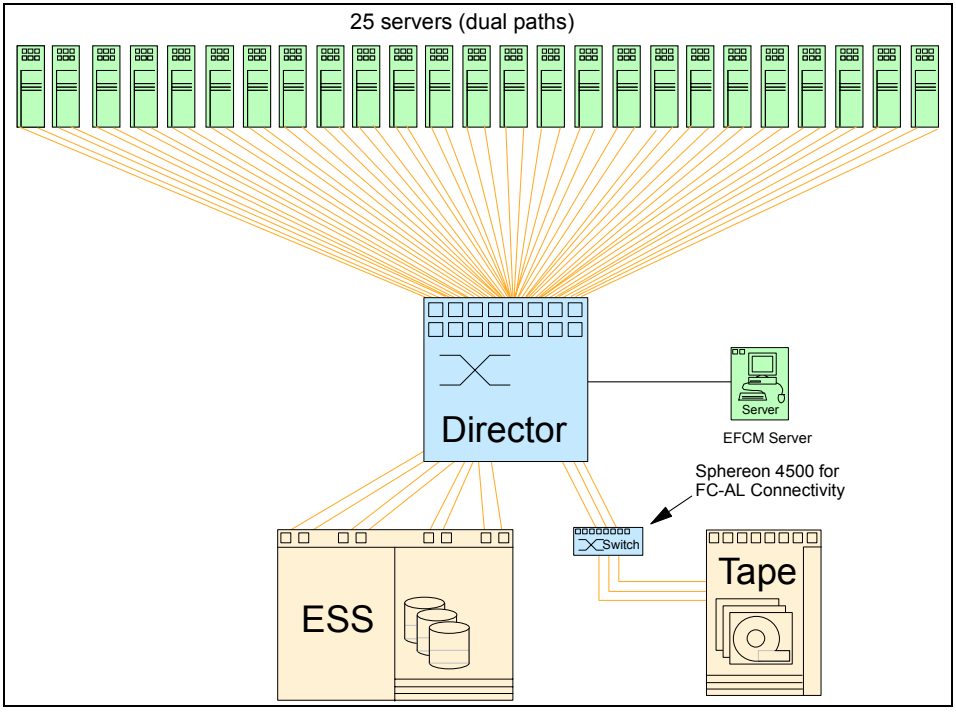


Figure 15-1 High performance design

Fibre Channel will operate at up to 200 MB/s. Of course this depends on many factors such as the type of data access, whether it is read or write intensive, the blocksize of the data, etc. For this example we will assume that the average is

130 MB/s. If we configured eight connections from the director to ESS, we would have a maximum SAN peak bandwidth capability of 1040 MB/s (8 x 130 MB/s).

If we connected 25 dual attach servers to a McDATA Intrepid 6064 Director, and all servers were processing at the same time, we would potentially have a maximum SAN peak bandwidth of 41.6 MB/s per server (1040 MB/s / 25).

This throughput assumes that all 25 servers are able to generate this level of I/O at the same time. This could be categorized as a *high performance profile*.

Based on this theory, for a high performance profile, we have a server connection to ESS port ratio of 6.25 which we round down to 6. So our ratio in this case is 6:1.

Note: The high performance profile is calculated by determining the ratio between the number of server ports (or HBAs) and ESS Fibre Channel ports.

In our example above: 25 servers with dual paths = 50 server ports / 8 ESS ports = ratio of 6.25:1

For *low performance profiles*, such as file and print servers, we will use a rule-of-thumb of 12 server connections to one ESS port. In this case we would use a ratio of 12:1.

Tape device functions such as serverless backup and/or the servers the tape device is connected to must be taken into consideration as well. For best performance, separate HBAs should be used for tape access.

These profile ratios are recommended as a starting point when there are no server performance details available. These rules are very generic and should only be applied at the initial design stage. Prior to any final design a detailed performance profile should be conducted using open systems performance measuring tools such as IOMETER and IBM's Disk Magic.

In our solution we will connect 25 dual attach high performance profile servers to a single ESS.

Components

- SAN fabric:
  - McDATA Intrepid 6064 Director configured with 64 ports
  - One McDATA Sphereon 4500 Fabric Switch for FC-AL connectivity
- Servers:
  - 25 servers each configured with dual FC HBAs.
- Storage:

  - One IBM TotalStorage Enterprise Storage Server Model 800 configured with 8 x Fibre Channel adapters
  - 3584 Automated Tape Library configured with 6 x FC drives
- Software:
  - McDATA EFCM
  - SDD installed on servers
  - IBM StorWatch Enterprise Storage Server Specialist for configuring the ESS
  - 3584 with StorWatch Specialist
  - IOMETER and Disk Magic for performance modelling

Checklist

We checked the following items:

- Spread dual connected ports across UPM cards to minimize the affect of a card failure within the director
- Spread ESS connections across different UPM cards
- Consider the impact of losing a UPM card and balance the server groups to minimize impact
- Leave some ports spare for contingency
- Monitor the performance using Tivoli
- Collect MIB information to determine busy ports
- Conduct a detailed server performance profile
- All storage devices, server HBAs, switches, or directors are configured with the latest IBM supported versions of drivers/firmware levels

Performance

As detailed in our solution description a detailed server performance profile needs to be undertaken to be confident that there is no under- or over-utilization of the SAN's bandwidth. Due to the performance of the director any SAN performance bottlenecks will likely be at the ISLs (if configured), or more likely at the HBAs of the storage device.

Based on this theory, the performance of the SAN will be determined on how much traffic will be moved through the E_Port or HBA. With detailed server profiles it is possible to balance this accordingly.

Scalability

Based on our performance profiling we could expand our solution and connect two directors together using dual E_Ports, as shown in Figure 15-2. Each director now has four connections to the ESS and 3 connections to the tape library. We have now created a higher availability SAN that could support 100 device connections (assuming 50 servers with dual HBAs) attached. This design provides protection against any possible complete failure of a director.

If the number of required ports is expected to grow, then starting with a half populated McDATA Intrepid 6140 Director rather than the McDATA Intrepid 6064 Director could be a more cost-effective decision.

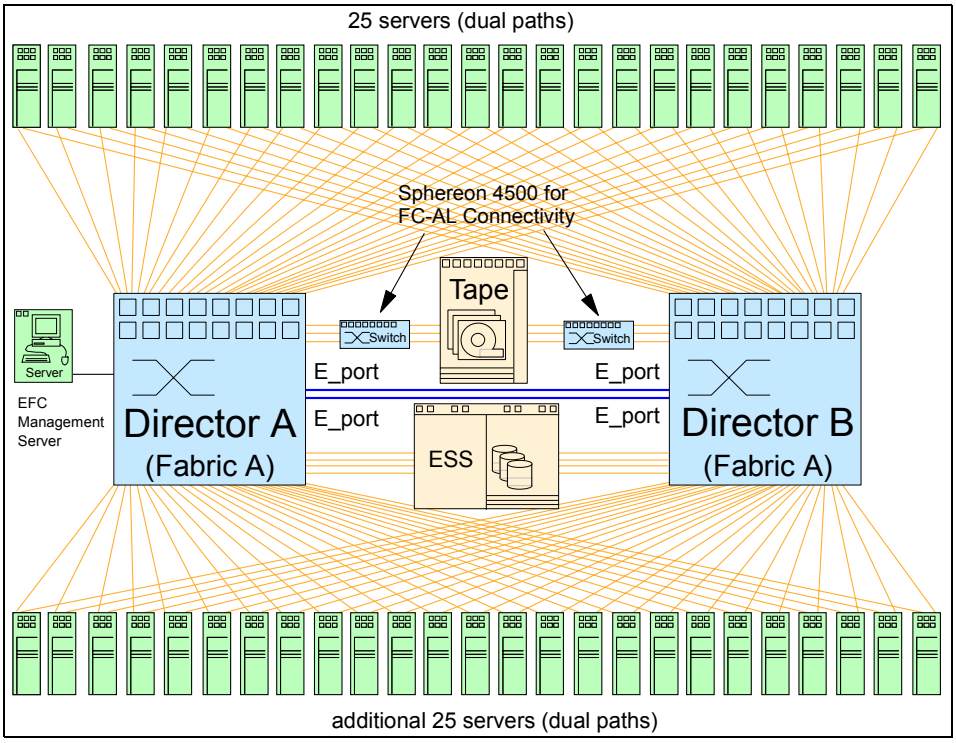


Figure 15-2 Expanded SAN fabric via E_Ports

In the example in Figure 15-2 we have connected two 64 port directors together. We have kept a spare card (4 ports) in each for the re-cabling or re-connecting of devices immediately in the event of a port or UPM card failure.

As you can see we have doubled the number of servers without changing the number of storage ports. This has increased the server port to storage port ratio to 12:1, and reduced the maximum SAN server bandwidth to 20.8 MB/s per server. This design is a much more cost effective solution.

Availability

While this design provides a higher-availability design than for the single director model, a failure in the SAN fabric could result in all hosts losing access to the devices. For example, if an invalid zoning change was made to the fabric or the fabric was corrupted this would effect all devices in the SAN.

Security

We have not considered any security issues with this solution. These will be addressed in some of the following solutions.

"What If" failure scenarios

These are some theoretical assumptions based on Figure 15-1 on page 608:

- **Cable:** If a cable fails between director and ESS, an alternate route will be used, although we would lose 12.5% of the available bandwidth.
- **Switch:** If a UPM card fails, we still have connectivity, as we have dual connections, but we would lose 50% bandwidth to any connected servers, and up to 12.5% bandwidth to the ESS.
- **Switch:** If CTP2 card fails, there would be no effect, there would be a failover to the redundant CTP2 card.
- **Switch:** If the backplane was damaged we would lose connectivity to all servers at that site. The solution in Figure 15-2 provides protection against a complete director failure.
- **Server HBA:** If a server HBA fails, we lose 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.
- **Storage:** If an ESS HBA is unavailable, we have multiple other connections that will automatically be used. We would lose 12.5% of the available SAN bandwidth per HBA.
- **Storage:** If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used. We would lose 25% of the available bandwidth per bay, The ESS has 4 adapter bays each holding up to 4 interface cards. In our example we have configured the ESS with 2 x FC adapter cards per bay.
- **Fabric:** For both Figure 15-1 and Figure 15-2 a failure in the SAN fabric itself will cause a loss of connectivity for all devices.

15.2 Availability solutions

The solution in 15.1, "Performance solutions" on page 608 used a single director. Given the availability characteristics associated with directors this may be enough for many installations, but when the nature of business applications involved makes it necessary to eliminate any possible single point of failure we may want additional protection.

15.2.1 Dual fabric

In the solution shown in Figure 15-3 we have a high availability design. We have two fabrics, two directors, and we have redundant paths to servers and storage. With a high availability design we have redundancy. Although dual directors offer high availability, in a single fabric you would still have a single point of failure i.e. the fabric. A single SAN fabric could be affected by a number of events including the following:

- Incorrect zoning change
- Overlaying a zone configuration
- RSCN storm

By implementing a solution based on dual fabrics we can avoid the impact of a SAN fabric failure. Such a solution is shown in Figure 15-3. In this scenario every device in the SAN has a connection to both fabrics. In the event of a director or fabric failure we would still have connections to all storage devices.

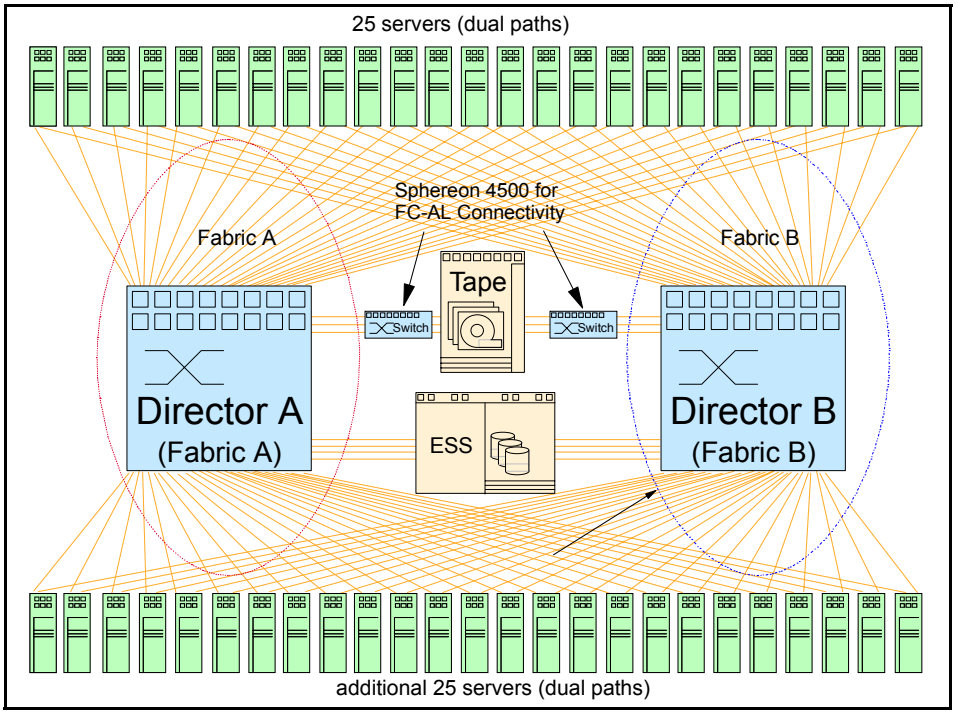


Figure 15-3 Redundant fabrics

Components

To achieve the dual fabric design we require no extra equipment when compared to our expanded fabric shown in Figure 15-2, although in this case we are connecting to LTO generation 1 drives requiring Arbitrated Loop connectivity.

We have included two McDATA Sphereon 4500 Fabric Switches to provide the FC-AL connectivity, in addition to our previous fabric.

Checklist

All the considerations in the previous example should be taken into account here. We should also consider the following additional points:

- Different preferred ID assignment to each director
- Redundancy, and spread across different host bays
- Documented procedures for recovery

Performance

A detailed server performance profile needs to be undertaken to be confident that there is no under- or over-utilization of the SAN's bandwidth.

It is important to monitor the links into the tape library and the ESS as more hosts are connected to the SAN. This is in order to ensure that the links into the devices do not become saturated as more load is placed on the SAN.

Scalability

Based on our performance profiling we could expand our solution further through the addition of extra directors into each separate fabric. In this case we would connect the directors within the fabric together using E_Ports but we would not join the fabrics.

Keeping a spare card (four ports) on each director available in the event of a failure is a good practice if it is practical in your environment. This allows for the re-cabling or re-connecting of devices immediately in the event of a UPM card failure.

Security

The security requirements are basically the same as those in the previous example.

Availability

While this design provides a higher-availability design than for the single SAN fabric model it does not protect us against a complete site failure.

“What If” failure scenarios

The What if scenarios are basically the same as those listed in the previous example, although we now have two directors to cover a complete switch failure:

- **Switch:** If the backplane of one of the directors was damaged and required replacing,requiring a director outage, then SDD would redirect the I/O through to the surviving fabric. This would lead to a possible performance issue as half of the SAN bandwidth is no longer available. This is why it is very useful to understand the performance requirements of your environment - it will allow you to predict the affect on performance should one fabric entirely fail.

15.2.2 Dual sites

In Figure 15-4 we have installed redundant hardware to avoid any possible single point of failure, but our installation is limited to a single site. If we need to be able to continue operations, for example a disaster has shut the site down, we may have to consider a dual site solution.

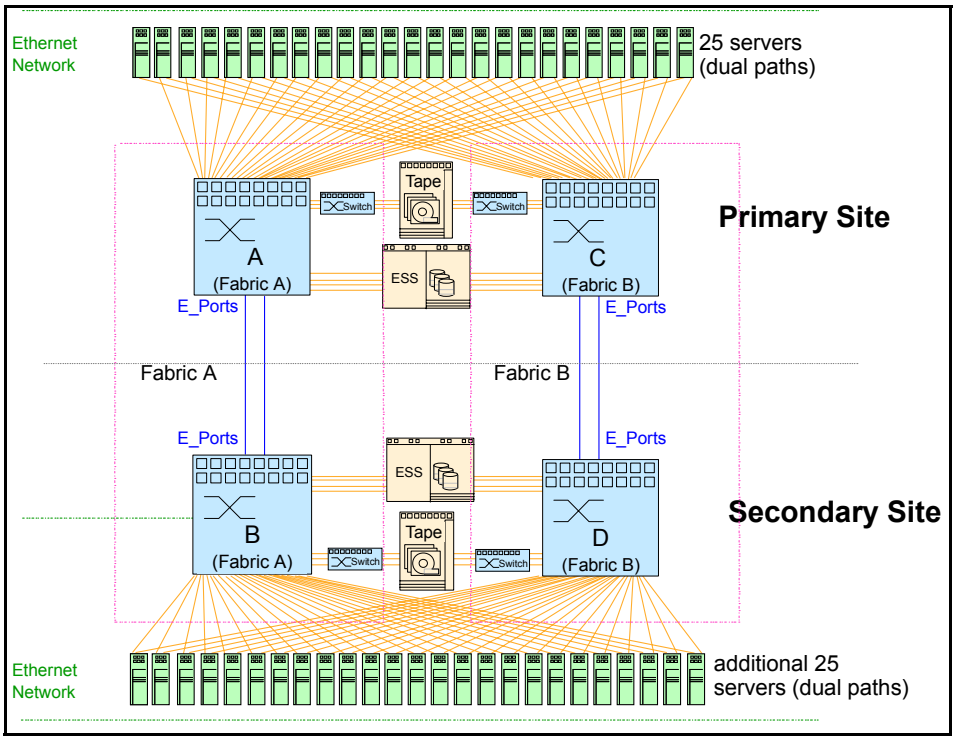


Figure 15-4 Dual sites

Our highly available clusters now have servers and storage subsystems installed in each site.

We need to install longwave SFPs in the directors to be able to move them more than 300 m apart (assuming 2 Gb/s). With ESS PPRC V2 synchronous, shared Fibre Channel connections may be used instead of dedicated ESCON links required with the initial version of PPRC. This raises the distance limit from 3 km to 103 km, or 300 km with Metro Mirror synchronous PPRC available with ESS LIC 2.4.0 and above. We are also limited by the maximum distance allowed by the clustering implementation and the connection used for the heartbeat transmission.

The number of ISLs is based on the bandwidth requirement between sites. We need a minimum of two for availability purposes. With two ISLs a link failure will reduce the available bandwidth to 50%. If we have implemented four directors to all our storage connections, four ISLs will allow us to keep the same bandwidth available for local and remote devices.

For full details on distance solutions, see Chapter 24, “IBM TotalStorage SAN m-type family distance solutions” on page 909.

We have installed an EFC Server at each site with each connected to their local directors on the private LAN. Remote workstations can access both EFC servers through the public LAN.

We zone by WWN so we can replicate zone information to directors on both sites. We will keep director network addresses in the private LAN the same on both sites so the same backup zip disk can be restored on both sites.

With this implementation in the case of a disaster on the Primary Site we can continue working from the Secondary Site if all data is mirrored, otherwise we will be able to resume operation after restoring from backups.

Components

To achieve the dual site design we are duplicating the single site redundant fabric design shown in Figure 15-3, and described in “Components” on page 614.

Checklist

In addition to the requirements of the previous examples we now have to consider:

- Dark fiber availability between sites for Fibre Channel with separate primary and alternate paths of approximately the same length
- Clusters heartbeat connection supported distance
- EFC Servers network addresses

- Different preferred ID assignment to each director
- E_D_TOV, R_A_TOV, and BB_Credit settings equal on both directors
- SAN Management. EFC remote workstation connections or host management software able to manage the SAN from both sites
- Long-wave lasers are installed and ports configured for long distance

Performance

The factors affecting performance introduced in this solution are the number of ISLs. ISL traffic will depend on how the data is distributed between the ESSs.

With four ISLs we have a theoretical 800 MB/s bandwidth between sites. If we assume that our FC link runs at about 130 MB/s we are assured of our 520 MB/s sustained data rate between sites.

We should try to keep heavy workload inside a director but, for example, cluster failover or backup operations may cause large variations in the amount of ISL traffic.

We should implement some measurement methodology, or use statistical data or EFC Product Manager Performance View, to ascertain if we have a bottleneck and need additional ISLs.

Scalability

We can scale this solution by using spare ports on the McDATA Intrepid 6140 Director. We can add new ESS arrays, directors, and servers at each site very easily. Keep in mind that we need to have redundancy for every device added to keep the same availability level.

Security

In addition to the points discussed in previous examples we should now consider the physical security of the patch panels with the fiber optic cables and the connections between sites.

“What If” failure scenarios

- **Site:** Should a complete site fail the servers can failover (manually or automatically) depending on the failover technology being deployed.
- **Site:** It is also important to consider the implications of fail-back procedures. Assuming the Primary Site failed over to the Secondary Site all updates are now occurring at the Secondary Site. At some stage the data on the ESS at the Secondary Site will need to be failed back to the ESS at the Primary Site to bring the data back in-sync. This may place increased load on the SAN infrastructure.

- **Local Storage:** In the event that a server needs to access the storage at the remote site this can put more pressure on the ISL bandwidth so this needs to be considered as part of the SAN planning process
- **Tape:** More ISLs could be required if remote tape vaulting was to be implemented so this also needs to be factored in to the SAN design process.

15.3 Clustering solutions

The diagram in Figure 15-5 shows a clustering solution for several pSeries servers designed to create a high availability clustered environment.

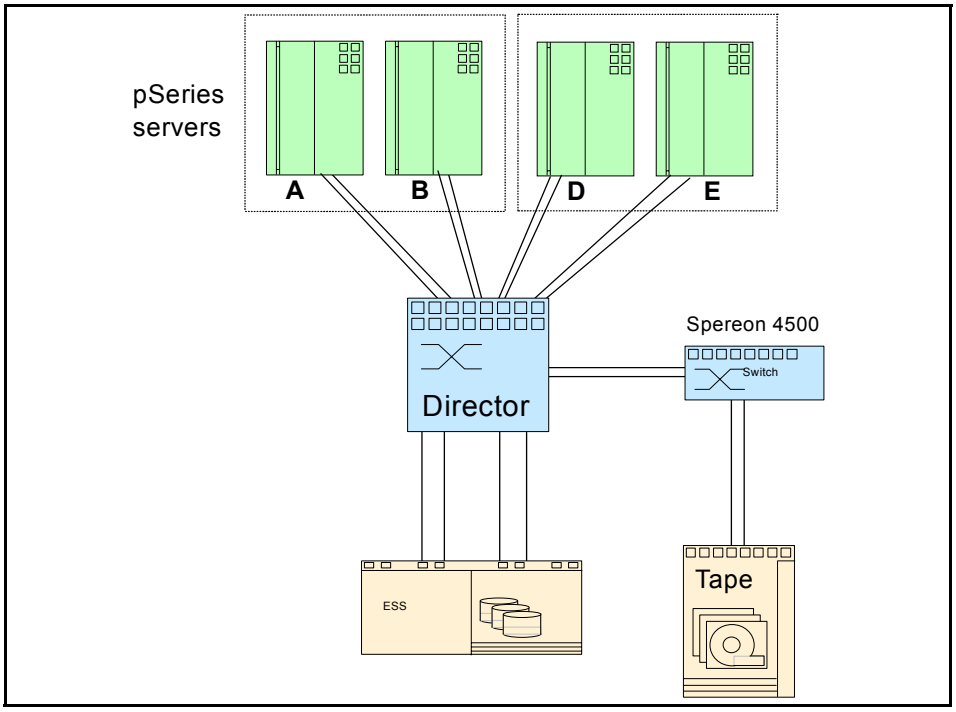


Figure 15-5 Single director clustering solution

In this example we show two separate clusters. Servers A and B are pSeries 660, connected as a high available cluster using HACMP, and we will call them an online cluster. Servers D and E are in a separate cluster. Server E is also the normal backup server.

In order to have redundant connections to storage we have installed two HBAs in each server, and IBM SDD is implemented to provide dual pathing and load balancing.

LUN masking in the ESS is implemented so cluster members are able to see the same LUNs. The backup server must be able to see its own LUNs and the LUNs where the backup data is copied.

We have installed four HBAs in the ESS, one on each host bay, in order to provide a high bandwidth and redundancy.

We have used a McDATA Intrepid 6140 Director allowing us to spread storage connections across different cards. Also, each server connection is attached to two different cards to avoid single a point of failure.

Components

- SAN fabric:
  - One McDATA Intrepid 6064 Director configured with 64 ports
  - One McDATA Sphereon 4500 Fabric Switch for FC-AL connectivity
- Servers:
  - Clustered pSeries servers each configured with dual FC HBAs
- Storage:
  - One IBM TotalStorage Enterprise Storage Server Model 800 configured with 4 Fibre Channel adapters
  - One 3590 Tape Drive
- Software:
  - SDD installed on servers
  - IBM StorWatch Enterprise Storage Server Specialist for configuring the ESS
  - Clustering software
  - IOMETER and Disk Magic for performance modelling

Checklist

We checked the following items:

- AIX operating system, clustering software (HACMP), dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Space required for FlashCopy of backup data
- ESS features and LIC level to support copy services and proposed configuration
- McDATA Director high availability features
- Procedures in place to backup zoning configuration after each change
- Nickname assignments so we can quickly cross reference WWNs to devices

- ESS LUN definitions performed
- Backup software and tape device driver levels to support proposed configuration
- EFC Server and Manager userids and passwords defined
- SANpilot and Telnet CLI default passwords changed

Performance

In this simple implementation, performance will basically depend on the number of HBAs available on each server and the number of storage connections.

The McDATA director supports any to any connectivity so it will not affect performance by itself. Latency in the director is less than 2.0 microseconds.

We have installed four HBAs in the ESS. Since 2 Gb/s Fibre Channel can run at 200 MB/s, we have a potential bandwidth of 400 MB/s. Although 200 MB/s is supported our typical throughput is expected to be about 130 MB/s so a reasonable bandwidth to expect is 520 MB/s.

Knowing the requirements of our servers we can decide whether it is enough, or if we need to add more connections or even more storage devices.

Without having the exact requirements, we may consider that for a high profile server, a server to storage ratio of 6:1 is acceptable. This is only a starting point and we will then implement some measurement system or use statistical data to decide whether we need to add more connections, or if we have more bandwidth than required.

In order to reserve bandwidth for the online and backup clusters, we have performed zoning, restricting development servers to access only two ESS ports. We performed zoning by WWN so it does not depend on the director port the fiber optic cable is connected to.

Scalability

The McDATA director supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation.

Security

The following are some security considerations:

- ESS LUN masking by WWN will allow each server access only to configured LUNs.

- EFC Manager userids, passwords and user rights are defined and default passwords removed so only authorized personnel can perform management functions.
- SANpilot and Telnet CLI default passwords changed.
- Remote access to EFC Manager is configured to limit access to authorized workstations.
- Physical director security is ensured by locked cabinet and restricted access site.
- Zoning has been implemented to restrict access.

"What If" failure scenarios

These are some theoretical assumptions:

- **Clustered server failure:** The paired server will take over. Access to data will not be affected. Performance may be affected since a single server will take the cluster workload.
- **Primary backup server failure:** Backup can still be performed from the alternate server, both access same data and share backup devices.
- **Host HBA failure:** SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. EFC Manager access and IBM StorWatch Enterprise Storage Server Specialist access required.
- **ESS host adapter failure:** The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce available bandwidth by 25%. For an average workload and five servers as shown it should not impact performance. When the host adapter is replaced we do not need to update the zoning, as the WWN is maintained by the ESS.
- **Director port failure:** The impact will depend on whether it is a server or storage port. It will be similar to Host HBA or ESS Host adapter failure. The cable can be moved to a spare port. AIX and SDD will have to be reconfigured to pickup the new path information if it was a storage port. Physical access to director and EFC Manager user with maintenance rights required. AIX root access may be required.
- **Fiber optic cable failure:** Impact will depend on whether it is a host attachment or storage attachment fiber optic cable. Only action required is cable replacement. Physical access to director and attached device required.

- **EFC Server failure:** No management access unless we are using in-band management. Operation is not affected until we need to alter zoning information, for example. Call home support will also be unavailable.
- **Director completely down, ESS completely down, or site down failure:** These will cause an interruption in normal operation.
- **Physical damage to ESS causing data loss:** We will need to restore data from backup copies.

15.4 Secure solutions

Any adverse effect to a SAN will typically have an impact on multiple servers within the SAN fabric. To minimize this impact it is important to ensure every possible security measure is incorporated into the SAN design. In Figure 15-6, we show an implementation where we have concentrated on illustrating the security features available in the McDATA portfolio to create a secure solution.

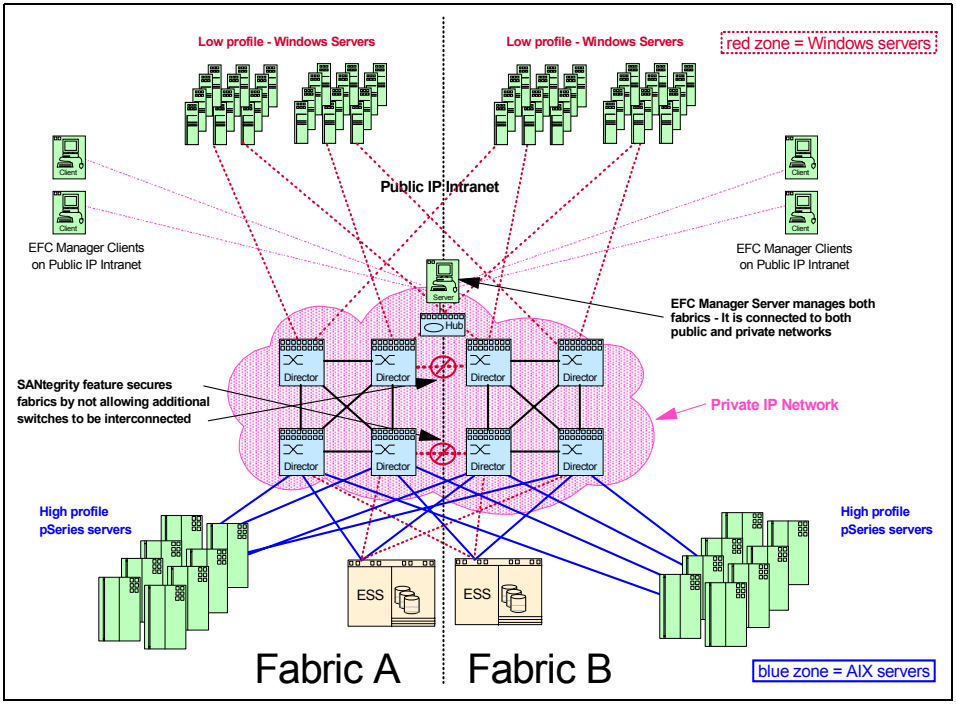


Figure 15-6 Secure solution

Components

- SAN fabric:

  - Eight McDATA Intrepid 6140 Director configured with 140 ports
  - Two fabrics (Fabric A and Fabric B)
- Servers:
  - Windows and pSeries servers each configured with dual FC HBAs
- Storage:
  - Two IBM TotalStorage Enterprise Storage Server Model 800 configured with 4 x Fibre Channel adapters
- Software:
  - SDD installed on servers
  - IBM StorWatch Enterprise Storage Server Specialist for configuring the ESS
  - EFC Manager Server 7.1 and Clients
  - SANtegrity feature enabled.

Checklist

- EFC Manager userids and passwords defined and the default password changed
- SANpilot and Telnet CLI default passwords changed
- EFC Manager Clients installed on management workstations or administrators PCs
- Switches and EFC Manager are on a private network to prevent unauthorized access
- SAN devices are in a locked secured environment with restricted access
- SANtegrity feature ordered to lock down fabric: This can prevent any unauthorized access and prevent fabric segmentation from accidental switch interconnection
- Ensure that Default zoning has been disabled

Security

The following are some other security considerations:

- ESS LUN masking by WWN will allow each server access only to configured LUNs.
- Persistent binding has been employed at each server so that the server only looks or knows about the LUNs it is supposed to see.
- EFC Manager user IDs, passwords and rights defined and default passwords removed so only authorized personnel can perform management functions.
- SANpilot and Telnet CLI default passwords changed.

- Remote access to EFC Manager configured to limit access to authorized workstations.
- Physical director security is ensured by locked cabinet, restricted access site.
- Zoning has been implemented to restrict access. We have two options for our zoning. Each one offers high security, but different levels of flexibility. We can have one type of zoning, or have all three implemented. We can have the zone type based on each server, or we can also have the zone type based on each server operating system, as illustrated in Figure 15-6. Careful consideration should be given as to what type of zoning is used, based on the level of flexibility and security that is required.
  - **Port zoning:** Use this type of zoning if you want to only allow a specific port to communicate with another specific port on a switch. In the event of a bad port, we would lose access until the port was replaced.
  - **WWN zoning:** Uses the WWN of each device and specifies which devices are allowed to communicate with each other. In the event of a port failure, we could just swap to a different port on the switch.
  - **Port binding:** This is a feature that can be employed at the switch/director. The WWN is bound to a specific port. If there was a port failure, devices would not be able to communicate if switched to a different switch port.
- **SANtegrity:** An optional feature with EFC Manager that allows for a fabric lock down. The fabric can be locked so as not to allow any type of unauthorized connection, such as switch interconnection, that might segment the fabric.

Performance

In this example we used a tiered approach. All high profile servers are connected to the same directors as the storage to provide high locality. These high performance servers will not need to traverse an ISL. The traffic needs to be monitored to ensure sufficient bandwidth. Lower profile servers were connected to the directors where they will traverse an ISL. The ISL traffic will need to be carefully monitored to identify and prevent congestion. The Open Trunking feature can be enabled to provide trunking of available ISLs if firmware and EFCM supports Open Trunking. All servers are dual connected with load balancing, and 2 Gb/s HBAs are used in all servers.

Scalability

All servers are dual connected to multiple fabrics. SAN devices such as directors, switches, servers, and storage can all be added non-disruptively.

“What if” security scenarios

- **Host HBA failure:** SDD will move all load to remaining paths. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the ESS host definition will have to be updated with the new WWN.
- **Zoning changes:** These may be needed depending on the type of zoning implemented.
- **Access:** EFC Manager and IBM StorWatch Enterprise Storage Server Specialist access are required.
- **ESS host adapter failure:** The available paths to storage will be reduced impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce the available bandwidth by 25%. When the host adapter is replaced we do not need to update the zoning as the WWN is maintained by the ESS.
- **Director port failure:** The impact will depend on whether it is a server or storage port that is connected to it. It will be similar to a Host HBA or ESS Host adapter failure. The fiber optic cable can be moved to a spare port. AIX and SDD will have to be reconfigured to pickup the new path information if it was a storage port. Physical access to the director and EFC Manager access is required. AIX root access may be required.
- **Fiber optic cable failure:** Impact will depend on whether it is a host attachment or storage attachment cable. The only action required is cable replacement. Physical access to the director and attached device is required.
- **EFC Server failure:** No management access unless we are using in-band management. Operation is not affected until we need to alter zoning information. Call home support will also be unavailable.
- **Director failure:** Reduced bandwidth but there will no loss of access. Each device is connected to multiple directors and fabrics.
- **Incorrect zoning:** All devices in the zone can be impacted.
- **Someone connects two switches together with fiber optic cable:** SANtegrity will prevent a zone merge and ISL segmentation.
- **Director completely down:** No loss of connectivity, since all devices have redundant paths.
- **ESS completely down, or site down:** These will cause an interruption in normal operation.
- **Physical damage to ESS causing data loss:** We will need to restore data from backup copies

- **Fabric failure:** All SAN devices are dual attached with a connection to each fabric. No loss of connectivity will occur, however, there will be a 50% bandwidth reduction.

15.5 Loop solutions

In the following example we discuss a loop solution for sharing FC attached tape drives among several systems to exploit the enhanced bandwidth and alternate path capabilities of SAN attached devices.

The IBM 3590 E11/E1A and H11/H1A tape drives provide the option of FC attachment. When the FC attachment feature is installed, each drive has two independent FC interfaces or ports.

The diagram in Figure 15-7 is our heterogeneous server example, but now we are also sharing tape devices.

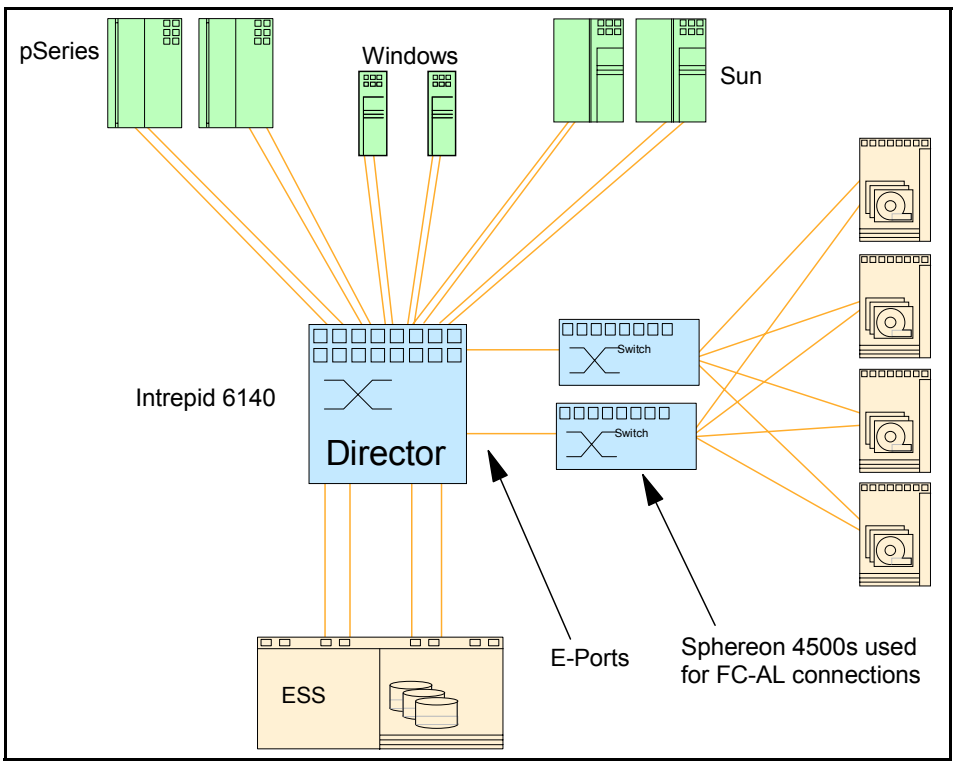


Figure 15-7 Tape attachment using Sphereon 2026-224 switches

Since the 3590 FC ports run in arbitrated loop (FC-AL), they cannot be directly attached to the director, so we have attached them to McDATA Sphereon 4500 Fabric Switch switches which then connect via an E_Port to the director.

Each 3590 drive port is attached to a different Sphereon 2026-224 switch for redundancy.

The Sphereon 2026-224s are also full fabric switches and can have any fabric device (FC-SW) connected to them as well.

In Figure 15-8 we show an example of tape zoning. In our case the pSeries server can access all four tape drives, the SUN servers two and the NT server only one.

We should be careful not to introduce single points of failure when zoning. In our example, if the only tape drive in the NTTAPE zone fails, the NT servers have no tape drive available. We can have an alternate zone defined, and in case of failure, we can activate that zone.

In order to avoid human errors that can affect operation of other servers, zone changes should only be performed by designated personnel. Procedures must be in place to make sure that personnel are aware of the devices available to each server according to the zones currently active.

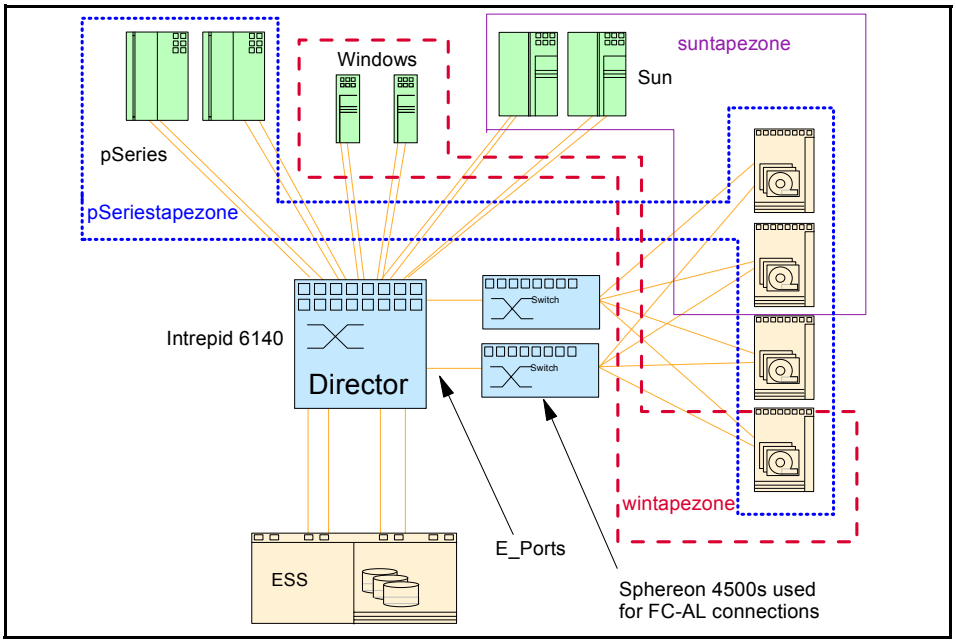


Figure 15-8 Tape zoning

Components

- SAN fabric:
  - McDATA Intrepid 6140 Directors configured with 140 ports
  - Two McDATA Sphereon 4500 Fabric Switches for FC-AL connectivity
- Servers:
  - Windows, SUN and pSeries servers each configured with dual FC HBAs
- Storage:
  - One IBM TotalStorage Enterprise Storage Server Model 800 configured with 4 x Fibre Channel adapters
  - Four 3590 Tape Drives
- Software:
  - SDD installed on servers
  - IBM StorWatch Enterprise Storage Server Specialist for configuring the ESS
  - EFC Manager Server

Checklist

In addition to the items already considered we must now consider:

- Host HBA supported for 3590 attachment
- Host tape device driver levels
- Host operating system levels compatible with 3590 FC requirements
- Sphereon 2026-224s unique switch IDs
- Sphereon 2026-224s priority values higher than director (not principal)
- Sphereon 2026-224s E_D_TOV, R_A_TOV, and BB_Credit compatible with director
- Zone configuration allowing tape access to required servers
- Host software tape sharing capabilities
- Switches LAN connection to EFC Server and remote workstations

Performance

The 3590 E11/E1A and H11/H1A tape drives have a 14 MB/s device data rate; with around 3:1 compression ratio we can have sustained data rates to the host of about 40 MB/s. All drives connected to a single Sphereon 2026-224 switch will share the 200 MB/s bandwidth of the single ISL connection.

Since the Sphereon 2026-224 uses a switched architecture (FC-SW) the full bandwidth is available for each loop. Additional ISLs can be added as needed. Each port can function as an FL_Port or E_Port. Each drive can sustain 40 MB/s. The maximum number of devices connected will depend on traffic and if additional ISLs are added. With this example the ISL bandwidth is sufficient to handle each drive running at 40 MB/s.

It is generally recommended to utilize separate HBAs for tape access than for disk access, and to zone the host to disk and host to tape paths separately.

Scalability

Additional tape drives can be added to Sphereon 2026-224s. ISLs can also be added if ISL congestion occurs, and Open Trunking added to optimize ISL usage.

Security

- Zoning can be used to restrict access to devices to specific servers when required.
- Proper tape management procedures to avoid servers contending for the same tape device.
- EFC manager and Sphereon 2026-224 Web access users and passwords configured and defaults removed.

"What If" failure scenarios

These are some theoretical assumptions:

- **ISL or switch failure:** Access available through the other switch but performance may be impacted depending on the number of drives attached. Traditionally tape failover is a manual operation. Multiple path devices are configured as several logical devices, one per path. Only one of these logical devices is made active. If there is a failure the application aborts and it can then be restarted using a different logical device. Latest levels of a tape device driver provide alternate pathing support and tape failover for Fibre Channel connections. With this support enabled, if an error occurs the device driver will automatically initiate error recovery and the operation will continue using the next logical path.
- **Director failure:** Since there is only one director the whole SAN would be unavailable.
- **Director port failure:** Depends on which port fails and if the ports fail that the switches connect to they will become unavailable but the tape drive should be accessible because of the redundant connections.
- **Device link or device port failure:** Alternate path remains operational. Recovery may be manual or automatic depending on operating system and driver level.
- **Switch port failures:** SFPs are hot swappable. FL_Ports and E_Ports can be moved to a spare port.
- **Tape drive failure in a single tape zone:** An alternate zone should be made active to obtain access to a working device.

Switch capable tape drives

The IBM TotalStorage Enterprise Tape Drive 3592 and IBM TotalStorage Ultrium 2 Tape Drives (LTO 2) both support direct switched fabric attachment over Fibre Channel, and as such could be connected directly to McDATA switches and directors. However, due to the higher costs of director ports, attachment via McDATA Sphereon 4500 Fabric Switch may still be preferable. The drive throughput (see Table 15-1) should be used to calculate the optimal port utilization.

Table 15-1 FC-SW capable tape drives

Tape drive	Native data rate	Compression	Effective data rate
3592	40 MB/s	3:1	120 MB/s
LTO 2	35 MB/s	2:1	70 MB/s

16

Chapter 16.

IBM TotalStorage SAN n-type family solutions

In this chapter we illustrate and describe solutions that can be implemented using the CNT director products included in the IBM portfolio.

We will base our solutions on the availability characteristics of each product described in Chapter 9, “IBM TotalStorage SAN n-type family and replication appliance” on page 345. The solutions are categorized as follows:

- Performance solutions
- Availability solutions
- Clustering solutions
- Secure solutions
- Loop solutions

© Copyright IBM Corp. 2004. All rights reserved.

16.1 Performance solutions

When designing a SAN careful thought must be given on how to design the SAN so that performance does not suffer. One concept that needs to be evaluated is oversubscription.

One type of oversubscription would be the number of servers associated with a storage port. It is very difficult to work out the ratio of server ports to storage ports. The solution we show in Figure 16-1 illustrates how a general high performance profile could be applied to a SAN design using a director and a single IBM TotalStorage Enterprise Storage Server.

If we do not have accurate performance data from the servers we need to employ a high level methodology to help come up with a baseline. This methodology should only be used to generate a high level design. Final designs must be based on performance data collected from the servers.

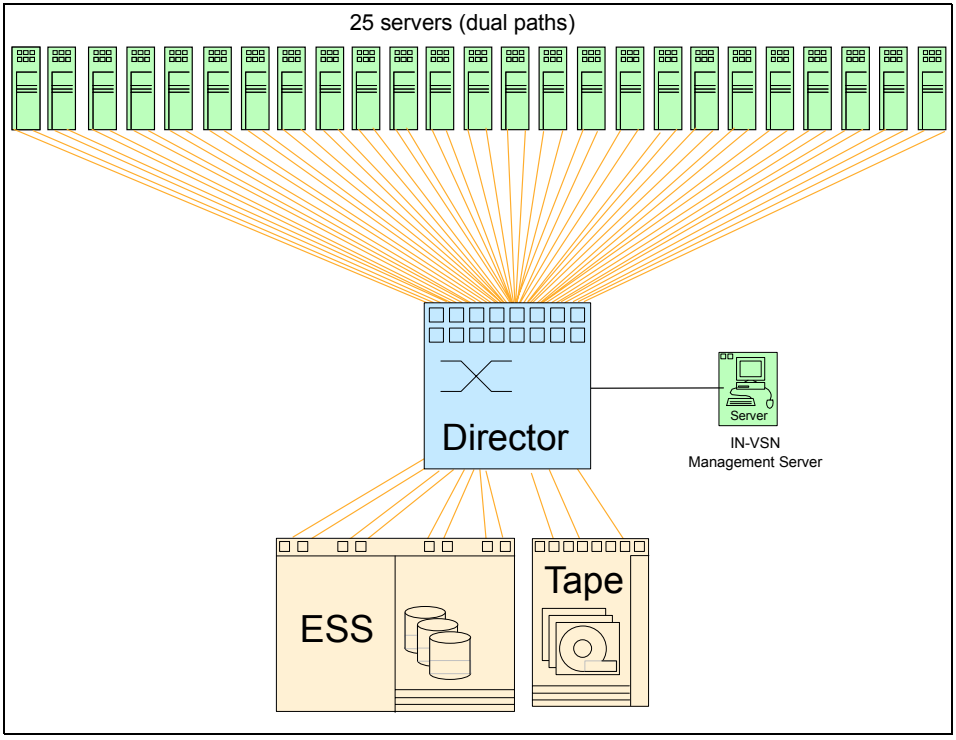


Figure 16-1 High performance design

Fibre Channel will operate at up to 200 MB/s. Of course this depends on many factors such as the type of data access, whether it is read or write intensive, the

blocksize of the data, etc. For this example we will assume that the average is 130 MB/s. If we configured eight connections from the director to ESS, we would have a maximum SAN peak bandwidth capability of 1040 MB/s (8 x 130 MB/s).

If we connected 25 dual attach servers to a CNT UltraNet Multi-service Director Model N16, and all servers were processing at the same time, we would potentially have a maximum SAN peak bandwidth of 41.6 MB/s per server (1040 MB/s / 25).

This throughput assumes that all 25 servers are able to generate this level of I/O at the same time. This could be categorized as a *high performance profile*.

Based on this theory, for a high performance profile, we have a server connection to ESS port ratio of 6.25 which we round down to 6. So our ratio in this case is 6:1.

Note: The high performance profile is calculated by determining the ratio between the number of server ports (or HBAs) and ESS Fibre Channel ports.

In our example above: 25 servers with dual paths = 50 server ports / 8 ESS ports = ratio of 6.25:1

For *low performance profiles*, such as file and print servers, we will use a rule-of-thumb of 12 server connections to one ESS port. In this case we would use a ratio of 12:1.

Tape device functions such as serverless backup and/or the servers the tape device is connected to must be taken into consideration as well. For best performance, separate HBAs should be used for tape access.

These profile ratios are recommended as a starting point when there are no server performance details available. These rules are very generic and should only be applied at the initial design stage. Prior to any final design a detailed performance profile should be conducted using open systems performance measuring tools such as IOMETER and IBM's Disk Magic.

In our solution we will connect 25 dual attach high performance profile servers to a single ESS.

Components

- SAN fabric:
  - CNT UltraNet Multi-service Director Model N16 configured with 64 ports
  - inVSN management server
- Servers:

  - 25 servers each configured with dual FC HBAs.
- Storage:
  - One IBM TotalStorage Enterprise Storage Server Model 800 configured with 8 x Fibre Channel adapters
  - 3584 Automated Tape Library configured with 6 x FC drives
- Software:
  - CNT inVSN
  - SDD installed on servers
  - ESS Specialist for configuring the ESS
  - 3584 with StorWatch Specialist
  - IOMETER and Disk Magic for performance modelling.

Checklist

We checked the following items:

- Consider port groups for zoning requirements
- Spread dual connected ports across TFIO modules to minimize the affect of a TFIO failure within the director
- Spread ESS connections across TFIO modules
- Consider the impact of losing an TFIO module, balance the server groups to minimize impact
- Leave some ports spare for contingency
- Monitor the performance using the inVSN software
- Collect MIB information to determine busy ports
- Conduct a detailed server performance profile
- All storage devices, server HBAs, switches or directors are configured with the latest IBM supported versions of drivers/firmware levels

Performance

As detailed in our solution description, a detailed server performance profile needs to be undertaken to be confident that there is no under- or over-utilization of the SAN's bandwidth. Due to the performance of the director, any SAN performance bottlenecks will likely be at the ISLs (if configured), or the HBAs of storage device.

Based on this theory, the performance of the SAN will be determined on how much traffic will be moved through the E_Port or HBA. With detailed server profiles, it is possible to balance this accordingly.

Scalability

Based on our profiling, we can expand our solution by simply adding more TFIO modules.

If the number of required ports is expected to grow, then starting with a half populated CNT UltraNet Multi-service Director Model N16 rather than fully populating the module slots would be a more cost-effective decision.

As you can see we have doubled the number of servers without changing the number of storage ports. This has increased the server port to storage port ratio to 12:1, and reduced the maximum SAN server bandwidth to 20.8 MB/s per server. This becomes much more cost effective.

Availability

In this SAN design we have used a director, which provides High availability features, although a failure in the SAN fabric could result in all hosts losing access to the devices. For example, if an invalid zoning change was made to the fabric or the fabric was corrupted, this would affect all devices in the SAN.

Security

We have not considered any security issues with this solution. These will be addressed in some of the following solutions.

"What If" failure scenarios

These are some theoretical assumptions based on Figure 16-1. They assume that the director is configured with the High Availability Option:

- **Cable:** If a cable fails between director and ESS an alternate route will be used, although we would lose 12.5% of the available bandwidth.
- **Switch:** Regarding the director, if a TFIO module fails, we still have connectivity, as we have dual connections, but we would lose 50% bandwidth to any connected servers, and up to 12.5% bandwidth to the ESS.
- **Switch:** If a TSW module fails, there would be no effect. First there is spare switching logic on the TSW module, then if that also fails, the other TSW modules share the I/O load in an active-active configuration with load balancing.
- **Switch:** If a TCM module fails, there would be no effect, as the spare TCM module would be automatically invoked.
- **Switch:** If the backplane was damaged, we would lose connectivity to all servers at the site. The solution in Figure 16-2 provides protection against a complete director failure.
- **Server HBA:** If a server HBA fails, we lose up to 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.

- **Storage:** If an ESS HBA is unavailable, we have multiple other connections that will automatically be used. We would lose 12.5% of the available SAN bandwidth per HBA.
- **Storage:** If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used. We would lose 25% of the available bandwidth per bay. The ESS has 4 adapter bays each holding up to 4 interface cards. In our example we have configured the ESS with 2 x FC adapter cards per bay.
- **Fabric:** A failure in the SAN fabric itself will cause a loss of connectivity for all devices.

16.2 Availability solutions

Continuing on from the solution design in the previous section, we now look at two solutions aimed at providing the highest possible availability. These may not be applicable to all environments, but they illustrate the issues associated with building a highly available SAN infrastructure.

16.2.1 Dual fabric

One of the issues we have mentioned previously is that a failure in the SAN fabric can cause the entire SAN to become unstable. A single SAN fabric could be affected by a number of events including the following:

- Incorrect zoning change
- Overlaying a zone configuration
- Broadcast storm

By implementing a solution based on dual fabrics, we can avoid the impact of a SAN fabric failure. Such a solution is shown in Figure 16-2. In this scenario, every device in the SAN has a connection to both fabrics.

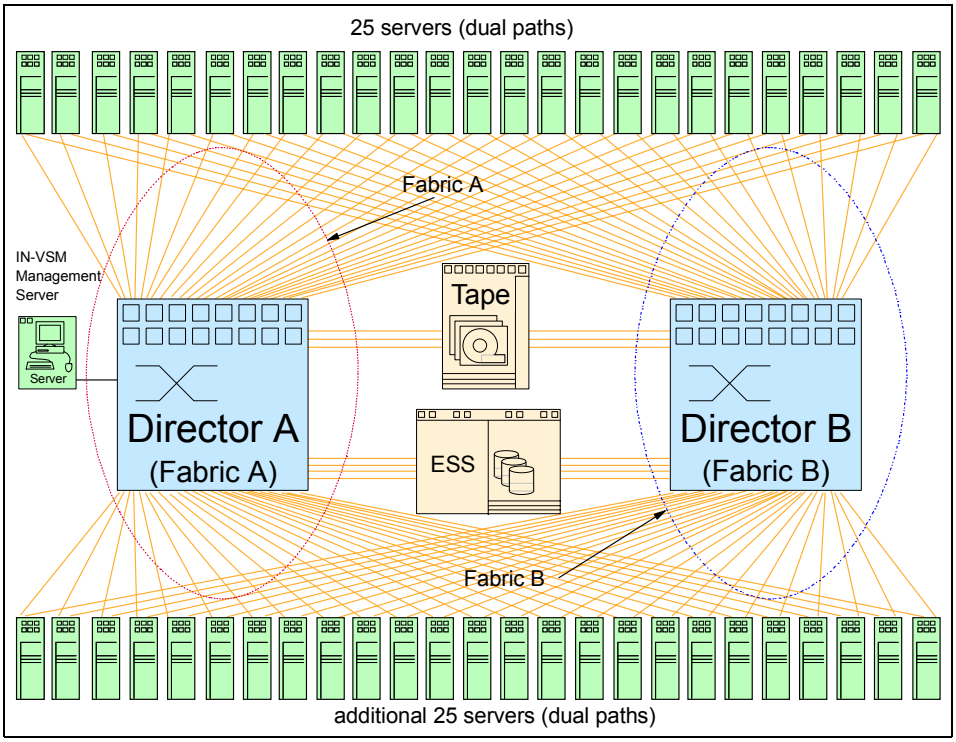


Figure 16-2 Dual fabric design

Components

- SAN fabric:
  - 2 x CNT UltraNet Multi-service Director Model N16 configured with 64 ports each
  - inVSN management server
- Servers:
  - 50 servers each configured with dual FC HBAs
- Storage:
  - 1 x ESS-800 configured with 8 x FC adapters
- Tape:
  - 2 x 3584 Automated Tape Library configured with 6 x FC drives for each site
- Software:
  - SDD installed on servers
  - ESS Specialist for configuring the ESS
  - 3584 with StorWatch Specialist
  - IOMETER and Disk Magic for performance modelling

Checklist

We checked the following items:

- Adjacent ports give optimum performance
- Spread dual connected ports across blades
- Spread ESS connections across TFIO module cards to minimize any impact, should a blade fail
- Depending on the whether the tape drive supports FC-SW protocol, we may need an IBM SAN switch (as TFIO blades do not support FC-AL)
- The 3590 is currently an FC_AL device, as are the native Fibre Channel Ultrium LTO Gen 1 drives, so they would need to connect via an IBM SAN switch
- The native Fibre Channel Ultrium LTO Gen 2 drive supports FC-SW and can be plugged directly into the TFIO module
- Consider the impact of losing an XFIO2 blade and balance the server groups to minimize impact
- Leave some ports spare for contingency
- Monitor the performance using the inVSN software
- Collect MIB information to determine busy ports
- Conduct a detailed server performance profile
- All storage devices, server HBAs, switches, or directors are configured with the latest supported versions of drivers/firmware levels

Performance

As detailed in our previous solution in 16.1, “Performance solutions” on page 632, a detailed server performance profile needs to be undertaken to be confident that there is no under- or over-utilization of the SAN’s bandwidth.

It is important to monitor the links into the tape library and the ESS, as more hosts are connected to the SAN in order to ensure that the links into the devices do not become saturated as more load is placed on the SAN.

Scalability

Based on our performance profiling, we could expand our solution further through the addition of extra directors into each separate fabric. In this case, we would connect the directors within the fabric together using E_Ports, but we would not join the fabrics (as this would form a single fabric).

Keeping a spare ports available in the event of a failure is a good practice if it is practical in your environment. This allows for re-cabling or re-connecting devices immediately in the event of a TFIO module or SFP failure.

Availability

While this design provides a higher-availability design than for the single SAN fabric model, it does not protect us against a site failure.

Security

We have not considered any security issues with this solution. These will be addressed in some of the following solutions.

"What If" failure scenarios

These are some theoretical assumptions based on Figure 16-2:

- If all high performance profile servers are on the same TFIO module, the director is a non-blocking device, so there should be no performance impact, although it would be sensible to spread the load.
- If a cable fails between director and ESS, an alternate route will be used, although we would lose 12.5% of the available bandwidth.
- Regarding the director, if a TFIO module fails, we still have connectivity, as we have dual connections, but we would lose 50% bandwidth to any connected servers, and 12.5% bandwidth to the ESS.
- If a TSW module fails, there would be no effect. First there is spare switching logic on the TSW module, then if that also fails, the other TSW modules share the I/O load in an active-active configuration with load balancing.
- If an TCM module fails, there would be no effect, as the spare TCM module would be automatically invoked.
- If the backplane of one of the directors was damaged and required replacing, then SDD would redirect the I/O through to the surviving fabric. This would lead to a possible performance issue, as half of the SAN bandwidth is no longer available. This is why it is very useful to understand the performance requirements of your environment; it will allow you to predict the effect on performance should a fabric fail.
- If a server HBA fails, we lose up to 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.
- If an ESS HBA is unavailable, we have multiple other connections that will automatically be used. We would lose 12.5% of the available SAN bandwidth per HBA.
- If an ESS bay is unavailable, multiple connections in other bays will be used automatically. We would lose 25% of the available bandwidth per bay.

- Although both fabrics can be managed from the one workstation, it will increase the number of zones and definitions that need to be managed by the administrator.

16.2.2 Dual fabrics over distance

The solution shown in Figure 16-3 is designed to provide protection against a site failure.

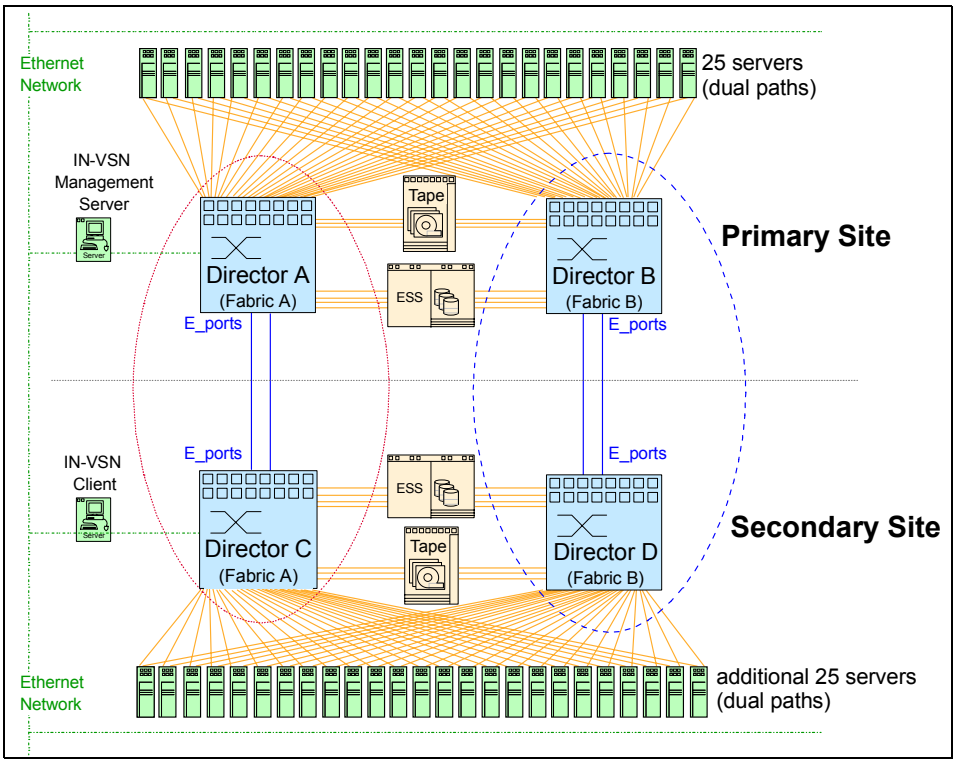


Figure 16-3 Dual fabric design over distance

In this scenario, every device in the Primary Site has a connection to both SAN fabrics. Each of the directors is then connected via an ISL or E_Port to another director located remotely at the Secondary Site. The ESSs are connected using PPRCv2, so all updates are mirrored between sites.

With this scenario, a failure at the primary site can cause our clustered servers to fail over to the secondary site. A SAN fabric failure will cause a server to fail over to the surviving SAN fabric; while performance could be affected, the servers would still have access to their data.

With this solution, the servers could be individual servers with a warm standby server at the remote location made possible via a manual failover process — or they could be clustered systems with an automated failover to a hot machine.

Components

- SAN fabric:
  - 4 x CNT UltraNet Multi-service Director Model N16 configured with 64 ports
  - inVSN management server
  - inVSN client for the remote site
- Servers:
  - 50 servers each configured with dual FC HBAs
- Storage:
  - 2 x ESS-800 configured with 8 x FC adapters
  - PPRCv2 license for remote copy between ESS's
- Tape:
  - 2 x 3584 Automated Tape Library configured with 6 x FC drives for each site
- Software:
  - SDD installed on servers
  - ESS Specialist for configuring the ESS
  - 3584 with Storwatch Specialist
  - IOMETER and Disk Magic for performance modelling
- Cross site links:
  - It is assumed that private fibre links are available to go the required distance

Checklist

We checked the following items:

- Spread dual connected ports across blades
- Spread ESS connections across TFIO modules to minimize any impact, should a blade fail
- Depending on the whether the tape drive supports FC-SW protocol, we may need an IBM SAN switch (as TFIO blades do not support FC_AL)
- The native Fibre Channel Ultrium 2 LTO drive supports FC-SW and can be plugged directly into the TFIO blade
- Consider the impact of losing an TFIO blade, balance the server groups to minimize impact
- Leave some ports spare for contingency — ideally, a spare blade

- Monitor the performance using the inVSN software
- Collect MIB information to determine busy ports
- Conduct a detailed server performance profile
- Ensure that diverse routes are used for the ISLs (E_Ports)
- Ensure that ISLs (E_Ports) are located on different TFIO modules
- All storage devices, server HBAs, switches or directors are configured with the latest supported versions of drivers/firmware levels

Performance

As detailed in our solution description a detailed server performance profile needs to be undertaken to be confident that there is no under- or over-utilization of the SAN's bandwidth.

It is important to monitor the links into the tape library and the ESS as more hosts are connected to the SAN. This ensure that the links into the devices do not become saturated as more load is placed on the SAN.

The performance of the ISLs will also need to be monitored over time to ensure that they are not overloaded. As the SAN increases it is possible that the bandwidth between the directors is no longer sufficient to meet the performance requirements. This can be extended by adding another ISL to the directors to increase the total bandwidth available.

This would be more of an issue if, for example, we were implementing a remote tape vaulting solution in which case a significant amount of data would flow across the ISLs. Our advice is to monitor the ISL traffic over time and have an action plan ready to implement if the ISL links start to become saturated.

Scalability

Based on our performance profiling, we could expand our solution further through the addition of extra directors into each separate fabric. In this case we would connect the directors within the fabric together using E_Ports, but we would not join the fabrics.

Keeping spare ports available in the event of a failure is a good practice, if it is practical in your environment. This allows for the re-cabling or re-connecting of devices immediately in the event of a TFIO module or SFP failure.

Availability

This design provides the highest possible availability, but at the cost of having redundant directors and servers at the secondary site.

Security

We have not considered any security issues with this solution. These will be addressed in some of the following solutions.

"What If" failure scenarios

These are some theoretical assumptions based on Figure 16-3.

- If all high performance profile servers are on the same TFIO module, the director is a non-blocking device so there should be no performance impact, although it would be sensible to spread the load for availability reasons.
- If a cable fails between director and ESS, an alternate route will be used, although we would lose 12.5% of the available bandwidth.
- Regarding the director, if a TFIO module fails, we still have connectivity, as we have dual connections, but we would lose 50% bandwidth to any connected servers, and 12.5% bandwidth to the ESS.
- If a TSW module fails, there would be no effect. First there is spare switching logic on the TSW module, then if that also fails, the other TSW modules share the I/O load in an active-active configuration with load balancing.
- If a TCM module fails, there would be no effect, as the spare TCM module would take over the work.
- If the backplane of one of the directors was damaged and required replacing, then SDD would redirect the I/O through to the surviving fabric. This would lead to a possible performance issue, as half of the SAN bandwidth is no longer available. This is why it is very useful to understand the performance requirements of your environment — it will allow you to predict the affect on performance should a fabric fail.
- If a server HBA fails, we lose up to 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.
- If an ESS HBA is unavailable, we have multiple other connections that will automatically be used. We would lose 12.5% of the available SAN bandwidth per HBA.
- If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used. We would lose 25% of the available bandwidth per bay.
- Should a complete site fail the servers can failover (manually or automatically) depending on the failover technology being deployed.
- It is also important to consider the implications of fail-back procedures. Assuming the primary site failed over to the secondary site all updates are now occurring at the secondary site. At some stage, the data on the ESS at the secondary site will need to be failed back to the ESS at the primary site to

bring the data back in-synch. This may place increased load on the SAN infrastructure.

- In the event that a server needs to access the storage at the remote site, this can put more pressure on the ISL bandwidth, so this needs to be considered as part of the SAN planning process.
- More ISLs could be required if remote tape vaulting was to be implemented, so this also needs to be factored into the SAN design process.

16.3 Clustering solutions

In Figure 16-4 we create a highly available SAN with two Microsoft Exchange WIN2K servers that are linked together using Microsoft Exchange installed on MSCS. We will utilize the Broadcast zone function to transfer the TCP/IP address of the server across the fibres.

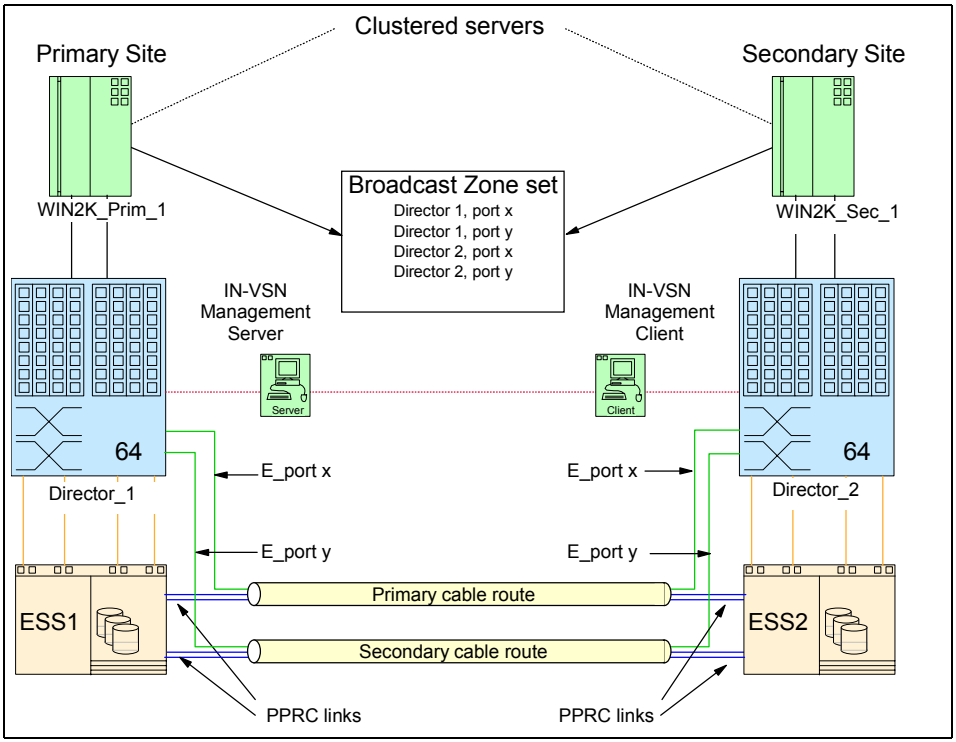


Figure 16-4 High availability solution using CNT Broadcast feature

Both WIN2K_Prim_1 and WIN2K_Sec_1 have dual HBAs. We are running dual pathing software, Subsystem Device Driver (SDD) to enable dynamic failover and workload balancing.

We are using 2 x 64 port CNT SAN256N Directors. The CNT SAN256N Directors are connected using the E_Ports. The E_Ports x and y are both longwave SFPs and can be located on any two separate TFIO modules (all ports on a TFIO blade support up to 128 Buffer Credits so we won't need to change this setting to support the increased distance).

We have defined the inVSN management server and a remote client which is located at the secondary site. In the event of a site disaster the inVSN client will be reconfigured as the inVSN Enterprise Manager.

As we have only a single server and storage device there is no requirement to introduce any hard or name server zones. These would be required if we had a larger number of servers and we wanted to create SAN islands. We have defined a broadcast zone to link Director 1's x and y ports to Director 2's x and y ports.

We have called our Production Exchange server broadcast zone PRDEXCHBZ1.

The CNT SAN256N provides 99.999% availability so we have only configured one for each site. For additional resilience we could have two Directors at each site linked using multiple E_Ports.

To prevent connectivity issues due to an ESS bay being unavailable, we have defined a fibre connection into separate bays in the ESS. The ESSs are linked together using Peer to Peer Remote Copy (PPRC). We would need a minimum of four fc links between ESS1 and ESS2 for pprc. The number of required links would need to be calculated based on the performance profile of the server.

The ESS requires LUN masking to be defined to enable servers to see their respective LUNs. We need to associate all WIN2K_Prim_1 LUNs to it's WWN, we then must perform an ESS modify volume assignment and also define the WWN of the second HBA in WIN2K_Prim_1 to the LUNs. We need to repeat this modify process for both WIN2K_Sec_1 WWNs. We now have 4 WWNs all allocated to the same LUNs on ESS1.

We must repeat this LUN masking process for ESS2, so ESS2 also has 4 WWNs defined.

To support MS Clustering the ESS microcode must be G3 or higher.

If we wanted to use WIN2K_Prim_1 and WIN2K_Sec_1 both as 'live' servers (that is to say, users connected and updating both servers) we could utilize the bi-directional function of PPRC. Changes made on WIN2K_Prim_1 would be

copied to ESS2; changes made on data could be copied from the primary to the secondary and the changed data from the secondary to the primary.

We have diverse primary and secondary fibre cable routes to protect from cable damage. If one route was to become unavailable the secondary route would be used without losing services.

Our solution does not include any extenders or repeater products, and therefore would be more suited to a campus solution. Without extenders, repeaters, or high power laser SFPs, our solution is limited to the maximum supported fibre channel distance of 10km. With extenders and repeaters we could increase these distances to 103 km or even more with special one-off approval.

Components

- SAN Fabric
  - 2 x CNT 2045-N16 configured with 64 ports (each with 2 longwave ports)
  - inVSN management server
- Servers
  - Servers configured with dual FC HBAs.
- Storage
  - 2 x ESS-800 configured with 8 x fibre channel adapters in each machine (four long-wave adapters for pprc links)
  - PPRC license
- Software
  - SDD installed on servers
  - ESS Specialist for configuring the ESS
  - IOMETER and DiskMagic for performance modelling.

Checklist

We checked the following items:

- Version of cluster is supported with ESS
- Version of cluster is supported with SDD
- Applied recommended parameter changes to the HBA's to support clustering
- Distance the cluster service is supported over
- Cluster server failover time
- Primary and secondary routes are diverse
- All cables have been clearly labelled
- A meaningful naming standard has been applied to the broadcast zone.
- Distances between primary and secondary routes are similar
- All storage devices, server HBAs, switches or directors are configured with the latest supported versions of drivers/firmware levels

Performance

In normal operation the ISLs will only be used for replicating Director configuration changes. Due to the FSPF principle the servers at both sites will only connect to their respective site Director. There are no performance considerations for the ISL under normal operations.

If the Cluster Service was activated the TCP/IP address of WIN2K_Prim_1 would be sent across the ISL's and WIN2K_Sec_1 would automatically attach to ESS1 using the two E_Ports. In this scenario WIN2K_Sec_1 would be reduced to 25% of its available SAN bandwidth connected to ESS1 and is accessing ESS1 via the ISLs. Depending on how heavily utilized the paths are to the ESS normally, this could cause congestion on the ISLs. This is why it is very important when designing SAN solutions that you take into account the performance impact when a cluster or site failover occurs.

We have a single director and the Director is a full non-blocking device, there are no performance considerations for the Director.

The important performance considerations for this solution is the number of fibres used between the Director and ESS. In our example we have two server connections to two ESS connections. For more servers, this number will depend on the performance profile of the servers, and the potential performance impact of losing one or more TIO blades.

In Section 16.1, "Performance solutions" on page 632 we look in more detail at performance considerations of having a larger number of servers connected using a Director to a single ESS.

From an availability aspect, if we configure a single connection from the ESS to each TIO module (8 connections). In the event of a single TIO blade failure we have only lost 12.5% of the available bandwidth to the ESS.

Scalability

We can scale this solution through the addition of more servers and devices into the fabric at each site. We would need to monitor the performance of the links between sites to ensure that there in no congestion over the ISLs.

Security

When operating with Primary and Secondary sites we need to insure all related SAN documentation is in a secure location that can be accessed at the recovery site. We also need to ensure we have enough userids of the correct type that are able to make any required changes to the zones.

"What If" failure scenarios

These are some theoretical assumptions:

- Server — If WIN2K_Prim_1 fails the Microsoft cluster Service would be activated and WIN2K_Sec_1 will run WIN2K_Prim_1's workload; this would result in a degradation of the available bandwidth.
- Server HBA — If an HBA fails, SDD will automatically failover workload to the alternate HBA, dependent on application this could result in 30-40% performance degradation. When the HBA has been replaced the WWN information of the new HBA will need to be added to the Broadcast Zone.
- Cable — If a cable between the server and Director fails, SDD will automatically failover workload, and dependent on application this could result in 30-40% performance degradation.
- Cable — If a cable fails between Director and ESS, an alternate route will be used, and we would lose 25% of the available bandwidth.
- Cable — If a cable fails between E_ports there would be no effect in normal operation. In a Cluster failover situation we would lose 50% of the available bandwidth.
- Director — If a TIO blade fails we still have connectivity as we have dual connection, but we would lose 50% bandwidth to any connected servers, and 25% bandwidth to the ESS (assuming each ESS connection is on a separate blade).
- Blade —If a TSW module fails, there would be no effect. First there is spare switching logic on the TSW module, then if that also fails, the other TSW modules share the I/O load in an active-active configuration with load balancing.
- Module — If an TCM module failed there would be no effect as the spare TCM module would be automatically invoked.
- Backplane — If the backplane was damaged we would lose connectivity to all servers at that site. Clustered servers would fail-over to the secondary site.
- ESS — If an ESS HBA is unavailable, we have multiple other connections that will automatically be used. We would lose 25% of the available SAN bandwidth.
- Bay — If an ESS bay is unavailable we have multiple connections in other bays that will automatically be used. We would lose 25% of the available bandwidth per bay.
- ESS — If the ESS is unavailable we would automatically failover to ESS2
- If a problem occurs with the SAN fabric then this could affect all servers in both the Primary and Secondary sites. It represents a single point of failure.

16.4 Secure solutions

Any adverse effect to a SAN will typically have an impact on multiple servers within the SAN fabric. To minimize this impact, it is important to ensure every possible security measure is incorporated into the SAN design.

16.4.1 Logical Domain security solution

For certain organizations, there is a legal or auditors requirement to physically separate data. In Figure 16-5 we utilize the CNT logical domain feature to ensure that only the required server has access to the data without compromising the management benefits of a SAN. Logical domains allows the CNT director to be partitioned into a maximum of four isolated logical switches if desired.The customer can separate different protocols like FICON and FCP instead of utilizing physically separated switches.

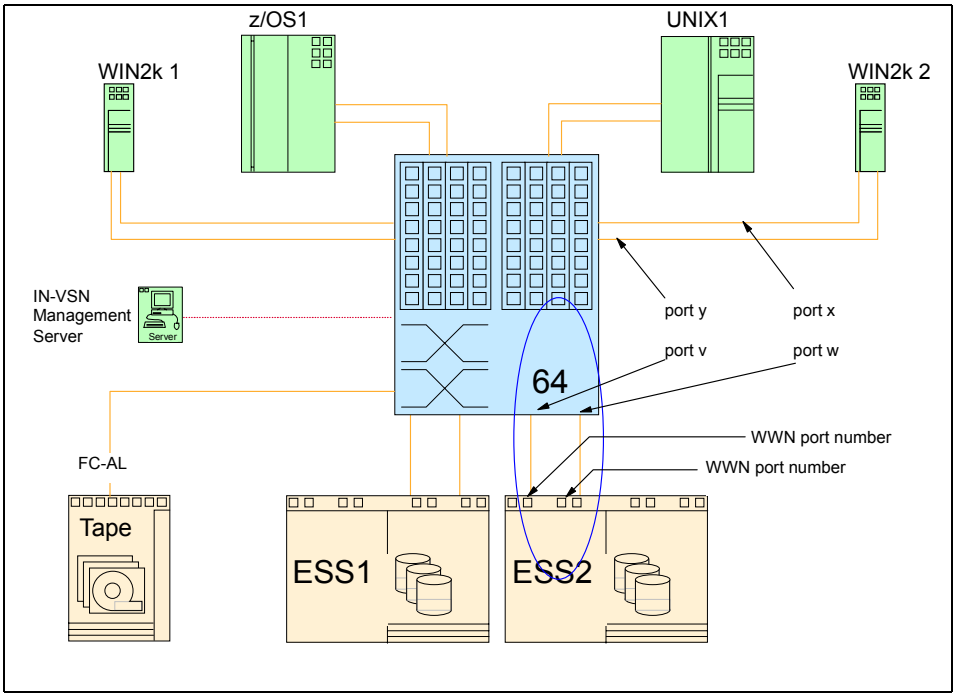


Figure 16-5 Solution using logical domains on the CNT(INRANGE) director

All four servers are sharing the same CNT SAN256N, so we are able to maintain a single point of management, but by using logical partitioning we have guaranteed only WIN2K2 has access to ESS2. We have isolated ESS2 to WIN2K2 only.

We would need to implement two logical domains on the director. The first logical domain (domain 1) will define the ports v, w, x and port y to the WWN port number of ESS2. A logical domain must consist of a minimum of one TIO module, but we want redundant TIO modules. The logical domain must consist of two separate TIO blades for redundancy and contain the four ports v, w, x, and y. It is important the ports x and y are split across the separate TIO blades.

The second logical domain (domain 2) will define the remaining ports on the rest of the TIO blades to the WWN port numbers of ESS1 and the WWN port number of the tape device.

Our use of logical domains has provided us with the required level of security, but we have isolated ESS2 and have prevented WIN2K2 from seeing our tape device.

The CNT director has a locking cabinet, so cables cannot be re-patched without authorization, and the inVSN management software has password protected userids to prevent unauthorized alterations of the hard zone settings.

To prevent connectivity issues due to an ESS bay being unavailable, we have defined a fiber optic cable connection into separate bays in the ESS.

Components

- SAN fabric:
  - CNT UltraNet Multi-service Director configured with 64 ports
  - inVSN management server
- Servers:
  - 2 x WIN2K servers each configured with dual FC HBAs
  - zSeries host with dual FICON adapters
  - pSeries UNIX server with dual FC HBAs
- Storage:
  - Dual ESS-800 configured with 2 x FC adapters in each machine
- Tape:
  - 3584 Automated Tape Library configured with multiple FC drives
- Software:
  - SDD installed on servers
  - ESS Specialist for configuring the ESS
  - 3584 with StorWatch Specialist

Checklist

We checked the following items:

- All TIO blades are defined in domain 1 or domain 2, which isolates the ports.

- The WWPN only of ESS2 have been used in the first hard zone (zone 1)
- The two physical ports WIN2K2 is connected to plus the entire TIO module are isolated in domain 1.
- WWPN hardware enforced or domain/portid zoning is also used in each logical domain to further protect the two separate applications.
- Ports x and y are on separate TFIO modules
- WIN2K1, z/OS1, and UNIX1 do not have LUNs defined on ESS2
- WIN2K2 has no LUNs defined on ESS1
- Dual HBAs have been used, so the LUNs defined on ESS1 and ESS2 have been associated with both HBA WWNs of all servers
- WIN2K2 does not require access to the tape device
- The CNT UltraNet Multi-service Director is locked and in a secure location
- The userids and passwords of the inVSN server software have been changed from the default
- Only SAN administrators have access to the inVSN userid and passwords
- A maintenance window is available when z/OS1, WIN2K1, WIN2K2 and UNIX1 will be unavailable, so the hard zone can be implemented.
- All storage devices, server HBAs, switches, or directors are configured with the latest supported versions of drivers/firmware levels.

Performance

The main performance consideration when introducing logical domains is the under utilization of a shared device. In our example, WIN2K2 has the dedicated bandwidth of ESS2. In reality it is unlikely this would be cost effective and ESS2 would have either direct connections or would be connected to another director or SAN island.

If we lose one of the connections to port x or port y, by either cable or HBA, the servers SAN bandwidth for WIN2K2 will be reduced by 50%, the actual server performance degradation will depend on WIN2K2s applications and how well it was load balancing across the dual HBAs.

Scalability

We could define an additional two logical domains, or separate within the domains further via normal WWPN hardware-enforced zoning.

Security

Security considerations have been covered in the solution and checklist sections.

“What If” failure scenarios

These are some theoretical assumptions:

- **Access is obtained to director cabinet:** If the fiber optic cables belonging to WIN2K2 were removed and replaced with the fibre cables belonging to WIN2K1, WIN2K2 would physically lose sight of the LUNs defined on ESS2. WIN2K1 would not be able to see any LUNs on ESS2. As the ESS uses LUN masking, all LUNs on ESS2 have been defined to the WWN of WIN2K2.
- **Access is obtained to inVSN server:** If the hard zone was amended and the port address of WIN2K2 is replaced with the port address of WIN2K1, we would have the same symptoms as swapping cables. The ESS LUN masking links the LUNs to the WWN of WIN2K2s HBAs only. In addition, the update to the hard zone would have resulted in a re-login for all servers connected to the director.
- **Access is obtained to inVSN and ESS Specialist:** If the logical domains, zoning, and ESS LUN masking information has been switched between WIN2K1 and WIN2K2, WIN2K1 would now see WIN2K2’s LUNs.
- **Port x or y fails:** These ports have been spread across two TFIO modules so our dual path will ensure we have connectivity to ESS2, and we also have spare ports in this domain.
- **Logical Domains were activated during normal operation:** All servers connected to the director would have to re-login, and depending on the operating system this may cause the server to shut down.

16.4.2 Soft zoning security solution

In Figure 16-6 we look at a SAN design that could be used to provide additional security to the SAN. The most flexible way to implement added security is with the use of name server zoning or soft zoning.

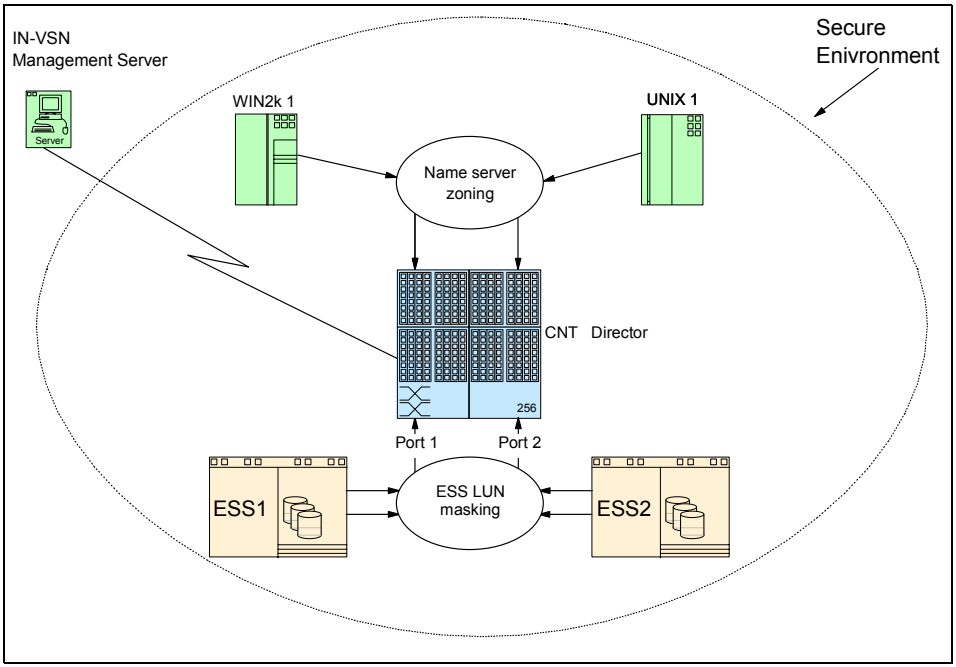


Figure 16-6 Protecting your data from both human error and sabotage

To prevent unauthorized access to our SAN fabric, all SAN components are contained in a secure environment. The management interfaces to the ESS and the director are userid and password protected, and it is not possible to define or amend these userids or password directly.

To prevent unauthorized cable changes, the director has been installed in a locking cabinet. For added protection against cable changes, we have also incorporated name server zoning.

In this solution we have used the WWNs of WIN2K 1 and UNIX 1 and created a name server zone through director 1 to the WWPN of the HBAs in the ESS. This compliments the LUN masking facility provided by the ESS.

Checklist

We checked the following items:

- All fabric components are in a secure location.
- The WWNs of the HBAs Port names of WIN2K1 and UNIX have been defined to a name server zone.
- If dual HBAs have been used, the LUNs defined on ESS1 and ESS2 have been associated with both HBA WWNs of WIN2K 1 and UNIX 1.

- The userids and passwords of the inVSN server software have been changed from the default.
- Only SAN administrators have access to the inVSN userid and passwords.
- All storage devices, server HBAs, switches, or directors are configured with the latest supported versions of drivers/firmware levels.

Performance

This solution is looking at SAN security, and therefore, performance has not been considered in this case.

Scalability

We can currently define up to 256 name server zones within this fabric.

"What If" failure scenarios

These are some theoretical assumptions:

- **Access is obtained to director cabinet:** If the fiber optic cables belonging to WIN2K 1 or UNIX 1 are switched, all the correct LUNs would still be visible at the server. Name server zoning is independent of the director port position.
- **Access is obtained to the inVSN server:** If WIN2K 1's Port WWN is removed from the name server zone, all LUNs will be visible as normal. Once WIN2K 1 ports are not defined to a name server zone they will automatically be defined to the name zone orphans zone, so the ESS will still be visible. The ESS LUN masking links the LUNs to the WWN of WIN2K 1s HBAs only.
- **Access is obtained to ESS Specialist:** If ESS LUN masking information is modified and UNIX 1 HBA WWN is added to WIN2K 1s LUN list, it would be possible for UNIX 1 to see WIN2K 1's LUNs.

16.4.3 Heterogeneous zoning solution

One of the major benefits of a SAN is to enable groups of servers to access shared devices. Typically, these shared devices will comprise of a mixture of equipment from many different vendors. With the ability to intermix FICON, FCP, and loop devices within a single director, we have the maximum flexibility to realize these benefits with an CNT UltraNet Multi-service Director.

The solution in Figure 16-7 is an example of a site where IBM and other vendors (in this case EMC, HDS, and STK) devices exist, and you want to incorporate them into the SAN fabric.

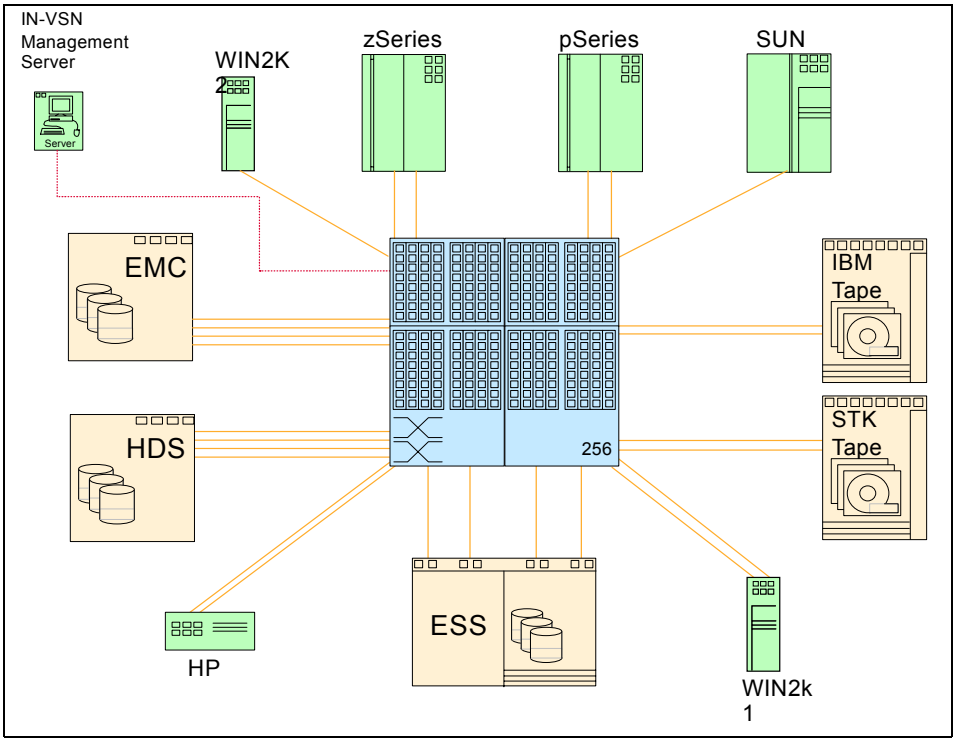


Figure 16-7 A heterogeneous solution connected to a single SAN fabric

To successfully implement this type of solution, there are a number of issues that must be considered in addition to those discussed previously.

In our solution, we have decided that:

- WIN2K 1 will connect to the HDS and STK tape.
- SUN will connect to EMC, ESS, STK tape and IBM tape.
- WIN2K 2 will connect to EMC, HDS, ESS, STK tape and IBM tape.
- zSeries to connect to ESS and IBM tape.
- pSeries to connect to ESS and IBM tape.
- HP to connect to HDS and IBM tape.

Our site naming standards label the disks:

- HDS Dev disk1
- EMC Dev disk 2
- ESS Prod disk1
- STK Unit1 tape 1
- IBM Unit 2 tape 1

We have decided to use name server zoning to achieve this, as follows:

- **Zone1:** N1D3U1
- **Zone 2:** S1D2P1U1U2
- **Zone 3:** N2D1D2P1U1U2
- **Zone 4:** O1P1U2
- **Zone 5:** R1P1U2
- **Zone 6:** H1D1U2

As we will be connecting SUN and WIN2K 2 to a mix of ESS and other storage devices, we cannot run dual pathing software on the server. The dual pathing software supplied by IBM, HDS, and EMC is supported for each vendor's equipment only. This means we cannot run multiple versions of dual pathing software on a server. We can only have a single connection from these environments.

We are using Emulex HBAs in the Windows WIN2K and SUN servers, and the standard shipped HBAs in the pSeries and HP. The zSeries is using FICON cards.

We have to select a common device driver for SUN and WIN2K 2: IBM and EMC support different versions of the device driver. We must raise an RPQ with IBM for support of the Mini Port device driver.

The supported firmware level of the CNT director may vary from vendor to vendor to reach a common ground, and RPQ may be required.

Components

- SAN fabric:
  - CNT UltraNet Multi-service Director Model N16 configured with 256 ports
  - inVSN management server
- Servers:
  - WIN2K servers each configured with dual FC HBAs.
  - WIN2K server with a single FC HBA
  - Sun server with a single FC HBA
  - pSeries with dual FC HBAs
  - zSeries server with dual FICON HBAs
  - HP server with dual HBAs
- Storage:
  - ESS-800 configured with 4 x FC adapters
  - EMC disk subsystem configured with 4 x FC adapters
  - HDS disk subsystem configured with 4 x FC adapters
- Tape:
  - 3584 Automated Tape Library configured with 6 x FC LTO drives

  - STK Tape Library configured with 2 drives
- Software
  - SDD installed on servers to attach only to the ESS
  - ESS Specialist for configuring the ESS
  - 3584 with Storwatch Specialist

Checklist

We checked the following items:

- The version of the operating system is supported on the device we will be connecting to
- Servers that will connect to multiple vendor devices will be single path only
- Any required RPQs have been raised and approved
- Correct device drivers have been downloaded
- Zones have meaningful names applied

Performance

The storage port to server ratio is very low. so performance of the SAN should not be an issue. We have configured four paths to each storage device. which is acceptable for a low number of servers. A server performance profile would need to be conducted to determine an accurate number of required fiber optic cables for each storage device. We have three drives defined in the IBM library and two drives in the STK library.

Security

As we have several name server zones defined, backups of the inVSN should be performed on a regular basis. and at least when the information is changed or changes.

“What If” failure scenarios

These are some theoretical assumptions:

- **If we attach a dual HBA to multi vendor storage devices:** The dual pathing software will probably not work and we will invalidate support.
- **If we run the wrong version of the device driver:** We will invalidate support.
- **If we run a lower level of firmware:** We will invalidate support from some vendors.

A more practical solution might be to have a server(s) access only one vendor’s disk subsystem. This could be implemented via separate zones and would allow

for the use of that vendors multipathing software as there is only one type of disk subsystem in the zone common drivers can be then be used.

17

Cisco solutions

In this chapter we illustrate and describe solutions based on the Cisco MDS 9000 family of switches and directors.

We will base our solutions on the features available for each of the Cisco MDS 9000 Multilayer product family as described in Chapter 11, “Cisco switches and directors” on page 489.

The solutions are categorized as follows:

- Performance solutions
- Availability solutions
- Clustering solutions
- Secure solutions
- Loop solutions

© Copyright IBM Corp. 2004. All rights reserved.

17.1 Performance solutions

When designing a SAN careful thought must be given on how to design the SAN so that performance does not suffer. One concept that needs to be evaluated is oversubscription.

One type of oversubscription would be the number of servers associated with a storage port. It is very difficult to work out the ratio of server ports to storage ports. The solution we show in Figure 17-1 illustrates how a general high performance profile could be applied to a SAN design using a director and a single IBM TotalStorage Enterprise Storage Server.

If we do not have accurate performance data from the servers we need to employ a high level methodology to help come up with a baseline. This methodology should only be used to generate a high level design. Final designs must be based on performance data collected from the servers.

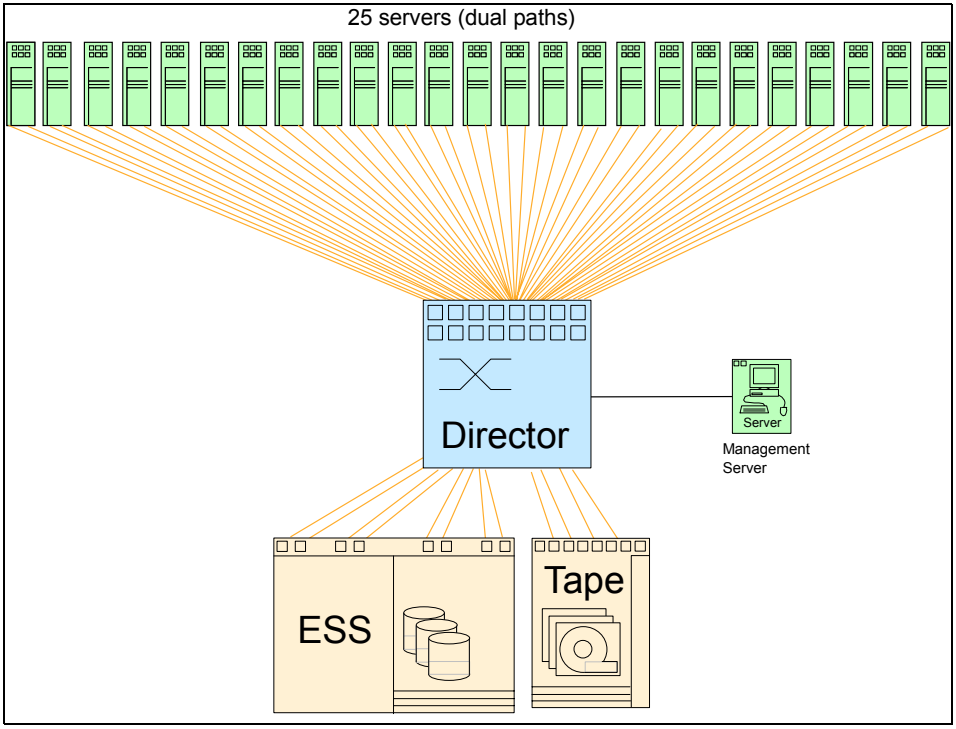


Figure 17-1 High performance design

Fibre Channel will operate at up to 200 MB/s. Of course this depends on many factors such as the type of data access, whether it is read or write intensive, the blocksize of the data, etc. For this example we will assume that the average is

130 MB/s. If we configured eight connections from the director to ESS, we would have a maximum SAN peak bandwidth capability of 1040 MB/s (8 x 130 MB/s).

If we connected 25 dual attach servers to a McDATA Intrepid 6064 Director, and all servers were processing at the same time, we would potentially have a maximum SAN peak bandwidth of 41.6 MB/s per server (1040 MB/s / 25).

This throughput assumes that all 25 servers are able to generate this level of I/O at the same time. This could be categorized as a *high performance profile*.

Based on this theory, for a high performance profile, we have a server connection to ESS port ratio of 6.25 which we round down to 6. So our ratio in this case is 6:1.

Note: The high performance profile is calculated by determining the ratio between the number of server ports (or HBAs) and ESS Fibre Channel ports.

In our example above: 25 servers with dual paths = 50 server ports / 8 ESS ports = ratio of 6.25:1

For *low performance profiles*, such as file and print servers, we will use a rule-of-thumb of 12 server connections to one ESS port. In this case we would use a ratio of 12:1.

Tape device functions such as serverless backup and/or the servers the tape device is connected to must be taken into consideration as well. For best performance, separate HBAs should be used for tape access.

These profile ratios are recommended as a starting point when there are no server performance details available. These rules are very generic and should only be applied at the initial design stage. Prior to any final design a detailed performance profile should be conducted using open systems performance measuring tools such as IOMETER and IBM's Disk Magic.

In our solution we will connect 25 dual attach high performance profile servers to a single ESS.

Components

- SAN fabric:
  - Cisco MDS 9509 Director configured with 64 ports (4 x 16 port switching modules)
- Servers:
  - 25 servers each configured with dual FC HBAs.
- Storage:

  - One IBM TotalStorage Enterprise Storage Server Model 800 configured with 8 x Fibre Channel adapters
  - 3584 Automated Tape Library configured with 6 x FC drives
- Software:
  - Cisco Fabric Manager
  - SDD installed on servers
  - IBM StorWatch Enterprise Storage Server Specialist for configuring the ESS
  - 3584 with StorWatch Specialist
  - IOMETER and Disk Magic for performance modelling.

Checklist

We checked the following items:

- Spread dual connected ports across blades to minimize the effect of a blade failure within the director.
- Spread ESS connections across switching modules in order to protect against a module failure.
- Consider the impact of losing a switching module and balance the server groups to minimize impact.
- Leave some ports spare for contingency.
- Monitor the performance of the environment with Cisco Fabric Manager.
- Collect MIB information to determine busy ports.
- Conduct a detailed server performance profile.
- All storage devices, server HBAs, switches, or directors are configured with the latest IBM supported versions of driver/firmware levels.

Performance

As detailed in our solution description, a detailed server performance profile needs to be undertaken to be confident that there is no under- or over-utilization of the SAN's bandwidth. Due to the performance of the director, any SAN performance bottlenecks will likely be at the ISLs (if configured), or more likely at the HBAs of the storage device.

Based on this theory, the performance of the SAN will be determined on how much traffic will be moved through the E_Port or HBA. With detailed server profiles it is possible to balance this accordingly.

The Cisco MDS 9000 family of switches and directors provide unique flexibility in designing a performance based solution due to the underlying architecture of the equipment. For example, we could configure a director using 4 x 16-port switching modules to provide 64-ports with a full non-blocking implementation. Alternatively, if we know some servers do not require such high performance, as they are not able to generate such high bandwidth requirements, we could

consolidate these servers onto the 32-port switching modules with shared bandwidth capability. This allows us to design SAN solutions on a performance basis while ensuring we get the best cost-per-port and best port density.

Scalability

Based on our performance profiling, we could expand our solution and connect two directors together using dual E_Ports, as shown in Figure 17-2. Each director now has 4 connections to the ESS and 3 connections to the tape library. We have now created a higher availability SAN that could support 100 device connections (assuming 50 servers with dual HBAs) attached. This design provides protection against any possible complete failure of a director.

If the number of required ports is expected to grow, then starting with a half populated Cisco MDS 9509 Multilayer Director rather than the Cisco MDS 9506 Multilayer Director could be a more cost-effective decision.

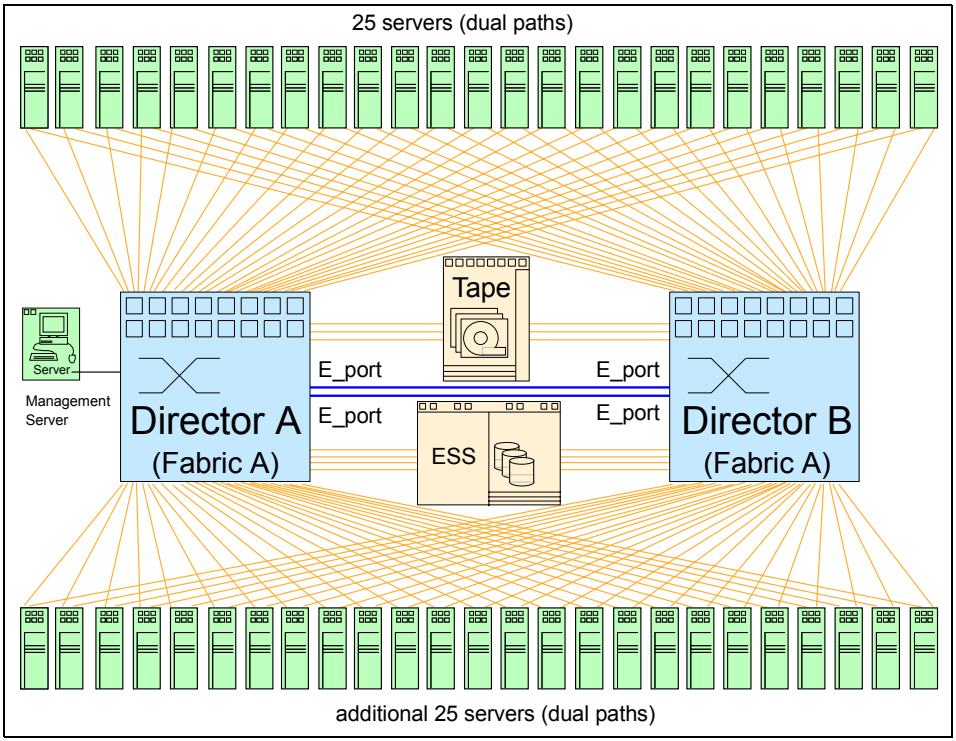


Figure 17-2 Expanding the SAN fabric via E_Ports

As you can see we have doubled the number of servers without changing the number of storage ports. This has increased the server port to storage port ratio

to 12:1, and reduced the maximum SAN server bandwidth to 20.8 MB/s per server. This design is a much more cost effective solution.

Availability

While this design provides a higher-availability design than for the single director model a failure in the SAN fabric could result in all hosts losing access to the devices. For example, if an invalid zoning change was made to the fabric or the fabric was corrupted this would affect all devices in the SAN.

The Cisco MDS 9000 family provides Virtual Storage Area Networks (VSAN) technology which is a way of creating totally independent fabrics across the same hardware platform. We will illustrate solutions based on VSAN technology in the solution in 17.2, “Availability solutions” on page 665.

Security

We have not considered any security issues with this solution. These will be addressed in some of the following solutions.

“What If” failure scenarios

These are some theoretical assumptions based on Figure 17-1:

- **Cable:** If a cable fails between the director and ESS, an alternate route will be used, although we would lose 12.5% of the total available bandwidth.
- **Switch:** Regarding the director, if a switching module fails we still have connectivity, as we have dual connections, but we would lose 50% bandwidth to any connected servers, and up to 12.5% bandwidth to the ESS.
- **Switch:** If a supervisor module fails, there would be no effect as the spare supervisor module will automatically take over.
- **Switch:** If the backplane was damaged, we would lose connectivity to all servers at that site. The solution in Figure 17-2 provides protection against a complete director failure.
- **Server HBA:** If a server HBA fails, we lose up to 50% of the server’s SAN bandwidth, and depending on the application, up to 30-40% of the server’s performance.
- **Storage:** If an ESS HBA is unavailable we have multiple other connections that will automatically be used. We would lose 12.5% of the available SAN bandwidth per HBA.
- **Storage:** If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used. We would lose 25% of the available bandwidth per bay. The ESS has 4 adapter bays each holding up to 4 interface cards. In our example we have configured the ESS with 2 x FC adapter cards per bay.

- **Fabric:** For both Figure 17-1 and Figure 17-2, a failure in the SAN fabric itself will cause a loss of connectivity for all devices. We can use Cisco's Virtual SAN Technology to protect against such a failure and this is discussed in the topics that follow.

17.2 Availability solutions

Continuing on from the solution design in the previous section, we now look at two solutions aimed at providing the highest possible availability. These may not be applicable to all environments, but they illustrate the issues associated with designing a highly available SAN infrastructure.

17.2.1 Dual fabric

One of the issues we have mentioned previously is that a failure in the SAN fabric can cause the entire SAN to become unstable. A single SAN fabric could be affected by a number of events including, but not limited to, the following:

- Incorrect zoning change
- Overlaying of a zone configuration
- RSCN storm

By implementing a solution based on dual fabrics, we can avoid the impact of a SAN fabric failure. Such a solution is described in Figure 17-3. In this scenario, every device in the SAN has a connection to the separate fabrics, and this is how a traditional SAN would be designed to protect against a single fabric failure.

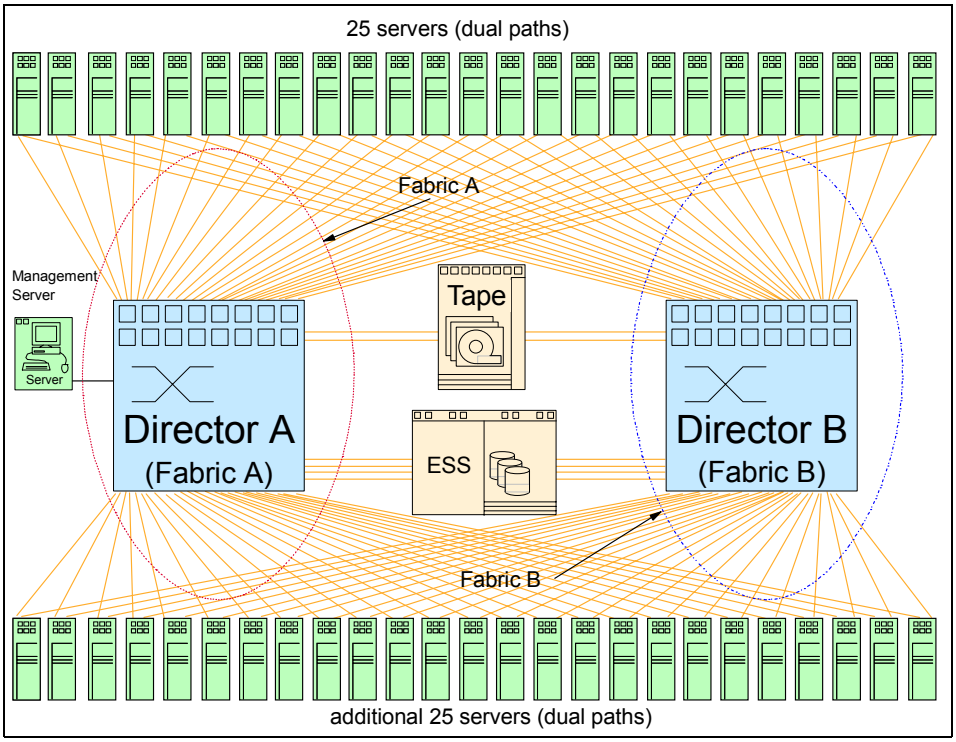


Figure 17-3 Traditional dual fabric design (without VSANs)

By using Cisco's Virtual SAN (VSAN) technology, we eliminate the requirement for the second director, as shown in the previous example. Shown in Figure 17-4 is an illustration of how we can provide dual SAN fabrics providing the same logical infrastructure as in the previous diagram. We have illustrated this with 25 servers, but this could be many more.

The diagram shows how two distinct VSANs are formed using the one physical infrastructure. The VSANs are logically separate from each other. Note that with VSANs it is not possible to have the same port defined in more than one VSAN.

This technology enables us to provide a dual-fabric solution without having to purchase multiple directors or switches to achieve this.

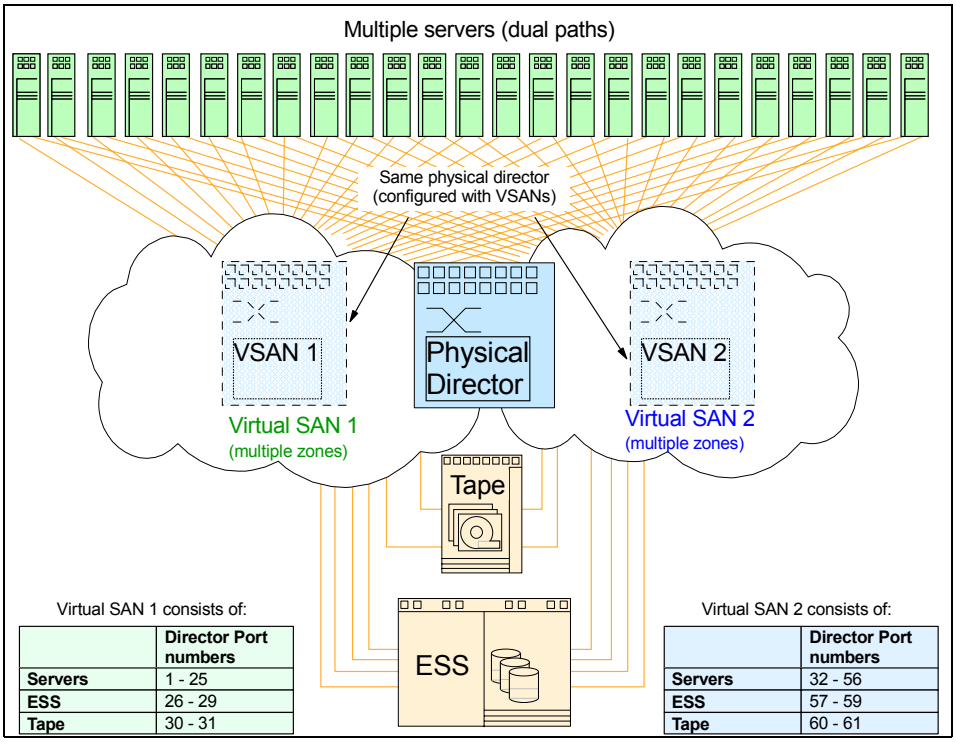


Figure 17-4 Dual fabric design with VSANs

Components

To achieve the dual fabric design by utilizing VSANs we require no extra equipment from our original implementation shown in Figure 17-1 on page 660.

Checklist

In addition to the “Checklist” on page 662 we also need to take the following into consideration:

- Ensure that all devices with multiple paths (servers, ESS, tape) have a connection to both VSANs

Performance

As detailed in our solution description in 17.1, “Performance solutions” on page 660, a detailed server performance profile needs to be undertaken to be confident that there is no under- or over-utilization of the SAN’s bandwidth.

It is important to monitor the links into the tape library and the ESS as more hosts are connected to the SAN, in order to ensure that the links into the devices do not become saturated as more and more load is placed on the SAN.

Scalability

Based on our performance profiling, we could expand our solution further through the addition of extra directors into each VSAN. In this case we would connect the directors within the fabric together using TE_Ports, but we would not join our fabrics (VSANs). TE_Ports enable us to pass multiple VSAN traffic across the ISL(s) while maintaining their segregation.

Availability

While this design provides a higher-availability design than for the single SAN fabric model, it does not protect us against a site failure such as those discussed in “Dual sites” on page 668.

Security

We have not considered any security issues with this solution. These will be addressed in some of the following solutions.

“What If” failure scenarios

The same What if scenarios are valid as those listed in the previous example, although we now have two individual Fabrics to provide redundancy, although the following also needs to be considered:

- If the director was physically damaged or destroyed, then obviously we would lose access to all the devices. This is one advantage of having redundant hardware in place.

17.2.2 Dual sites

So far we have installed fabric redundancy to avoid any possible single point of failure, but our installation is limited to a single site. If we need to be able to continue operations, for example a disaster has shut the site down, we may have to consider a dual site solution.

To implement a dual fabric solution across two sites we would require multiple directors at the Primary and Secondary sites, as shown in Figure 17-5. While this design provides the highest availability, it can be expensive to implement.

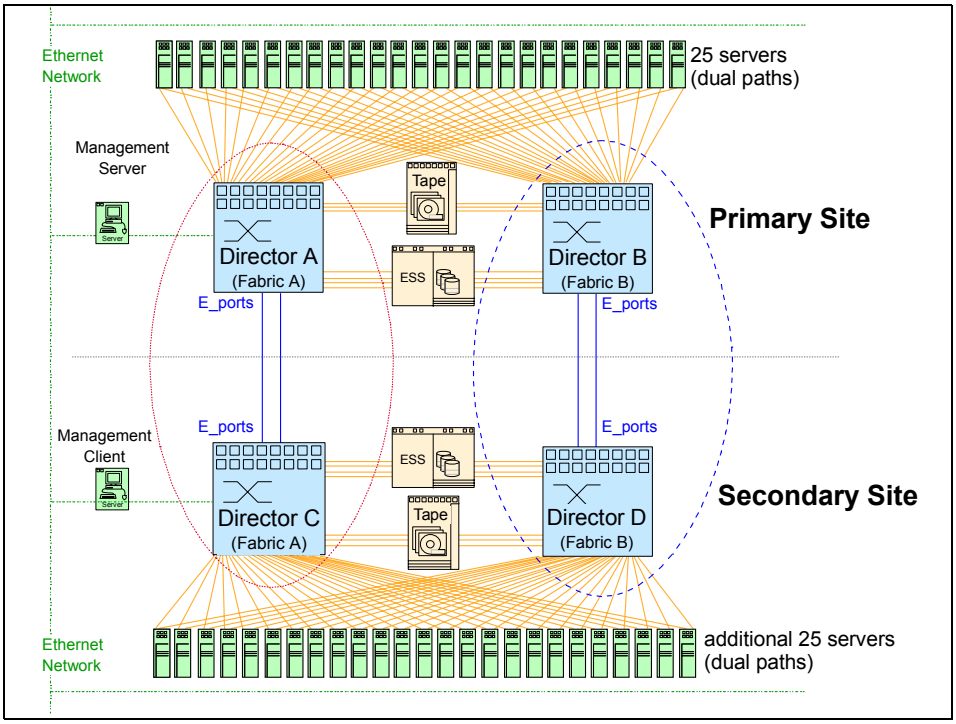


Figure 17-5 Traditional across site dual fabric design

Using the Virtual SAN feature of the Cisco MDS 9000 family, we can put together a dual fabric and dual site design using half the number of directors when compared to the solution in Figure 17-5.

The solution shown in Figure 17-6 is designed to provide the same protection against a site failure, while maintaining dual fabric redundancy, but also minimizing the amount of money required to implement the solution.

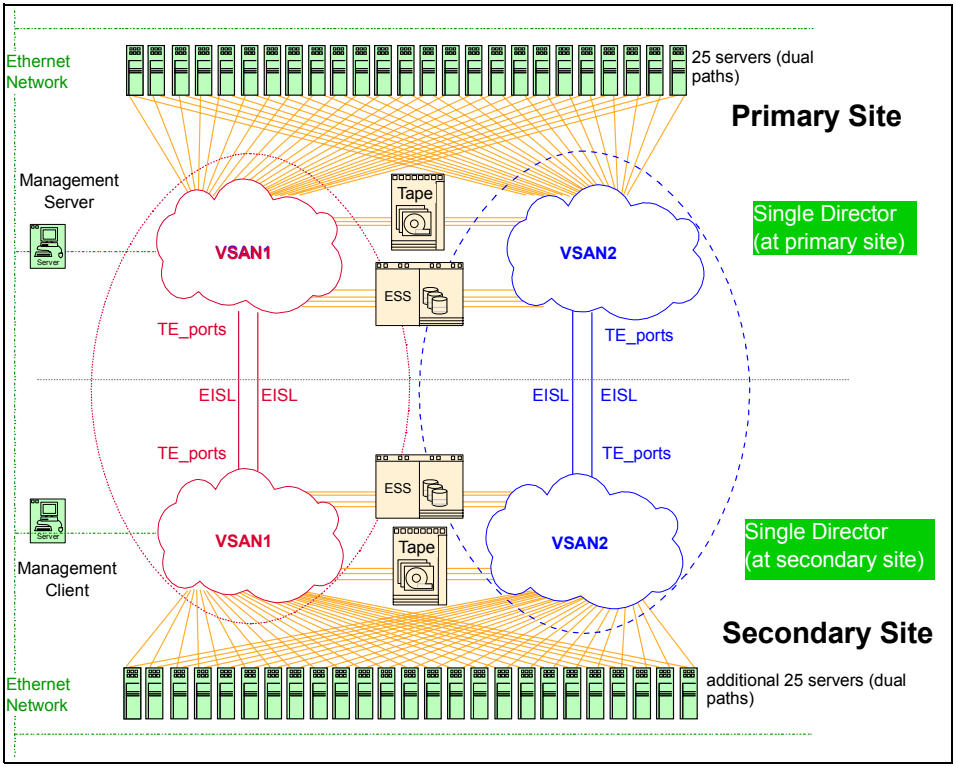


Figure 17-6 Across site Dual fabric design using VSANs

In this scenario, every device in the Primary Site has a connection to both VSAN1 and VSAN2 fabrics. The directors are connected via an EISL or TE_Port to a director located at the Secondary Site. The ESS's are connected using PPRC so all updates are mirrored between sites.

With this scenario, a failure at the Primary Site can cause our clustered servers to fail over to the Secondary Site. A VSAN fabric failure will cause a server to fail over to the surviving VSAN fabric. While performance could be affected, the servers would still have access to their data.

With this solution, the servers could be individual servers with a warm standby server at the remote location made possible via a manual failover process, or they could be clustered systems with an automated failover to a hot machine.

Components

To achieve the dual site design we are duplicating the single site redundant fabric design shown previously in Figure 17-4 on page 667, with the additional requirements:

- Cross site links:
  - Dark fiber less than 10km in length

For full details on distance solutions, see Chapter 26, “Cisco distance solutions” on page 969.

Checklist

In addition to the previous example we also considered the following items:

- Ensure that diverse routes are used for the EISL (TE_Ports)
- Ensure that EISLs (TE_Ports) for each VSAN are located on different switching modules. This ensures that if a switching module should fail, there would still be an active path between directors
- Ensure that EISLs and other devices requiring high performance are only attached using 16-port switching modules in order to utilize all available ports
- Ensure that any servers fitting the low performance profile are connected to the 32-port switching modules

Performance

The performance of the ISLs will need to be monitored over time to ensure that they are not overloaded. As the SAN increases it is possible that the bandwidth between the directors is no longer sufficient to meet the performance requirements. In this case, we would implement a Port-Channeling solution to form an aggregated logic path across the ISLs. This can also be extended by adding another ISL to the directors to increase the total bandwidth available. This would be more of an issue if, for example, we were implementing a remote tape vaulting solution, in which case a significant amount of data would flow across the ISLs. Our advice is to monitor the ISL traffic over time and have an action plan ready to implement if the ISL links start to become saturated.

Scalability

Based on our performance profiling we could expand our solution further through the addition of extra directors into each site and expanding the ports in each VSAN. In this case we would connect the directors at each site using TE_Ports but we would continue to maintain the dual fabric through the VSAN technology.

Keeping spare ports available in the event of a failure is a good practice if it is practical in your environment. This allows for the re-cabling or re-connecting of devices immediately in the event of a switching module or SFP failure.

Availability

This design provides the highest possible availability but at the cost of having redundant directors and servers at the secondary site.

Security

We have not considered any security issues with this solution. These will be addressed in some of the following solutions.

“What If” failure scenarios

All the same “What if” scenarios remain valid as previously covered, although we now have two separate sites to provide complete redundancy, although the following also needs to be considered:

- **Director:** In the event of a catastrophic failure of the director, the servers would fail over to the secondary site. This could be an automatic failure for clustered servers or a manual failover depending on the server requirements. Performance would not be an issue as the Secondary Site is configured to match the Primary Site requirements. This is not always the case with secondary sites so it is suggested that a performance analysis of the primary site is undertaken in order to ensure the secondary site can meet these requirements.
- **Site:** It is also important to consider the implications of fail-back procedures. Assuming the primary site failed over to the secondary site, all updates are now occurring at the secondary site. At some stage the data on the ESS at the secondary site will need to be failed back to the ESS at the primary site to bring the data back in-sync. This may place increased load on the SAN infrastructure.
- **ISL:** In the event that a server needs to access the storage at the remote site this can put more pressure on the EISL bandwidth so this needs to be considered as part of the SAN planning process.
- **ISL:** More EISLs could be required if remote tape vaulting was to be implemented so this also needs to be factored in to the SAN design process.
- **ISL:** If remote tape vaulting was required it would be a useful idea to put these ports and a number of dedicated ISL’s into a new VSAN only for tape activity. This would enable the organization to restrict or control the bandwidth that the tape drives were able to use and would prevent the tape activity from impacting on the production systems. By using VSANs we can ensure that traffic from one application does not affect another application.

17.3 Clustering solutions

These topics focus on clustering solutions.

17.3.1 Two-node clustering

In Figure 17-7 we show a typical basic high availability SAN design for a two-node clustering and fully redundant fabrics by using VSANs.

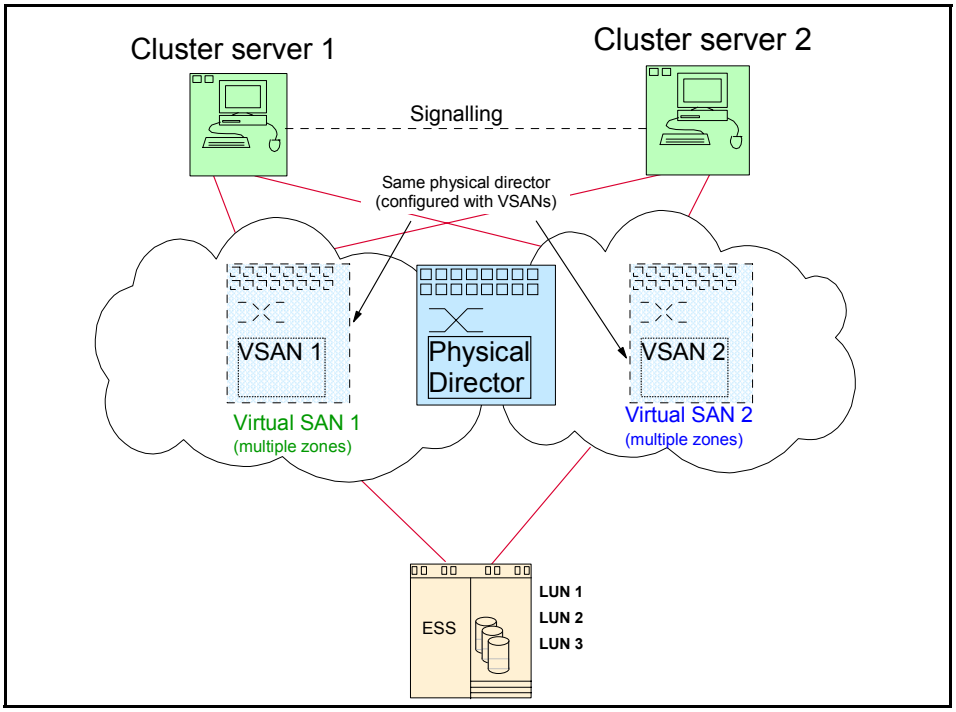


Figure 17-7 Microsoft cluster with redundant fabric

Components

- SAN fabric:
  - One 16-port Cisco MDS 9216 Multilayer Fabric Switch
- Servers:
  - Two IBM xSeries servers each configured with dual FC HBAs
- Storage:
  - One ESS 2105-800 with two FC Adapters,
- Software:
  - IBM Subsystem Device Driver (SDD)

Checklist

- Install and configure switch.
- Install Fibre Channel HBAs.

- Configure ESS.
- Attach ESS to switch.
- Attach servers to switch.
- Validate failover/failback operations.
- All storage devices, server HBAs, switches, or directors are configured with the latest IBM supported versions of driver/firmware levels.

Performance

Typically, for a low performance server, the recommended server to storage oversubscription is 12:1, and for a high performance server, the server to storage oversubscription is 6:1. With a 2:1 ratio (4 / 2), the above configuration is within ratio provided based on four server connections to two storage connections.

To increase the performance of the SAN, multiple connections may be added from the hosts to the switches and from the switches to the storage devices.

Scalability

A dual SAN fabric is a topology where you have two independent SANs that connect the same hosts and storage devices. This design is not one of the most highly scalable as all hosts and storage must be connected to both switches to achieve high availability.

Availability

In addition to the server high availability clustering, SAN high availability is provided with this dual VSAN design. Dual HBAs are installed in each host and the storage device must have at least two ports. Fail over for a failed path or even a failed switch is dependent on host failover software, namely, the IBM Subsystem Device Driver (SDD). The switch does not reroute traffic for a failed link as we have two individual fabrics with this type of design. Each VSAN is a single fabric.

Security

The ESS performs LUN masking by default so all devices with LUNs defined have two levels of security, LUN masking and zoning.

This dual SAN fabric design protects against a fabric wide outage such as inappropriate or accidental change to zoning information. This zoning would only affect one single SAN fabric, so the hosts could still be able to access their storage through the alternate SAN fabric.

By implementing security within the switch we can limit individual ports to only allow specific WWPN to login, also ports may be **`shut`** by default, therefore not allowing unauthorized servers from attaching to the SAN.

“What If” failure scenarios

Here we consider the following scenarios:

- **Server:** The clustering solution will failover to the passive server dynamically.
- **Server HBA:** If one of the HBA fails, IBM SDD software will automatically failover the workload to the alternate HBA. The active server of the cluster will lose up to 50% bandwidth.
- **Cable:** If a cable between a server and the switch fails, IBM SDD software will automatically failover workload to the alternate path.The active server will lose up to 50% bandwidth.
- **Cable:** If a cable between the switch and the disk storage fails, an alternate route will be used. The cluster will lose up to 50% bandwidth.
- **Switch port:** If one of the ports fails, you may replace it using a hot-pluggable SFP.
- **Switch power supply:** Should one redundant power supply fail, the other will take over automatically.
- **Fabric:** If a fabric fails, the server will use the alternate switch to connect to the storage. The cluster will lose up to 50% bandwidth. Once the VSAN is corrected.
- **Switch:** If the switch fails, there will be no connectivity to the storage. The cluster will fail. A Second switch could be used to provide greater redundancy.
- **Storage:** If the ESS fails, the servers will not be able to access the storage. A redundant ESS may be added to mirror data from the primary ESS.

17.3.2 Multi-node clustering

In Figure 17-8 we extend the environment in Figure 17-7 on page 673. It is extended to increase the number of nodes to up to eight nodes in a single cluster using Datacenter Server 2000 for Microsoft Cluster Server.

Microsoft 2000 clustering: The Windows Server 2003 family supports server clusters for up to eight nodes. Refer to http://www.microsoft.com for more details.

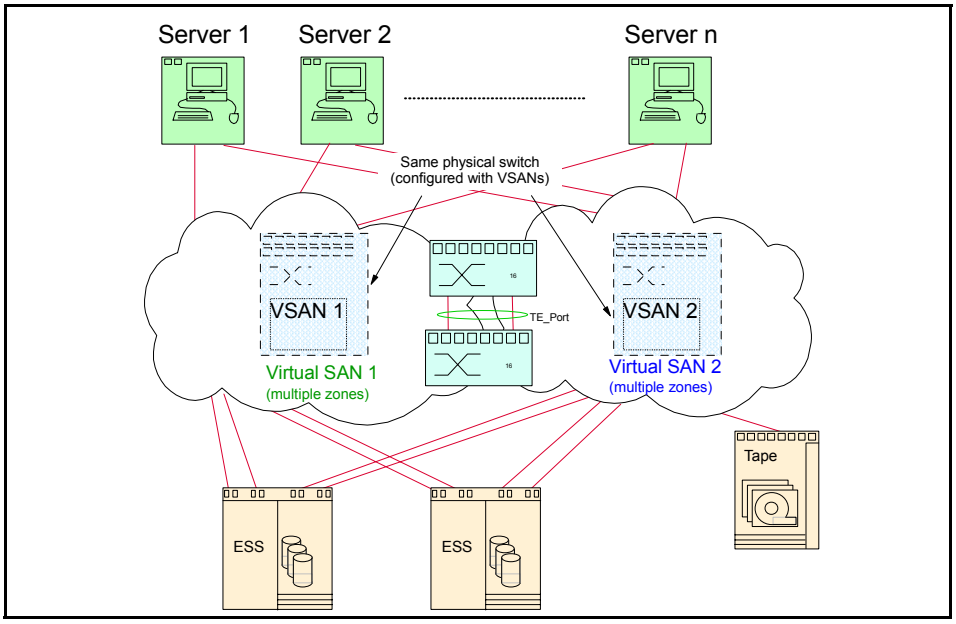


Figure 17-8 Datacenter Server 2000 for MSCS

In Figure 17-8, we show a high availability SAN design for an eight-node cluster using the Datacenter Server 2000 for Microsoft Cluster Server connected to two Cisco MDS 9216 Multilayer Fabric Switches with redundant fabric, which allows access to the two ESSs.

Apart from server failover, this design provides failover for HBAs and switches. Dual HBAs are installed in each host and each storage device must have at least two ports. Fail over for a failed path or even a failed switch is dependent on the host failover software, namely the IBM Subsystem Device Driver (SDD).

Components

- SAN fabric:
  - Two Cisco MDS 9216 Multilayer Fabric Switches
- Servers:
  - Eight IBM xSeries servers each configured with dual FC HBAs
- Storage:
  - Two ESS 2105-800 with four FC adapters
  - IBM 3590 Tape Subsystem with native FC Adapter
- Software:
  - IBM Subsystem Device Driver (SDD)

Checklist

- Install and configure switches.
- Install Fibre Channel HBAs.
- Install and certify Datacenter Server 2000.
- Configure tape subsystem.
- Configure ESS.
- Attach storage to switch.
- Attach servers to switch.
- Validate failover/failback operations.
- All storage devices, server HBAs, switches, or directors are configured with the latest IBM supported versions of driver/firmware levels.

Performance

This solution is configured based on the ratio of eight servers to two disk storage devices with a redundant fabric. Hence, the effective oversubscription will be 4:1 and is within our predefined 6:1 ratio. To increase the performance of the SAN more ESS ports may be connected to the switches.

Scalability

This design accommodates eight hosts and two Disk storage devices, a single tape connection and two Cisco MDS 9216 Multilayer Fabric Switches with no additional module installed. The sollution uses 25 available ports which may be used to connect to other storage devices. Additionally another switching module could be added to the second slot of each switch, therefore increasing the amount of available ports.

Availability

In addition to the server high availability clustering, SAN high availability is provided with this dual switch, dual VSAN design. Dual HBAs are installed in each host and the storage device must have at least two ports. Fail over for a failed path or even a failed switch is dependent on host failover software, namely, the IBM Subsystem Device Driver (SDD). Each VSAN is transverses each switch via the TE_ports, adding to the complete redundancy of the two fabrics.

Security

The ESS performs LUN masking by default so all devices with LUNs defined have two levels of security, LUN masking and zoning.

This dual SAN fabric, dual switch design protects against a fabric wide outage such as inappropriate or accidental changes to zoning information. This zoning would only affect one single SAN fabric, so the hosts could still access their storage through the alternate SAN fabric.

By implementing security within the switch we can limit individual ports to only allow specific WWPNs to login, also ports may be **shut** by default preventing unauthorized servers from attaching to the SAN.

"What If" failure scenarios

Here we consider the following scenarios:

- **Server:** The clustering solution will failover to the passive server dynamically.
- **Server HBA:** If one of the HBA fails, IBM SDD software will automatically failover the workload to the alternate HBA. The active server of the cluster will lose up to 50% bandwidth.
- **Cable:** If a cable between a server and the switch fails, IBM SDD software will automatically failover workload to the alternate path.The active server will lose up to 50% bandwidth.
- **Cable:** If a cable between the switch and the disk storage fails, an alternate route will be used. The cluster will lose up to 50% bandwidth.
- **Switch port:** If one of the ports fails, you may replace it using a hot-pluggable SFP.
- **Switch:** If a switch fails, the server will use the alternate switch to connect to the storage. The active server will lose up to 50% bandwidth.
- **Storage:** If one ESS fails, the servers will not be able to access that storage. The other ESS may be used to mirror data from the primary ESS.

17.4 Secure solutions

Any adverse effect to a SAN will typically have an impact on multiple servers within the SAN fabric. To minimize this impact, it is important to ensure that every possible security measure is incorporated into the SAN design.

17.4.1 Zoning security solution

In Figure 17-9 we look at a SAN design that could be used to provide additional security to the SAN. Using the Cisco VSAN technology we have created two isolated VSANs:

- VSAN_2 contains:
  - Our production UNIX server (Prod_UNIX), and all ESS1 ports
- VSAN_3 contains:
- Only the development UNIX server (Devl_UNIX)
- VSAN_4 contains:

  - Our NT server and ESS2 (to keep things simple, we have shown one NT server, but this could be multiple servers)

With multiple servers within each VSAN we further protect the servers, by creating a zone for each host to the ESS ports. As VSAN_3 does not contain any ESS ports, a separate IVR zone must be defined with the Devl_UNIX ports and the ESS1 ports.

By defining LUN masking from the ESS we further ensure that each server is protected and cannot access the other servers storage. For example, we don't want the Development server to ever be able to access the Production servers live data.

In our example we have defined a separate VSAN for our Development UNIX server, as we want to share ESS1 with the Production system we plan to use FlashCopy® to populate our test databases. With VSANs, we cannot have a port in more than one VSAN, as this would restrict the number of physical ports into the ESS that we would have access to, and in turn affect our ESS performance if we only had one path available. We define an IVR zone to enable an Inter VSAN route to allow traffic from our Development server in VSAN_3 to the ESS paths which are only in VSAN_2.

As we have not used the default VSAN (VSAN001) as one of our fabric environments, we create added security. When a new server or device is connected to an unsued port it will become available to VSAN001 by default. During the switch setup we choose for our default communication method to not allow communication between ports that are not within a zone. By using this method, each port must be manually defined to the correct VSAN by an authorized administrator.

For protection against cable changes, we have chosen WWPN zoning. This way, if someone does move cables, the WWPN will move with the cable and the zoning remains accurate.

In this solution we have used the WWPNs of the servers and created a zone through to the WWPN of the HBAs in each of the ESS. This compliments the LUN masking facility provided by the ESS.

The Cisco MDS 9000 family of products also allows for management authorization via a RADIUS server. In this example, we have utilized our existing RADIUS servers to authenticate any management request to update the MDS 9000 configuration. The following describes the authentication process:

- The Management client requests permission to perform a task on the director.
- The director passes the request to the RADIUS servers which check the User Name, Password and User Access Roles.

- The RADIUS server accepts or denies the access request.
- The authentication process is either accepted or denied to the management client.

.

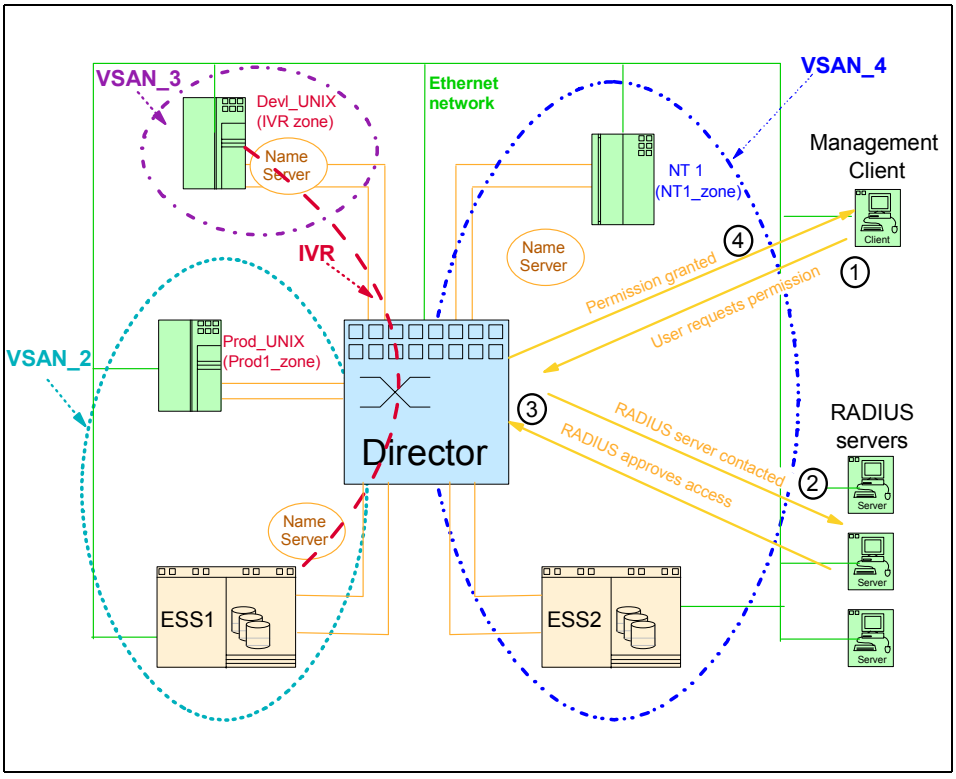


Figure 17-9 Protecting your data from both human error and sabotage

To prevent unauthorized access to our SAN fabric, all SAN components are contained in a physically secure environment. The management interfaces to the ESS and MDS 9000 Director are userid and password protected, and it is not possible to define or amend these userids or password directly at the product.

Components

- SAN fabric:
  - 1 x Cisco MDS 9506 Multilayer Director configured with 64 ports
- Servers:
  - 2 x UNIX servers configured with dual FC HBAs
  - 1 x NT server
  - RADIUS servers
- Storage:

  - 2 x ESS-800's configured with 2 x FC adapters
- Software:
  - Cisco Fabric Manager
  - SDD installed on servers
  - ESS Specialist for configuring the ESS

Checklist

We checked the following items:

- All fabric components are in a physically secure location.
- The director is installed in a lockable cabinet.
- If dual HBAs have been used, the LUNs defined on ESS1 and ESS2 have been associated with both HBA WWPNs of NT1, Prod_UNIX and Devl_UNIX.
- The userids and passwords of the management tools have been changed from the default.
- Only SAN administrators have access to the Cisco Fabric Manager userids and passwords.
- All storage devices, server HBAs, switches, or directors are configured with the latest supported versions of drivers/firmware levels.

"What If" failure scenarios

These are some theoretical assumptions:

- Access is obtained to director cabinet. If the fiber optic cables belonging to NT1 Prod_UNIX or Devl_UNIX are switched, all the correct LUNs would still be visible at the server. Name server zoning is independent of the director port position.
- Access is obtained to the Cisco Fabric Manager server. Authentication process protects the fabric from an un-authorized user.
- Access is obtained to ESS Specialist. Even if ESS LUN masking information is modified and Prod_UNIX's HBA WWN is added to NT1s LUN list, it would not be possible for UNIX1 to see NT1's LUNs as it is in a separate VSAN. The ports for ESS2 are defined to VSAN_2 and are not accessible to VSAN_1.
- If someone managed to obtain access to the equipment and install a new device into the SAN, then these ports would be part of the default VSAN and would still not be accessible to either VSAN_1 or VSAN_2.

17.5 Loop solutions

The Cisco MDS 9000 family supports a wide range of different port devices without the requirement of additional fabric components. In the solution that follows, we illustrate how the Cisco MDS 9000 family can handle FC_AL connections using the Translative Loop (TL_Port).

17.5.1 Utilizing the translative loop port

In Figure 17-10 we are able to create a single SAN fabric that incorporates a mix of devices without having to allow for the “bad habits” of a loop device, particularly a LIP, or add any additional bridging devices.

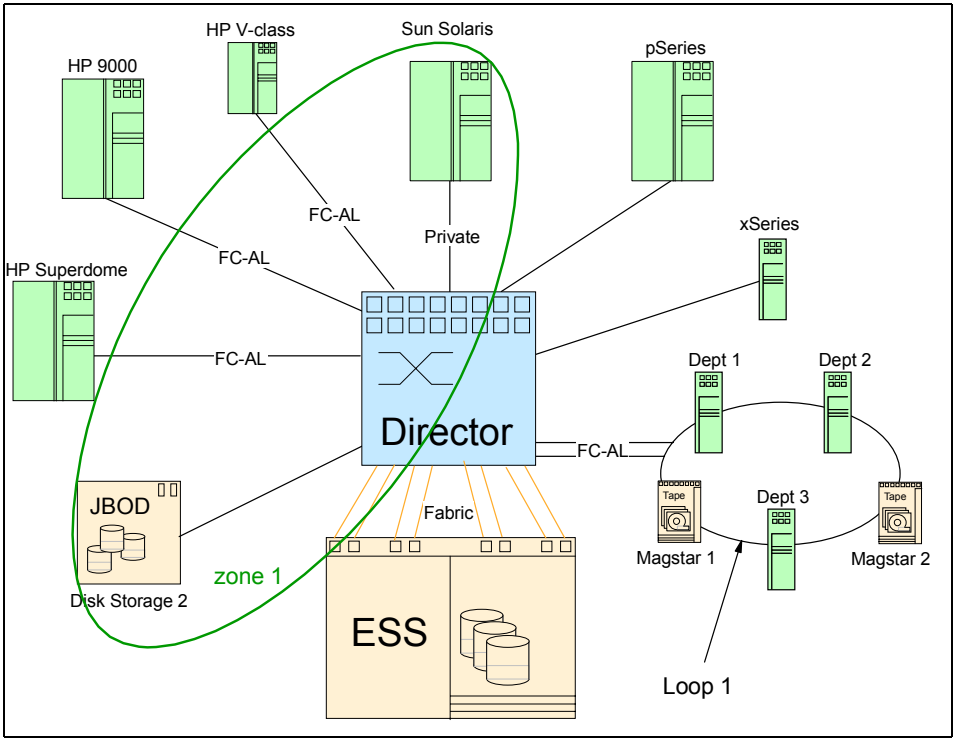


Figure 17-10 Utilizing the Cisco MDS 9000 TL_Port

For example, if Loop 1 was implemented as a traditional loop, that is to say a standalone Loop, not connected to the Cisco MDS 9509, an FC_AL event from Magstar 1 (which could be something as simple as someone powering on the tape drive), will result in the departmental servers 1, 2, and 3, and Magstar 2 devices to be reset. This is due to a LIP which is sent around the loop to all devices.

If the Departmental Servers were performing I/O, or Magstar 2 was performing a backup, this LIP would provide an unwelcome interruption to processing.

In our example, Loop 1 has been connected to a TL_Port of the director, allowing all Nodes in the loop to communicate to any of the other Nodes on the director being fabric or FC_AL devices. A LIP on Loop 1 will only affect devices on Loop 1 and not any others. If we wish a LIP from Magstar 1 to be transparent to the departmental servers and Magstar 2, each must be connected to individual TL_Ports on the director.

By manually defining each of the the director ports with FC_AL devices attached, as TL_Port, we are able to attach all our HP servers (which only support FC_AL in this example) and our IBM fabric attach pSeries and xSeries servers to the same SAN fabric, with no other additional SAN equipment required.

Components

- SAN fabric:
  - Cisco MDS 9509 Director configured with 4 x 16-port switching modules
- Servers:
  - 3 x HP servers each configured with dual FC-AL HBAs
  - 1 x SUN Server configured with dual JNI HBAs
  - pSeries configured with dual HBAs
  - 4 x xSeries servers configured with single HBAs
- Storage:
  - ESS-800 configured with 8 x FC adapters
  - JBOD disk array with dual FC-AL HBAs
- Software:
  - Cisco Fabric Manager
  - SDD installed on servers
  - ESS Specialist for configuring the ESS

Checklist

We checked the following items:

- Any disk devices that do not support LUN masking are zoned to their respective servers.
- All storage devices, server HBAs, switches, or directors are configured with the latest IBM supported driver and firmware levels.

Performance

For this solution we have no knowledge of our server's performance profile, so we assume our high performance profile of 6 server connections to each ESS storage port (refer to Section 17.1, "Performance solutions" on page 660 for

details). For the Departmental Servers, we are using our low performance profile ratio of 12 server connections to each ESS storage port. As the Magstar devices are part of the departmental loop, we will class the Magstar devices as low profile. We can ignore our Sun server, as this is attaching to its own device.

In our solution we have 5 high profile server connections to 1 ESS port and another 5 low profile device connections to one ESS port.

We are within our performance profile for both server groups.

We could also make use of the 32-port switching module by attaching servers that fit the low performance profile to this module. This will enable us to provide a better cost/port ratio and increase the port density while still managing the performance requirements of these servers.

Scalability

This solution can easily be scaled through the addition of more switching modules (either 16 or 32-port modules) and through the addition of extra switches connected via EISLs.

As there are no restrictions on how many TL_Ports can be defined or where they have to be located, we can easily build a solution to scale up to many hundreds of devices.

Security

The ESS performs LUN masking by default so all devices with LUNs defined have two levels of security: LUN masking and zoning. As Disk Storage 2 does not perform LUN masking, our only level of protection is zone 1. To guard against unauthorized changes to this zone, we could use the security management tools provided with Cisco's Fabric Manager — refer to Section 17.4, "Secure solutions" on page 678 for more details.

"What If" failure scenarios

These are some theoretical assumptions:

- If unauthorized access is obtained to the Cisco Fabric Manager, it would be possible to create a zone allowing any server to have access to Disk Storage 2's LUNs.
- If Magstar 2 is removed from Loop 1, it will effect all of the other devices in Loop 1.

18

Case studies

In this section we will portray actual end-user scenarios. Pertinent details have been extracted from RFIs/RFPs to identify the business and technology requirements. Most SAN related RFIs/RFPs surface with a contingent to consolidated storage: This may be reflected in the case studies examined, but we will endeavor to concentrate on the SAN components when recommending the solution.

We will analyze, design, and recommend solutions for each customer scenario based on three manufacturers; Brocade, Cisco, CNT, and McDATA. IBM either resells or OEM's (Other Equipment Manufacturer) these solutions.

We have attempted to use case studies that may be similar to your environment and requirements (based on field experience). Refer to the scenario that best fits your company situation/environment.

Note: Whenever we used the IBM ESS as the storage device in our case studies, we assumed that IBM ESS can sustain up to 400 MB/s and that each port can give up to 80 MB/s of bandwidth. This does not mean that you cannot use more than five ports per IBM ESS.

© Copyright IBM Corp. 2004. All rights reserved.

18.1 Case Study 1: Company One

We will detail the background and requirements of the company for this case study.

18.1.1 Company profile

Company One is a wholly owned subsidiary of MuchBigger Company, a multi-billion dollar, multi-national company that provides commercial agricultural products. Company One provides products and services to the agricultural community at various layers. Company One provides these services for a national market and manages its own datacenter following and setting standards where appropriate from the parent company. The parent company and other subsidiaries have not implemented any SAN technology to date. POPs (point of presence) are situated throughout 12 North American major metropolitan areas.

18.1.2 High-level business requirement(s)

As a new and isolated environment, Company One would like to introduce a consolidated storage solution for centralized e-mail and other ERP applications in the next two months. Initially, 10 servers will be attached but this will grow to 20 servers in the next 12 months, and potentially double the following year should the parent company adopt these applications and centralize to one location. Scheduled downtime must be restricted to four hours per month and must only occur over a weekend. Single points of failure should be kept to an absolute minimum. The environment must be available for testing in two months.

18.1.3 Current infrastructure

The existing network infrastructure will be used which is 100BaseT switched Ethernet.

18.1.4 Detailed requirements

Two new application environments need to be established:

- A new centralized Microsoft Exchange Cluster farm running initially on 8 Intel servers with Microsoft Windows 2003 operating system. The requirement for storage for these servers is 800GB of disk, evenly spread across the servers with relatively low access. This is defined as 10 I/O/s with I/Os typically having a block size of 4KB in size per server.
- A in-house ERP application server with no significant disk requirement will use two Microsoft SQL Server instances running on two clustered

(active/active) Intel servers running initially on Microsoft Windows NT SP6, but will migrate to Microsoft Windows 2003 when application testing is complete. The disk storage required for each database instance will initially be 200GB. Expected throughput requirements are 1000 I/Os per database with 4 KB blocks.

- All data will need to be backed up with the potential to restore at a remote location. The current backup mechanism will not change (i.e. Tivoli Storage Manager LAN based solution with LTO scalable library directly attached to pSeries).
- Implementation services and on-site training will be required.
- The customer does not require the latest and greatest technology and in fact, would prefer to implement a tried and tested solution.

18.1.5 Analysis (ports and throughput)

From the technical requirements defined, we will try to define the requirements for our SAN design. With regards to those requirements, we should identify the following information:

- Number of ports needed for connectivity to servers and storage devices. The best suited topology for our environment
- Throughput from each server to the storage device(s)
- How much redundancy/resilience we need (availability)
- Distances between devices, type of connectivity in the case that we have multiple sites
- Planned growth of resource needs (ports and throughput)
- The effect of maintenance upgrades (procedures, downtime, degraded bandwidth)
- What is acceptable downtime after introduction of new infrastructure?
- Implementation times
- Impact to current environment
- Required skills for production environment
- Do we need to implement some kind of disaster recovery plan including backup solution?
- Do we need to integrate any legacy devices?

You should refer to Chapter 7, “SAN design considerations” on page 243 for explanation of how various factors can affect your SAN design.

In our example we have the following requirements:

- We have to build a redundant SAN, because we can only afford scheduled downtimes. We will also use redundant paths from servers to the storage, providing higher availability on the connection level. The SAN has to be redundant in a way that will allow mandatory updates without causing downtime in the production environment and not cause performance degradation during maintenance.
- Because of planned growth we have to design a solution which will accommodate non-disruptive upgrades of the SAN infrastructure
- Eight Intel servers for Exchange farm, with two connections to the SAN, 2 x 8 = 16 SAN ports. Each server has an I/O throughput requirement of 40 KB/s. This means that we can cover all the traffic needed with only one adapter per server.
- The total capacity for these servers is 800GB; spread across all eight servers; this means 100 GB per server.
- All servers are running Microsoft Windows 2003 Server operating system.
- Two servers running Microsoft Windows NT SP6 in a cluster, with two connections to the SAN, 2 x 2 = 4 SAN ports. Each server has 4000KB which is approximately 4MB/s. We can cover all the traffic with one adapter per server.
- The total capacity for the two servers is 400 GB.
- The predicted growth in next year will be from the current 10 servers to 20 servers, thus adding 20 additional SAN connections. We assume the same bandwidth requirements for new servers.
- The possible growth in the following year is 100%. This will mean another 40 SAN connections in the second year after the initial implementation. Because of this reason our SAN design has to be ready for expansion in a non-disruptive manner. We expect that the bandwidth requirements will stay the same on new servers as they are on existing ones.
- The initial total throughput towards the storage device is 8 x 40KB = 320KB/s and 2 x 4MB = 8MB/s. In total we have 8.32MB/s required to/from storage device. For this capacity requirement, we will use two SAN ports for our storage device. We could use one, but the second one is for high availability.

To summarize, we have the following initial requirements for SAN infrastructure:

- 20 SAN ports on the server side and two SAN ports on storage side.
- The performance should not be degraded in the case of maintenance and SAN upgrades.

The growth in the first year is an additional 20 server SAN ports with the total 8.32 MB/s of throughput. In the second year we have potential expansion for an additional 40 server SAN ports with 16.64MB/s throughput.

As we can see from the bandwidth requirements we do not need to increase the throughput of the storage system. This means that we are not increasing the number of SAN ports on the storage side.

The solution used has to be certified from the vendor aspect, because we do not have any time for proof of concept. Implementation time is only two months.

18.2 Case Study 2: Company Two

We will detail the background and requirements of the company for this case study.

18.2.1 Company profile

Company Two is a regional health care provider for a large metropolitan city. Company Two provides hospital services in two hospitals (26 miles apart) for this geographic region along with 20 neighborhood clinics. Company Two is a community based non-profit organization.

18.2.2 High-level business requirement(s)

Company Two has three business objectives for this year and would like to employ economies of scale by using any overlap in these efforts. The objectives for this year are:

- Establish a disaster recovery plan using the two data centers (located at each hospital) to backup each other
- More effectively use storage by consolidating where appropriate
- Overcome apparent I/O related performance issues

18.2.3 Current infrastructure

The current Company Two IT environment is very complex. The two hospitals will be referred to as Getwell and Feelinbad. One hundred and fifty servers are spread across the two data centers (100 - Getwell, 50 - Feelinbad split on servers and storage resources). Each server has its own local internal disk, apart from 10 pSeries servers which use a cabinet of SSA disk (approximately 600 GB of which 300 GB is used) in total. IBM's HA/CMP is being used between two of the servers in active/active mode. The total disk capacity is 3.6TB, however, only

1.5TB is used. Backup's are not managed today but full file system dumps are performed on a weekly basis to local DLT and 8mm tape drives. Only 64 of the servers (56:8 between the data centers) use more than 5 GB of disk storage. Out of the remaining 86 servers, it is expected that 10% per year (for the next three years) will need to migrate to the SAN (due to disk allocation capacity, not I/O).

Servers are heterogeneous in nature: the most critical application (3D Image processing) is processing data on two SGI Origin servers (running IRIX) running independently but using two way database mirroring with Informix Continuous Data Replication. These servers each run at 5000 I/O/s (peak) with 32KB block sizes to SCSI attached disk on multiple buses. Cache is not very effective for the type of workload with only a 10% read hit ratio. The indirect costs associated with these servers being down is estimated at $80,000 per hour. The total storage allocated to the SGI complex is 400GB and this is growing at a rate of 10% month. All environments (SGI aside) are growing at an average of 10% per year. All Netware and NT servers are providing some form or file/print services to users at all locations. The total I/O workload for all servers, with more than 5 GB of disk, (except SGI) never exceeds 2000 per second at any one time; block sizes vary between 2K and 8K. No individual server (except SGI) exceeds 600 I/O/s.

In Table 18-1 we show the inventory.

Table 18-1 Case Study 2: Server and storage inventory

Hardware	OS	Total Storage for all	Block Size Range	Total I/O/S for all	Quantity	Location
HP 9000	HP/UX	70GB	4KB	4000	8	Getwell
HP 9000	HP/UX	300GB	4KB	450	3	Feelinbad
SGI Origin	IRIX	400GB	32KB	10,000	2	Getwell
pSeries	AIX	300GB	8KB	2000	10	Getwell
HP Tru64	OpenVMS	40GB	2KB - 4KB	400	6	Getwell
HP Tru64	Tru64	30GB	2KB - 4KB	600	4	Getwell
Compaq Proliant	MS/NT	60GB	2KB - 8KB	3800	10	Getwell
Compaq Proliant	Netware	60GB	4K B - 8KB	3200	10	Getwell
Compaq Proliant	Netware	120GB	4KB - 8KB	990	3	Feelinbad
Compaq Proliant	Netware	40GB	4KB - 8KB	660	6	Getwell
Compaq Proliant	Netware	80GB	4KB - 8KB	400	2	Feelinbad

The average read/write ratio across the whole server complex is 75:25.

In Figure 18-1 we show the server schematic for Case Study 2.

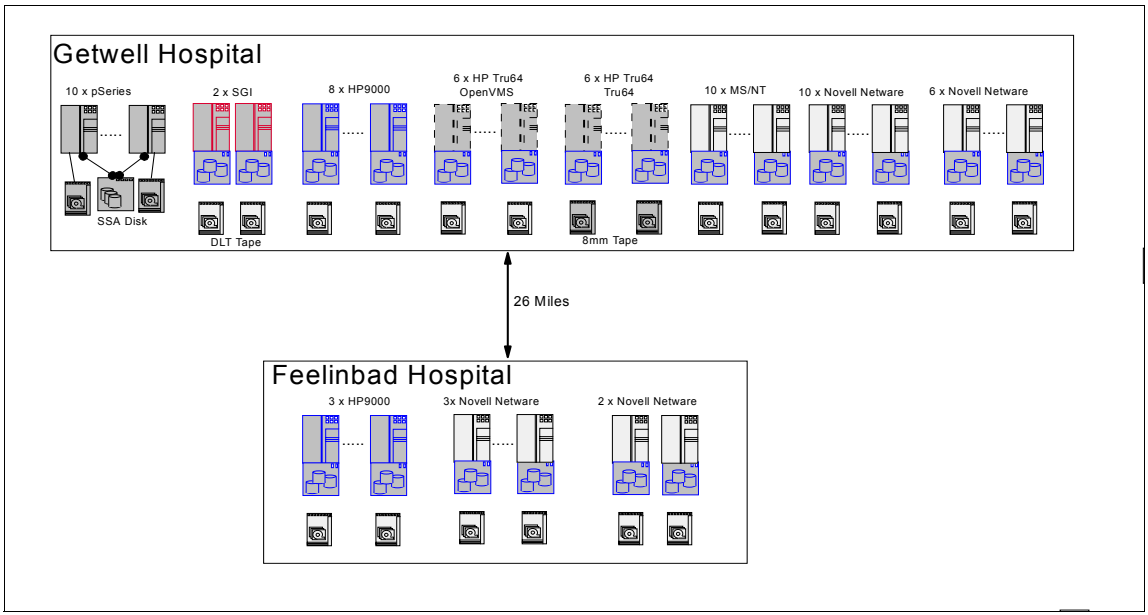


Figure 18-1 Case Study 2: Server Schematic

18.2.4 Detailed requirements

- Establish a disaster recovery plan using the two data centers (located at each hospital) as a failover site for each other.

  Company Two would like a recommendation as to whether it needs to buy additional IBM and SGI servers at the Feelinbad site or whether they should separate the existing cluster functions over distance between the two sites.

- More effectively use storage by consolidating where appropriate

  Allow for centralized management and reduce surplus of unused disk resources by allowing disk resource sharing. Use SAN functions to assist in disaster recovery where appropriate and cost effective. Re-use investment into SSA disk where possible.

- Overcome apparent I/O related performance issues

  Provide sufficient bandwidth to accommodate existing and expected growth for the next year, with consideration for the next three years. This growth will be both logical and physical (workload and servers).

18.2.5 Analysis (ports and throughput)

From the technical requirements defined we will try to define the requirements for our SAN design. With regards to those requirements we should identify the following information:

- Number of ports needed for connectivity to servers and storage devices. The best suited topology for our environment
- Throughput from each server to the storage device(s)
- How much redundancy/resilience we need (availability)?
- Distances between devices, type of connectivity in the case there are multiple sites
- Planned growth of resource needs (ports and throughput)
- The effect of maintenance upgrades (procedures, downtime, degraded bandwidth)
- What is acceptable downtime after the introduction of new infrastructure?
- Implementation times
- Impact to current environment
- Required skills for production environment
- Do we need to implement some kind of disaster/recovery plan including backup solution?
- Do we need to integrate any legacy devices?

You should refer to Chapter 7, “SAN design considerations” on page 243, for explanation of how various factors can affect your SAN design.

In our example we have the following requirements:

- We have 100 servers in first center (Getwell) and 50 servers in second center (Feelinbad). From the storage point of view we will only connect 64 of them to SAN, as the remainder are using 5 GB or less storage. The remaining servers will be included in the SAN in the following years when they exceed a certain amount of used storage, for example 10 GB.
- 56 of the servers are located at Getwell and eight of the servers are located at Feelinbad.
- 26 miles equates to 41.8 KM.
- Storage capacity at Getwell is 1000 GB and at Feelinbad is 500 GB. As a requirement stipulates a disaster recovery plan, we need to plan combined capacity of storage on both sides. In our example, we will use one IBM Enterprise Storage Server on each side.

- We have the following bandwidth requirements in Getwell site:
  - Eight servers with no more than 2.34 MB/s per server and no more than 5.62MB/s total
  - Two servers with no more than 157 MB/s per server and no more than 312.5 MB/s total
  - Ten servers with no more than 4.68 MB/s per server and no more than 5.62 MB/s total
  - Six servers with no more than 2.34 MB/s per server and no more than 1.56 MB/s total
  - Four servers with no more than 2.34 MB/s per server and no more than 2.34 MB/s total
  - Ten servers with no more than 4.68 MB/s per server and no more than 30 MB/s total
  - Ten servers with no more than 4.68 MB/s per server and no more than 25 MB/s total
  - Six servers with no more than 4.68 MB/s per server and no more than 5.16 MB/s total

  The total bandwidth for servers (excluding SGI) is 75.3 MB/s. These numbers represent peak values.
- We have the following bandwidth requirements in Feelinbad site, this:
  - Three servers with no more than 2.34 MB/s per server and no more than 7.73MB/s total
  - Three servers with no more than 4.68 MB/s per server and no more than 5.16MB/s total
  - Two servers with no more than 4.68 MB/s per server and no more than 3.13MB/s total

The total bandwidth for servers is 16.02 MB/s. These numbers represent peak values.

All together we have 91.32 MB/s data bandwidth addressing disk storage. This is important, because in the case that one storage device fails we will have all the bandwidth in one center.

As we can see from the bandwidth numbers, we can achieve the desired throughput with one FC adapter per server, except for the SGI servers. Because we need to implement a highly available redundant SAN, we will use two FC HBAs per server. This will cover the performance needs and accommodate maintenance and upgrade mechanisms. Therefore, we will need 54 x 2 = 108 SAN ports at the Getwell Center and 8 x 2 = 16 SAN ports in the Feelinbad center.

For the two SGI servers, we will use two 2 Gb/s adapters per server, which will cover the performance requirements (157MB/s per server), for maintenance and upgrades, and in the event that we lose one path. This will give us 2 x 2 = 4 SAN ports in the Getwell Center.

Note: 157 MB/s is reaching the practical limit of 2 Gb/s FC. We should consider adding an additional adapter to accommodate this.

Because SGI servers are not supported on all the storage platforms we recommend using separate storage platforms for it. This means that we can cover bandwidth needs for all non-SGI servers in the Getwell Center with one 1 Gb/s SAN port and two 2 Gb/s ports for SGI servers. Because of the high availability and redundancy requirement, we will double the storage SAN ports. With this in mind we will have 2 x 1 = 2, 1 Gb/s SAN ports and 2 x 2 = 4 with 2 Gb/s SAN ports in the Getwell Center.

In the Feelinbad Center we will have a slightly different situation. Because of the proposed disaster recovery plan we will introduce two new servers which will cover the critical applications in the event of disaster at one site. These two new servers will comprise of one SGI and one pSeries. These will need four additional SAN ports, where two of them are 2 Gb/s.

For the storage requirement, SAN ports in Feelinbad will need 2 x 1 = 2 SAN ports for all non-SGI servers, and 2 x 1 = 2 SAN ports, which will be 2 Gb/s for SGI.

Note: Because we will access the storage in Feelinbad Center from Getwell center in the case of storage failure, we need to accommodate the same speed on storage ports as Getwell. This means we need to add an additional two 2 Gb/s SAN ports for SGI servers. Therefore, we will need a total of 6 storage ports in the Feelinbad Center.

Because we have two locations we also need to plan the ISLs between those two sites. We need to plan for two types of ISLs:

- The ISLs which will be used to access the remote storage copy in the event that local storage fails.
- The ISLs which will be used for data replication between storage devices. The data replication can be also done using the operating system, but in our example we have decided to use the copy services on the level of the storage device. Performing data replication on the operating system level may introduce new performance problems, because you are utilizing the processor power and storage bandwidth of the application server.

In Figure 18-2, we show a schematic representation of dividing ISLs for data access and for data replication.

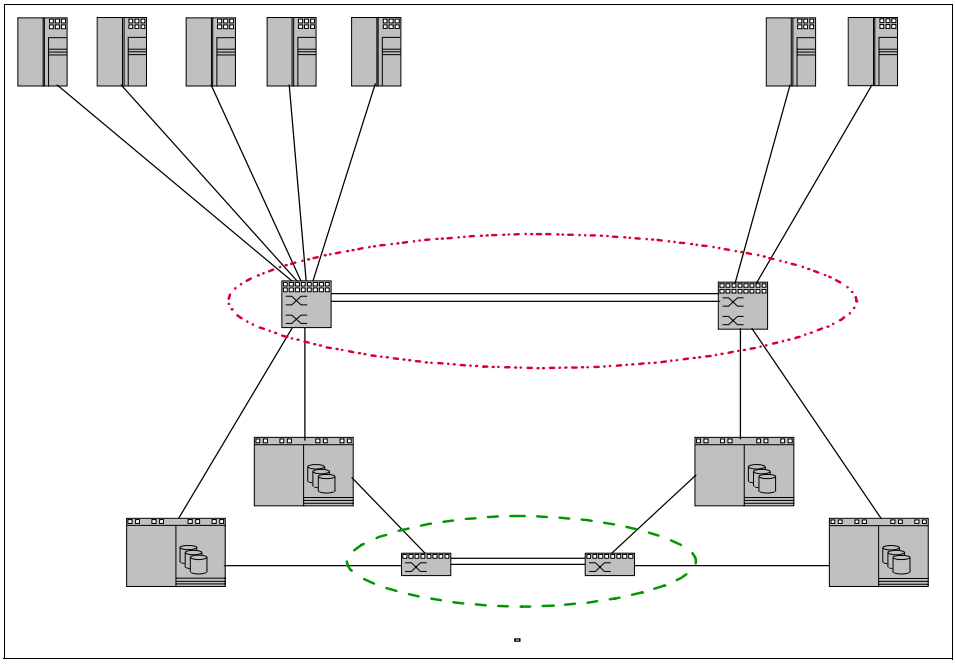


Figure 18-2 Different ISL for data access and for data replication

This does not mean that you cannot use the ports on the same switching device for both types of ISLs. In our example we used separate switching devices to clarify the point being made.

We will plan to include the existing SSA disk in an expansion frame with the new ESS.

Note: With regards to the IBM ESS, we would use Fibre Channel links for the data replication between two storage devices. If the distances are too long for standard Fibre Channel links we could use some mechanism for channel extension.

To conclude we have the following SAN port requirements:

- **Getwell Center**

  One hundred eighteen SAN ports for servers and storage access (112 for servers, 6 for storage). For those ports going to the Feelinbad site we need at least 6 ISLs, two of them to accommodate storage requirements for all non-SGI servers and four 2 Gb/s ISLs for accommodating SGI bandwidth.

  Because we will replicate the storage data to the Feelinbad Center we need to provide the same amount of ISL as we have the ports to access the storage - six.

  For data replication we will use two separate technologies:

  – For IBM ESS we will use Fibre Channel. We will use an adequate number of ESCON links. Because the sites are 26 miles apart we will use Fibre Channel directors as the extension mechanism.

  – For SGI we will use Fibre Channel connections. As we know, since only 25% of the bandwidth requirements if for writing we need only to provide one additional storage SAN ports, which will be used for data replication, and also one ISL for this purpose. In our example, we will use two storage ports going to two switches and two ISLs for redundancy purposes. All the links will be 2 Gb/s.

  The total SAN ports needed in Getwell Center for our example this is 118 + 6 + 2 + 2= 128.

- **Feelinbad Center**

  Twenty six SAN ports for server and storage access (20 for servers, 6 for storage). Of those ports going to the Getwell site we need at least 6 ISLs, two of them to accommodate storage requirements for all non-SGI servers and four 2 Gb/s ISLs for accommodating SGI bandwidth.

  The requirements for the data replications are the same as in Getwell Center.

  The total SAN ports needed in Feelinbad Center, for our example this is 26 + 6 + 2 + 2 = 36.

Some operating system levels in our example are not supported in a FC environment. We need to ensure that upgrades are performed to a correct (supported) level if the applications permit an upgrade. This can prolong the implementation phase or even the decision for the overall implementation. An option is to include only the servers which support this connectivity into SAN design initially, and add the others later.

We decided to go with a 2 Gb/s solution because of performance reasons for the SGI servers.

For connecting the remote sites (26 miles), we will use DWDM equipment.

The planned growth is 10% per year. To cover all non-SGI servers, we do not need to add any new ports, because 2 Gb/s technology can cover all growth. For the SGI servers, additional four 2 Gb/s ports on the storage side and four 2 Gb/s server ports on each side. We will not introduce any new ISLs between the sites to accommodate the new SGI bandwidth, because we have enough bandwidth in non redundant mode in that case. All the connections on both sites are duplicated, this means that in this case we will have four connections from each server to the storage, but only half of this will be utilized. This give us the need for only four ISLs to the backup center. This links will be only utilized when we will have storage failover and we assume that this failure will be resolved quickly. We also do not need any new ports for data replication. Therefore, an additional 8 SAN ports in Getwell Center and 8 SAN ports in Feelinbad Center will be needed in the following two to three years.

18.3 Case Study 3: Company Three

We will detail the background and requirements of the company for this case study.

18.3.1 Company profile

Company Three is a large independent insurance brokerage company representing a multi-regional market. They provide a wide spectrum of insurance products and services to businesses, individuals, and families. Company Three attracts over $150 million of revenue per year.

18.3.2 High-level business requirement(s)

- Implement a consolidated storage solution to satisfy growth of NT platform which is nearing capacity.

18.3.3 Current infrastructure

Seven servers: Five Microsoft Windows NT Server , one pSeries and an iSeries. Windows NT environment performs Microsoft Exchange and file/print server functions and must be available during office hours (Monday through Friday, 8am - 5pm). The pSeries needs to process data that is created on the iSeries and then needs to share or send the data to an NT file server. FTP is currently used for the data which is approximately 20 MB in size. The iSeries runs a proprietary core business application and needs to be available 24x7 for processing (batch and online). The cost to the business of the iSeries being unavailable during office hours for online processing is $30,000 per hour. The cost of any of the NT servers being down is estimated to be $1,000 per hour.

NT accounts for 57GB of data, pSeries has 8GB, and iSeries has 36GB. In Figure 18-3 we show the server schematic.

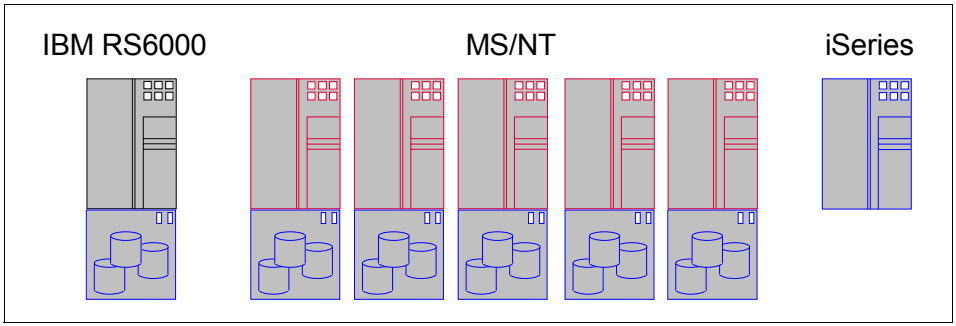


Figure 18-3 Case Study 3: Server Schematic

18.3.4 Detailed requirements

Implement a consolidated storage solution to satisfy growth of NT platform which is nearing capacity.

Three of the NT servers are rapidly reaching capacity with no additional disk bays available within the servers. The immediate need is to satisfy this disk requirement. Overall growth in storage appears to be at the rate of 15% per year. Company Three would like to pursue opportunities for server consolidation for file/print services and would also like to make the Exchange application more resilient. I/O has not been perceived to be a problem within the organization and a solution is sought that will not introduce new bottlenecks.

18.3.5 Analysis (ports and throughput)

From the technical requirements defined, we will try to define the requirements for our SAN design. With regards to the requirements we should identify the following information:

- Number of ports needed for connectivity to servers and storage devices. The best suited topology for our environment
- Throughput from each server to the storage device(s)
- How much redundancy/resilience we need (availability)
- Distances between devices, type of connectivity in the case that we have multiple sites
- Planned growth of resource needs (ports and throughput)
- The effect of maintenance upgrades (procedures, downtime, degraded bandwidth)
- What are acceptable downtimes after the introduction of the new infrastructure?
- Implementation times
- Impact to current environment
- Required skills for production environment
- Do we need to implement some kind of disaster/recovery plan including backup solution?
- Do we need to integrate any legacy devices?

You should refer to Chapter 7, “SAN design considerations” on page 243 for explanation of how various factors can affect your SAN design.

In our example we have the following requirements:

- Five Windows NT servers with 57 GB of storage, one pSeries server with 8 GB of data and iSeries with 36 GB of data.
- Because we are not solving any performance constrains we can provide enough bandwidth with one SAN port per server; to accommodate high availability we will use two SAN ports per server, thus having 7 x 2 = 14 SAN ports.
- We will also have two SAN ports on the storage side, providing high availability.
- An iSeries must have a minimum requirement of OS/400 V5.2 level to support Fibre Channel connections for SAN connectivity.

- There is a need to transfer 20 MB of data every night from the pSeries to the Windows NT platform. Because of this, we should consider the implementation of data sharing across the SAN, utilizing the infrastructure.

If we summarize, we have the following initial requirements for SAN infrastructure:

- Fourteen SAN ports on server side and 4 SAN ports on storage side.
- iSeries box needs a minimum OS/400 V5.2 level to support SAN connectivity via Fibre Channel.
- We do not have any performance issues, because the SAN capacity exceeds the needed bandwidth.
- The SAN should be designed with redundancy and high availability in mind because the downtime costs are significant and we have an application which has to be available 24 x 7. This means that SAN should be available during maintenance and upgrades.

The planned growth is 15% per year. This could imply we will need two more SAN storage ports in the following year.

18.4 Case Study 4: Company Four

We will detail the background and requirements of the company for this case study.

18.4.1 Company profile

Company Four is a large national financial investment company. They have two data centers in large cities (known as east and west). The heritage of the company's IT base has been with mainframe systems running z/OS, however, over the past years they have experimented with UNIX servers at the east facility to try to reduce costs. Key business applications remain on the mainframe systems at both sites, loss of revenue due to the failure of one of these systems is estimated to be $300,000 per hour. The UNIX servers run less critical applications with an estimated $10,000 per hour should any one be down.

18.4.2 High-level business requirement(s)

All disk storage at the east site is coming off a three year lease in six months. Company Four must provide a new solution that will satisfy the company's growth and performance requirements for the coming three years.

18.4.3 Current infrastructure

The servers at the east facility of Company Four consist of a mix of mainframe and UNIX servers. Three enterprise disk subsystems exist that are of the same type, two are used for mainframe storage and one is used for UNIX servers

The mainframe systems reside on two z900 units. These servers run CICS and DB2 databases with ESCON attached disks. Total disk space for these systems is 3.6 TB and I/O peaks at 4000 I/O/s (56 KB per I/O). Twelve ESCON channels are dedicated for DASD from each server to two ESCON directors which in-turn connect to two consolidated storage devices.

The UNIX servers comprise of 2x pSeries servers (production), 2x pSeries (test/QA) servers (all running AIX), 2x SUN UE6000 (production) and 2x SUN UE3000 (test). All of these servers directly connect via UltraSCSI connections to a consolidated storage device. All servers run in high availability mode (active/passive), the pSeries' use IBM's HACMP, and the Sun servers rely on Veritas for this function.

The production pSeries servers run DB2 and the active server currently use 1.1 TB of disk space. I/Os peak at 1000 I/O/s for this server and are typically 8 KB. The test pSeries servers have a similar configuration to that of production (to enable full *functional* testing of product) but only use 200 GB of disk space. I/Os reach 400 I/O/s during testing periods only, which can be scheduled during non-peak hours on a monthly basis.

The Sun server runs BEA Tuxedo applications. Tuxedo queues only account for 20GB of disk space, server cache is effectively used for these short lived transactions and the I/O rates peak at 80 per second (8 KB blocks). The test Sun servers have a similar configuration to that of production (to enable full *functional* testing of product), I/O rates and disk allocation is minimum. The application requires that both the product Sun server and the production IBM server be running together for full application availability. Nightly feeds from the mainframe populate parts of the DB2 database. Two GB of data is typically transferred over the backbone network.

All servers interconnect using a 1 Gb/s Ethernet backbone network. Backups are currently being performed over the network to Tivoli Storage Manager running on the mainframe: The same mainframe system is ESCON attached via 8 paths to an IBM tape library with a VTS (Virtual Tape Server) with twelve 3590 tape drives.

The average read/write ratio is 80:20 across the complex.

The east and west sites are currently interconnected using two T1 circuits.

In Figure 18-4 we show the case study schematic.

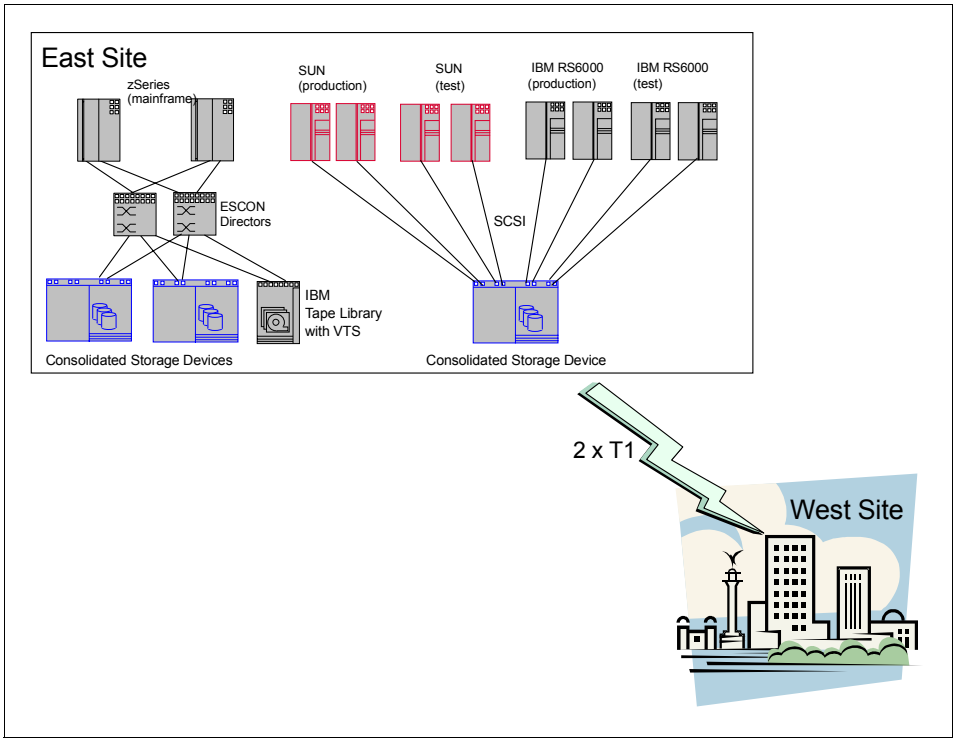


Figure 18-4 Case Study 4: Server schematic

18.4.4 Detailed requirements

- All disk storage at the east site is coming off a three year lease in six months. Company Four must provide a new solution that will satisfy the companies growth and performance requirements for the coming three years.

  As Company Four must replace the existing equipment, they would like to take advantage of Fibre Channel/FICON technology for both flexibility, performance, overcome cable distance limitations, and to position themselves better for a disaster recovery solution. The solution must position Company Four to enable them to accommodate emerging technologies as they become available without replacing the investment.

  If storage consolidation can be performed to a greater degree than exists today, without impacting performance, then this must be considered. Any opportunity to ease cable management must be taken advantage of along with the ability to have the servers 300 meters away from the storage device due to space limitations after pending construction.

Consideration to a better disaster contingency plan must be given, however the details are not yet available (assumptions should be made that all production servers will be replicated at the remote site). The east and west data centers are 300 Km apart; the proposal should reflect options that will allow for full mirroring (synchronous or asynchronous) and highlight any areas where data integrity issues (i.e. loss of data or data corruption) may surface.

18.4.5 Analysis (ports and throughput)

From the technical requirements defined, we will define the requirements for our SAN design. With regards to those requirements, we should identify the following information:

- Number of ports needed for connectivity to servers and storage devices. The best suited topology for our environment
- Throughput from each server to the storage device(s)
- How much redundancy/resilience we need (availability)
- Distances between devices, type of connectivity in the case that we have multiple sites
- Planned growth of resource needs (ports and throughput)
- The effect of maintenance upgrades (procedures, downtime, degraded bandwidth)
- What are acceptable downtimes after introduction of new infrastructure?
- Implementation times
- Impact to current environment
- Required skills for production environment
- Do we need to implement some kind of disaster/recovery plan including backup solution?
- Do we need to integrate any legacy devices?

You should refer to Chapter 7, “SAN design considerations” on page 243, for explanation of how various factors can affect your SAN design.

In our example we have the following requirements:

- There is 3.6 TB of ESCON attached zSeries storage for two zSeries servers in the east location. The I/O peak throughput is 218 MB/s
- Two pSeries database servers for a production system running AIX HACMP implementation (active/passive) with 1.1 TB of storage. The I/O peak throughput is 7.82 MB/s.

- Two test pSeries database servers for test system running AIX HACMP implementation (active/passive) with 200 GB of storage. The I/O peak throughput is 3.12MB/s.
- Two SUN production servers. They need 20 GB of storage with peak I/O throughput 640 KB/s
- Two SUN test servers. No significant requirements for storage space and I/O bandwidth.

As the customer is requiring consideration for a disaster recovery plan we will assume that we have all production servers also in the west location.

As you can see from the requirements we can satisfy the bandwidth requirements for all open servers with one SAN port per server. But because we have to implement the redundant fabric without single points of failure we will use two SAN ports for each server. This will give us 8 x 2 = 16 SAN ports for the open systems servers in the east location. To accommodate the storage device bandwidth we will use two SAN ports for open systems servers.

For the west servers we will have 4 x 2 = 8 SAN ports for open systems servers and two SAN ports for storage device.

We will use one IBM ESS at each location for open systems storage. We will connect these two storage devices with ESCON over ATM channel extenders to allow peer-to-peer Remote Copy. Because of the low bandwidth requirements, we can use asynchronous copying of data between storage devices.

We will not accommodate any Fibre Channel ISLs; in the event of server or storage device failure at the east site, we will move entire application to the west site.

Note: The latency of 500 miles Fibre Channel over ATM would be too high to have the servers in one site and data in the other site.

For backing up open system servers we will continue to use the TSM server on the zSeries. The backup will be performed over dedicated network backbone and only at one site.

For zSeries, we will use FICON connectivity for storage devices. We will have four FICON channels per server (recommended ESCON to FICON ratio is 6:1). We will use the same IBM ESS as for the open systems servers. For those eight FICON connections we will have eight FICON connections on IBM ESS.

For the disaster recovery site we will implement XRC for zSeries systems data replication from east to west site, using Fibre Channel over ATM channel extenders. For the XRC implementation we will add an additional four Fibre Channel adapters to the local ESS. On the remote site we will have an additional four Fibre Channel adapters in one of the servers. This server will replicate the data using XRC from the primary site (east) to the secondary site (west). The Fibre Channel connections for XRC will be attached directly from the IBM ESS at the east site to a zSeries at the recovery site (west).

We will replace ESCON connections for the zSeries tape backup with FICON, two ports for each mainframe server plus two FICON ports for each tape drive controller.

This gives us a total, in the east site, of:

- Eighteen SAN ports for open system servers and storage
- Eighteen FICON SAN ports for mainframe systems and their storage

For the west site we have total of:

- Eight SAN ports for open system servers and storage
- Eighteen FICON SAN ports for mainframe systems and their storage

We will accommodate six ESCON channels for the open systems' PPRC.

18.5 Case Study 5: Company Five

We will detail the background and requirements of the company for this case study.

18.5.1 Company profile

Company Five represents an autonomous research department within a large company. They have no direct revenue and rely on external sponsorship for funds. They collect data from surveys and create various analytical models. The department's strategic direction is to move their application to a Linux platform within the next 18 months to improve application availability.

18.5.2 High-level business requirement(s)

- Accommodate relatively large amounts of data in a scalable storage solution and overcome existing storage capacity limitations.
- Improve backup throughput to reduce the backup time by at least a factor of ten.

18.5.3 Current infrastructure

Server environment consists of 3 Dell servers. Two of the servers are running Microsoft Windows NT, the third is running Redhat Linux . One of the NT servers is used purely for file serving and uses all five of its internal 18 GB disks. This server is the only area of growth, which is 30% per year. The other NT server is running home grown applications that model the survey data, this data resides on five 36 GB drives and is reaching capacity.

The Linux platform has an identical hardware configuration to the second NT server and is being used to develop and port the applications from NT. I/O from the first NT server are very low — peaks in the region of 300 I/O/s with an average of 2 KB blocks.

The second NT server sustains 1000 I/O/s for periods of 2 hours at a time when a model is running. It appears to be CPU bound during this time and not I/O constrained; I/Os tend to be 16-32 KB. It is expected that the Linux implementation will perform the with same basic I/O characteristics. There will be a substantial period of time when both the NT and Linux versions of the application will need to process the same data, for QA purposes.

A lot of time appears to be wasted, with additional processing and effort, as the NT server tends to crash a couple of times a week mid-way through the model generation process.

Data arrives from sources every Monday and is added to the primary NT server. The data is maintained on a rolling two year basis, that is, as new data is added the oldest data is dropped. Before the new data is applied, a full backup is taken of both NT servers. The backup software (using MS/NT backup utility) runs on the file server, which backs up data from both NT servers to a locally attached tape drive (DDS-3), which sustains 5 MB per second. No data modeling occurs while he backup is taking place as performance is severely impacted.

There is one logical network which is 100BaseT switched Ethernet.

In Figure 18-5 we show the server schematic.

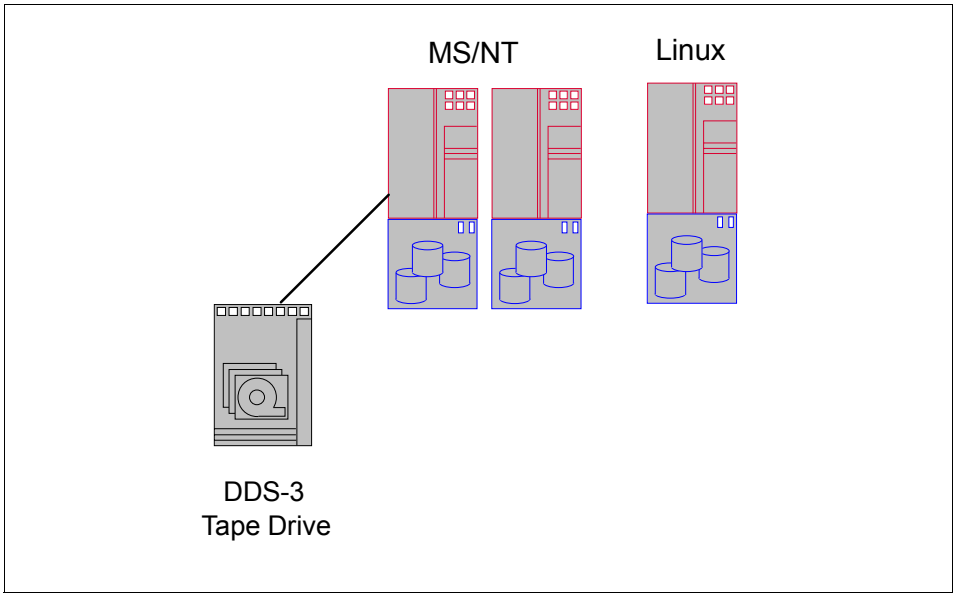


Figure 18-5 Case Study 5: Server Schematic

18.5.4 Detailed requirements

- Accommodate relatively large amounts of data in a scalable storage solution and overcome existing storage capacity limitations.

  The existing file server is using 65 GB of its 72 GB disk space (after RAID5). Each of the other servers (as the NT server is replicating data to the Linux server on a weekly basis) uses 120 GB of the 144 GB (after RAID5). The amount of data is expected to remain constant over the next 3 years along with the amount of processing.

- Improve backup throughput to reduce the backup time by at least a factor of ten.

  Company Five has concerns regarding the current backup technique (software and hardware) when they do eventually migrate to Linux as they would like to have one standard automated solution. All data must be retained for 10 years with the ability to access it on demand.

  They would like to have an environment that will support diskless servers.

18.5.5 Analysis (ports and throughput)

From the technical requirements defined, we will try to define the requirements for our SAN design. With regards to those requirements, we should identify the following information:

- Number of ports needed for connectivity to servers and storage devices. The best suited topology for our environment
- Throughput from each server to the storage device(s)
- How much redundancy/resilience do we need (availability)
- Distances between devices, type of connectivity in the case that we have multiple sites
- Planned growth of resource needs (ports and throughput)
- The effect of maintenance upgrades (procedures, downtime, degraded bandwidth)
- What are acceptable downtimes after the introduction of new infrastructure?
- Implementation times
- Impact to current environment
- Required skills for production environment
- Do we need to implement some kind of disaster/recovery plan including backup solution?
- Do we need to integrate any legacy devices?

You should refer to Chapter 7, “SAN design considerations” on page 243, for explanation of how various factors can affect your SAN design.

In our example, we have the following requirements:

- One Windows NT server with 72 GB of disk space and 600 KB/s peak bandwidth requirements.
- One Windows NT server with 144 GB of disk space and 31.25 MB/s peak bandwidth requirements.
- One Linux server with 144 GB of disk space and 31.25 MB/s peak bandwidth requirements.

As we can see, we can solve the bandwidth requirements for each server with one SAN port. Because we want to build redundant connections from servers to the SAN fabric, we will use two SAN ports per server.

Bases on the bandwidth and storage requirements, we will introduce a storage device with two SAN ports for redundancy.

For the backup we will introduce an additional Windows NT server running Tivoli Storage Manager Server and Fibre Channel attached backup device. For this we will need two additional SAN ports for the backup server and two FC-AL ports for Fibre Channel tape drives in an IBM UltraScalable Tape LTO Library. By using Tivoli Storage Manager, we will not perform full backups every time; and initial full backup will be followed with progressive weekly and/or daily backup(s).

The backup server will require additional storage on the storage device for a storage pool, which will be used for faster backup/restore and tape reclamation.

On the Windows NT platform we will use LAN free backup.

We do not recommend diskless servers at this time. The primary reason being attributed to the added complexity of troubleshooting when errors are encountered during the server's boot sequence. Multiple drives will be relieved during the disk consolidation process, these can be used for spare drives should an internal system drive fail.

The data between Windows NT and Linux application servers can be shared over the SAN using Tivoli SANergy. The Tivoli SANergy Meta Data Controller will run on the backup server. This also means that the shared data will be backed up from Windows NT server using LAN free backup. On the Linux application server only the system data will be backup using LAN based backup on a dedicated network.

Because the same backup platform will be used for both applications servers, the migration to the Linux platform could be eased with the use of the backup software.

So the total SAN port requirements are as follows:

- Eight SAN ports for all servers
- Two SAN ports for storage
- Two FC-AL SAN ports for tape drives

18.6 Case Study 6: Company Six

We will detail the background and requirements of the company for this case study.

18.6.1 Company profile

Company Six is a fast moving consumer goods (FMCG) purveyor with depots that are geographically dispersed. They rely upon the data that they gather to ensure that goods are delivered in a timely manner to their stores. Their system is also responsible for making sure that their transport infrastructure is reactive to changes in the market and road conditions.

18.6.2 High-level business requirement(s)

- To increase data availability and accommodate the storage requirements for another UNIX server.
- Mirror all data at a remote site which is 900 miles (1448 KM) away
- Establish better backup/restore capabilities that can share the existing tape resources with no manual intervention

18.6.3 Current infrastructure

Company Six invested in a consolidated storage device within the past year to accommodate the storage requirements of their two UNIX servers. The consolidated storage device has the capacity to grow to over 4TB, however it only has two host adapter connections with no further expansion capabilities. Each host adapter is directly connected using Fibre Channel to a server. One of the servers is a Sun UE4500 and the other is an IBM pSeries server. The Sun has 1.2 TB of data stored on the storage array, with I/Os peaking at 2000 per second with an average block size of 24KB. The IBM server has 800 GB of data with I/Os peaking at 1200 per second with a block size of 8 KB.

The workloads on these two machines tends to be time correlated. That is, peaks in the workloads will tend to occur simultaneously across the two servers.

The network is based on ethernet backbone.

The read/write ratio is 70:30 across the applications.

A direct Fibre Channel connect IBM tape library with two tape drives are available for use between the two servers for backup purposes.

In Figure 18-6 we show the server schematic for Case Study 6.

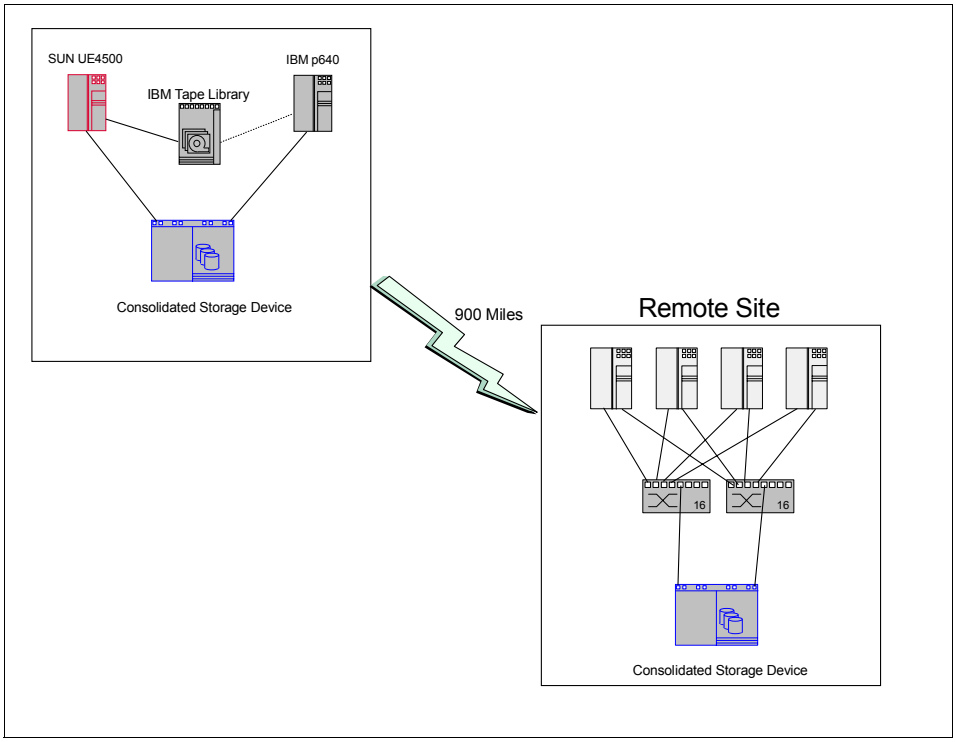


Figure 18-6 Case Study 6: Server Schematic

18.6.4 Detailed requirements

- To increase data availability and accommodate the storage requirements for another UNIX server.

  Company Six recently experienced an SFP failure in one of the host adapters in the storage array which resulted in an eight hour outage (primarily due to problem determination). This condition must not occur again. Effort must be made to increase the fault tolerance level and provide monitoring capabilities to the Systems Management application (HP/Openview) to identify any failing devices.

  An additional IBM UNIX server is planned to be deployed as an additional tier to this application. I/O and disk is expected to be relatively low (lt 100 I/O/s with a block size of 8K). The 40 GB required for this application should be placed on the consolidated storage device.

- Mirror all data at a remote site which is 900 miles away.

  Mirror all data at a remote site which is 900 miles away, minimizing the impact to the application with no data integrity issues. Performance degradation (in relation to the I/O/s) is acceptable but not above 20% of what is being achieved today. Servers are available at the remote site that can assume the workload, however, they have no SAN connectivity. There is a Compaq SAN at the remote site already using two Compaq's Fibre Channel SAN Switch 16-EL. Four servers (Compaq running NT), are currently dual connected to the switches. One RA8000-FC storage device with 200 GB of data is also connected to the switches with one path to each switch. There are currently no I/O bottlenecks at the remote site.

- Establish better backup/restore capabilities that can share the existing tape resources.

18.6.5 Analysis (ports and throughput)

From the technical requirements defined, we will try to define the requirements for our SAN design. With regards to those requirements, we should identify the following information:

- Number of ports needed for connectivity to servers and storage devices. The best suited topology for our environment
- Throughput from each server to the storage device(s)
- How much redundancy/resilience we need (availability)?
- Distances between devices, type of connectivity in the case that we have multiple sites
- Planned growth of resource needs (ports and throughput)
- The effect of maintenance upgrades (procedures, downtime, degraded bandwidth)
- What are acceptable downtimes after introduction of new infrastructure?
- Implementation times
- Impact to current environment
- Required skills for production environment
- Do we need to implement some kind of disaster/recovery plan including a backup solution?
- Do we need to integrate any legacy devices?

You should refer to Chapter 7, "SAN design considerations" on page 243, for explanation of how various factors can affect your SAN design.

In our example we have the following requirements on the primary site:

- One SUN server with 1.2TB storage and peak I/O throughput of 46.88 MB/s. Currently, this server is only using one Fibre Channel adapter, which is directly connected to the storage device.
- One IBM pSeries server with 800 MB of storage and peak I/O throughput of 9.38 MB/s. These is only one Fibre Channel connection from the server to the same storage device as SUN server.
- On the primary site we have a storage device which only has (and can only have) two SAN ports.
- An additional IBM pSeries server will be implemented with the need for 40 GB of the storage space and peak I/O throughput of 800 KB/s.
- Because of recent failure, we will design a redundant SAN with two ports in each server. This will give us six SAN ports per server, and two 1 GB ports for the storage device.

In our example, we have the following requirements at the secondary site:

- Four Windows NT server with two SAN ports each, connected to existing redundant SAN with two independent switches.
- We also have the SUN and IBM server available on the secondary site for case of disaster recovery. This gives us six additional SAN ports for the servers.
- The storage device on the remote site is currently using one SAN port per existing Fibre Channel switch, of which there are two.

For replicating the servers across the sites we will use operating system level mirroring. The AIX system will use its built-in LVM (Logical Volume Manager) and SUN we will use Veritas Volume Manager. The distance of 900 miles will give us 6.912 milliseconds latency. For establishing the Fibre Channel connection between the two sites, we will use DWDM products on each site. The Windows NT storage at the remote site will not be replicated back to the home site.

Note: Typically, latency is 4.8 microseconds for one kilometer.

For the data replication, we will use two ISLs, which should accommodate the bandwidth required, Non SGI

Backup will be accomplished using Tivoli Storage Manager. This software will reside on the new pSeries. Existing servers will use LAN-Free clients to backup their data through the SAN. The existing tape library will be connected to the SAN.

To summarize we have the following SAN ports requirements at the primary site:

- Six SAN ports for servers
- Two SAN ports for the storage device
- Two SAN port for the tape device
- Two ISLs for data replication

At the secondary site we have the following requirements:

- Six SAN ports for UNIX servers
- Eight SAN ports for Windows NT servers
- Two ports for storage device
- Two ports for ISLs for data replication

19

IBM TotalStorage SAN b-type family solutions

In this chapter we will describe solutions that are based upon the IBM TotalStorage SAN Switch range of products.

© Copyright IBM Corp. 2004. All rights reserved.

19.1 Case Study 1: Company One

If we consider the company and its requirements, as detailed in 18.1, “Case Study 1: Company One” on page 686, we will propose two designs. In one design, we will use director class products and in the other we will use switches.

19.1.1 Switch design

In the switch design we have decided to use two 16-port Fibre Channel switches for our design. You can see the proposed design in Figure 19-1.

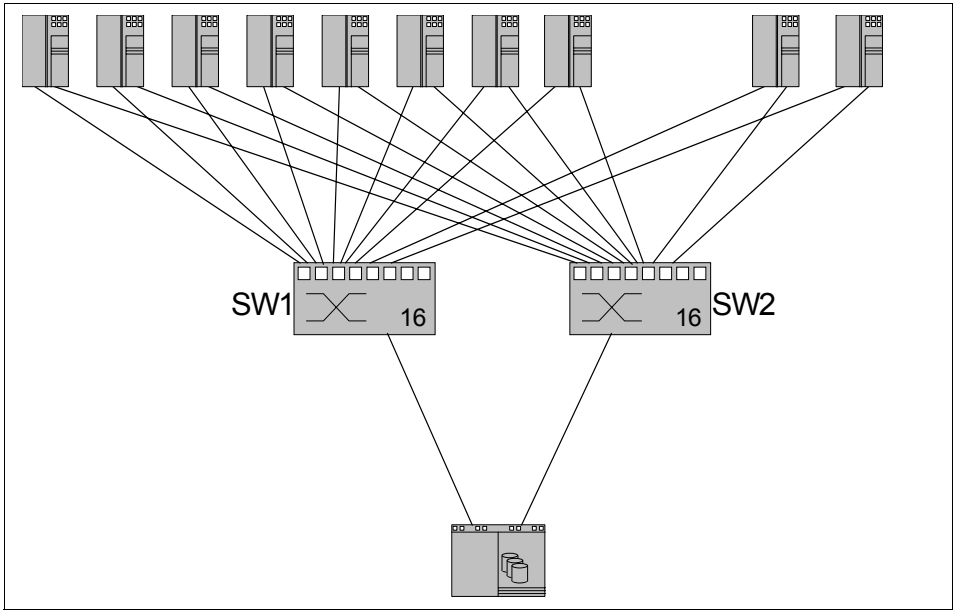


Figure 19-1 Core SAN design

In Table 19-1 we show the number of ports used for the design.

Table 19-1 Number of used ports

Ports	Servers	Storage	Spare
SW1	10	1	5
SW2	10	1	5
Total	20	2	10

Because we are using two paths from each server to the storage, we are providing redundancy and high availability. To utilize this physical setup we will use multipathing software on each of the servers. With such a design we are fulfilling these requirements:

- No single point-of-failure: redundant SAN
- All bandwidth requirements are met (40 KB/s and 4 MB/s from servers and 8.32 MB/s to storage)
- SAN components can be upgraded without impact on the servers. In case of upgrade or maintenance of switches, we can first upgrade one switch, while paths across the second one are still available. After the traffic is established back on the upgraded switch, we can upgrade the second switch.
- Possible growth without impact on the production. Because we are using a redundant SAN we can introduce additional switches without downtime on the existing servers.

Because our design will expand in the future we will use the other five ports for these functions:

- Three ports per switch will be reserved for future expansion, which will give us the option to connect three additional pairs of switches to the existing SAN without disturbance
- We will use one port per switch for maintenance purposes. For example, if one port fails this one can be used for replacement. It can be also used to diagnose problems in the SAN

This will leave us one additional port for expansion.

The expansion for the first year is ten additional servers. This means that we have to provide ports for nine of them because we already have one. We will achieve this by adding two new switches to the SAN. We will connect them through the ports we prepared on the first two switches, as we know that the design requirement was to expand without impact on the production servers.

In Figure 19-2, you can see that we have added an additional two switches to the SAN.

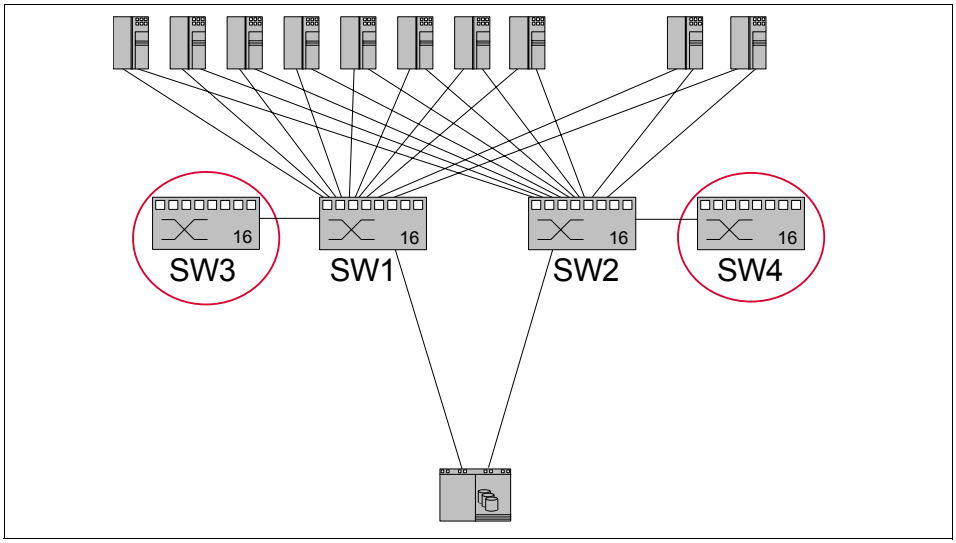


Figure 19-2 Adding additional two switches to the SAN

In this step, we added two additional switches without disturbing the current traffic on the SAN. Both paths from each server to the storage are still available, this means that during the expansion, our SAN is still in redundant operation.

Note: The configuration of SW3 and SW4 should be cleared before we introduce ISLs to SW1 and SW2.

In Figure 19-3, we show how new servers in the second year can be connected to the SAN.

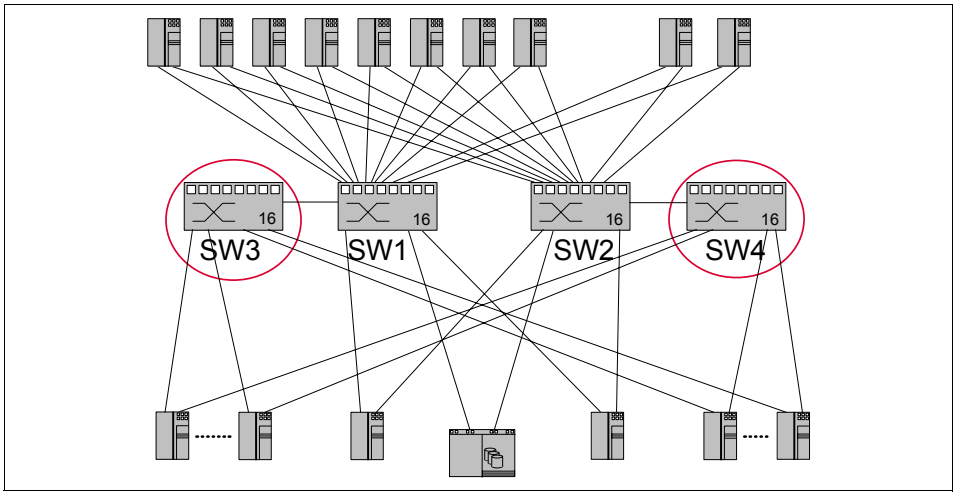


Figure 19-3 SAN after first year of expansion

In Table 19-2 we show the port usage.

Table 19-2 Number of used ports

Ports	Servers	Storage	Maintenance and expansion	ISLs	Spare
SW1	11	1	3	1	0
SW2	11	1	3	1	0
SW3	9	0	2	1	4
SW4	9	0	2	1	4
Total	40	2	10	4	8

We will assume the same bandwidth requirements for the new servers as the requirements for the original ones. We chose to use only one ISL between the switches, because it can handle the bandwidth requirements (8.32MB/s). We also do not need to expand the number of storage SAN ports.

Note: Sometimes it is recommended to introduce additional storage ports, because if we get a large number of servers accessing the same storage ports, we could get congestion on that port. This will not be a bandwidth problem, but the problem of handling a lot of connections at the same time.

We connected the new servers as follows:

- One server on switches SW1 and SW2
- Nine servers on switches SW3 and SW4

On the new switches we will reserve some ports for future use:

- One port for a storage connection
- One port for maintenance

With a separate storage connection we can actually separate the storage bandwidth from the servers on SW1 and SW2 from the servers on SW3 and SW4. After this setup, we will have four ports available on each switch for future expansion.

For the potential expansion to up to 40 servers, we need an additional 40 SAN ports. For this we will need to connect two additional switches to switches SW1 and SW2, as shown in Figure 19-4.

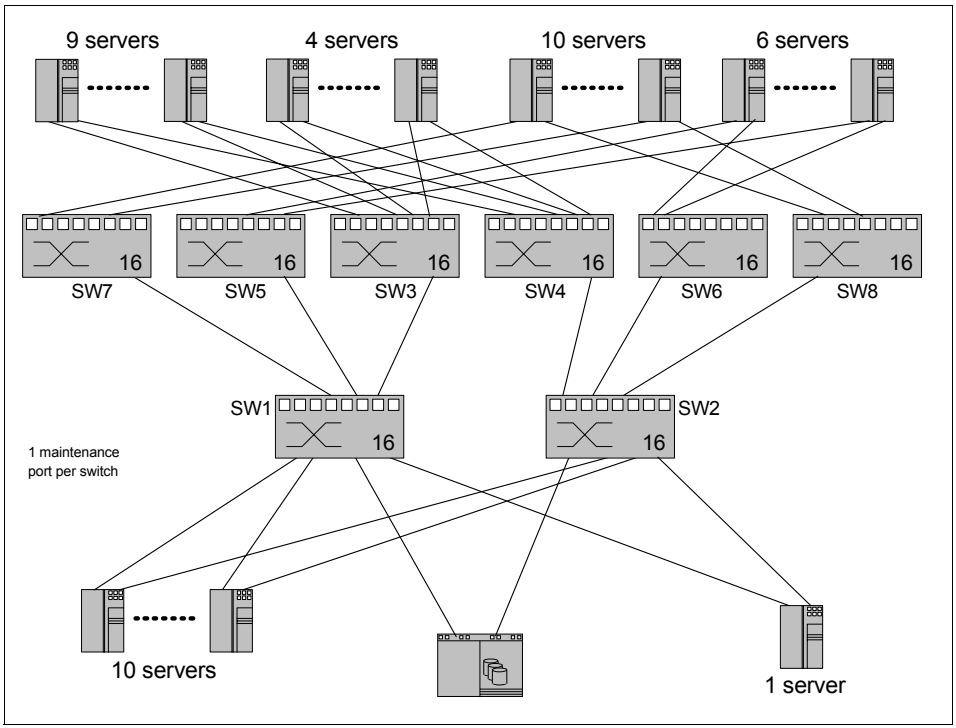


Figure 19-4 SAN design after three years of operation

In Table 19-3, we show the number of ports used for the solution.

Table 19-3 Number of used ports

Ports	Servers	Storage	Maintenance and expansion	ISLs	Spare
SW1	11	1	1	3	0
SW2	11	1	1	3	0
SW3	13	0	2	1	0
SW4	13	0	2	1	0
SW5	10	0	2	1	3
SW6	10	0	2	1	3
SW7	6	0	2	1	7
SW8	6	0	2	1	7
Total	40	2	14	12	20

As we can see, the design in Figure 19-4 is quite complex, because this has been growing over the years without interruption to the production systems. However, if it is acceptable to take the outage then the design can be changed, as shown in Figure 19-5, and this rearrangement will not cause any further expense for the purchase of new equipment.

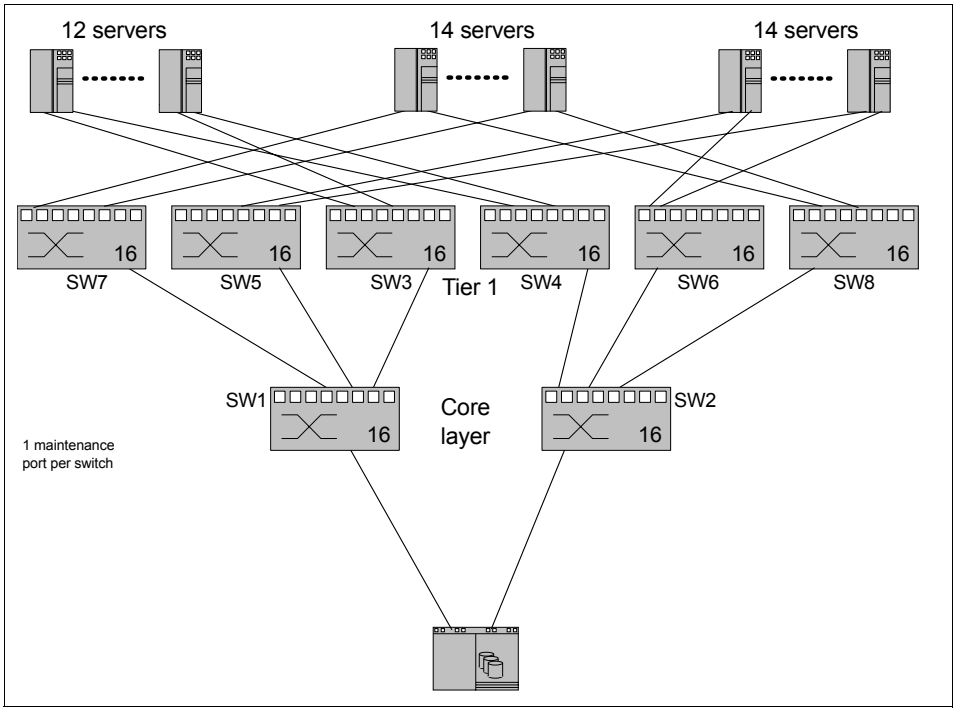


Figure 19-5 Core edge design

In Table 19-4 we show the number of ports used for the solution.

Table 19-4 Number of used ports

Ports	Servers	Storage	Maintenance and expansion	ISLs	Spare
SW1	0	1	1	3	11
SW2	0	1	1	3	11
SW3	12	0	1	1	2
SW4	12	0	1	1	2
SW5	14	0	1	1	0
SW6	14	0	1	1	0
SW7	14	0	1	1	0
SW8	14	0	1	1	0
Total	40	2	8	12	26

As you can see, we established a core-edge design approach. If you decide to re-cable before introducing the last two 16-port switches, you can replace them with 8-port switches and later use them as a core switches instead of SW1 and SW2, as shown in Figure 19-5.

In the following sections, we will outline some aspects of the design.

19.1.2 Performance

As we can see from the design we have an initial 10:1 fan-out ratio for servers accessing the storage device port. This means that 10 servers are accessing the same storage pool. As we only need 8.32 MB/s for the servers, therefore we have enough bandwidth on the storage port for all servers. By adding the next 10 servers we will get a fan-out ratio of 20:1. The bandwidth requirements will be 16.64 MB/s. Because of these requirements, one ISL is enough and one connection to the storage device will cover the performance needs. The one hop we use for ISL connections of additional switches will not greatly increase the latency, because those switches will be connected over a short distance.

Note: The rule of thumb is that the latency per 1 Km of fiber optic cable is 5 microseconds.

19.1.3 Availability

The redundant SAN design, with two paths from each server to the storage device, will give us the opportunity to perform maintenance on it, without downtime of the servers. We will also be able to perform upgrades or replacements without affecting the production servers.

19.1.4 Security

When implementing the switches, we need to accommodate some security related items:

- **Switch security**
  - Change the default passwords to access the switches.
  - Put switches in a separate management network if one is already in place for other functions.
- **Zoning**

In our case we have only one platform (Windows). Because we have only one storage port for all servers, there is no requirement to implement any zoning. If in the future we would expand the number of storage ports, we can group the servers by the storage ports and separate them into zones for performance reasons. This can be implemented non-disruptively using software zoning.

Note: There are no data integrity issues if you do not implement zoning. In our example, we would only use zoning for performance reasons.

19.1.5 Distance

As you can see from the designs, we are using a maximum of one hop between the switches. This means that there is no practical delay in the performance. For all the connections we will use shortwave SFPs as the servers are within a radius of 500 m (300 m for 2 Gbps SFP). It is possible to move some servers further away by using longwave SFPs for ISLs. Even with the fabric expansion we should have no problems with the delays in the fabric OS for name server and FSPF changes.

19.1.6 Scalability

Within our designs we have accommodated three years growth. The growth can be achieved without any interruptions in the production environment. If there is a need to grow after the predicted three years growth, we can switch the design to a core-edge approach, as shown in Figure 19-5 on page 722.

This will give us ample opportunity to grow as we can attach additional pairs of switches to core switches SW1 and SW2.

19.1.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: The clustering solution will failover to the passive server dynamically.
- HBA: If one of the HBA fails, multipath software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multipath software will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: Another redundant power supply may be added to the switch, and should one fails, other will take over automatically.

- Switch port: If one of the ports fails, you may replace using a hot-pluggable SFP, without outage or performance problem.
- Switch: If a switch fails, the server will use the alternate switch to connect to the storage, without outage or performance problem.

19.1.8 Manageability and management software

The switches have built in management capabilities:

- Telnet/serial interface to the OS with all the functions to manage the fabric
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use

To utilize these management features, the switches have to be connected to the network (Ethernet). It is possible to connect to only one switch and then manage all the others in the same fabric using inband management over Fibre Channel. This is not recommended, because if we lose an ISL or switch, we cannot manage the other switches.

We show an example of setting up the management network in Figure 19-6.

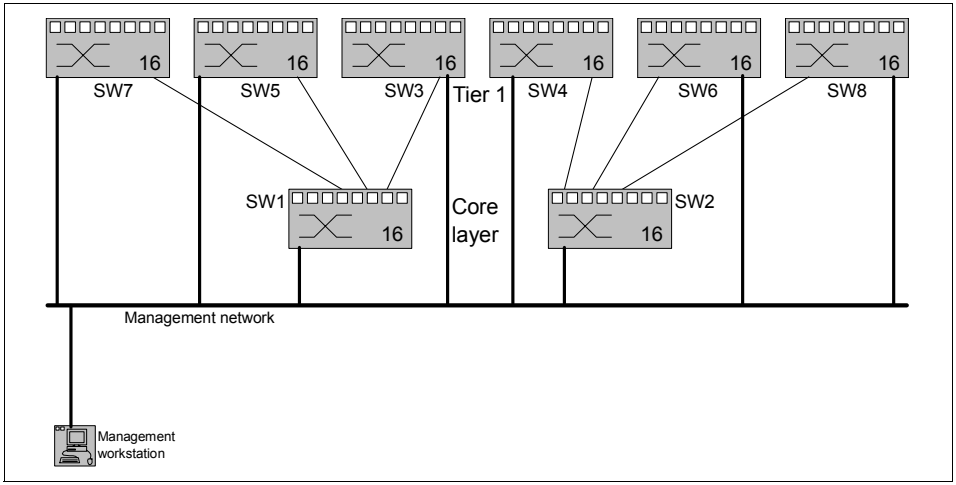


Figure 19-6 Management network

Note: This network can be part of your existing network infrastructure.

With growth, the complexity of the design grows. We recommend that you introduce some kind of enterprise integrated management software such as Tivoli Storage Network Manager.

19.1.9 Core switch design

For the core switch design, we decided to use two 24 port core switches. You can see the design in Figure 19-7.

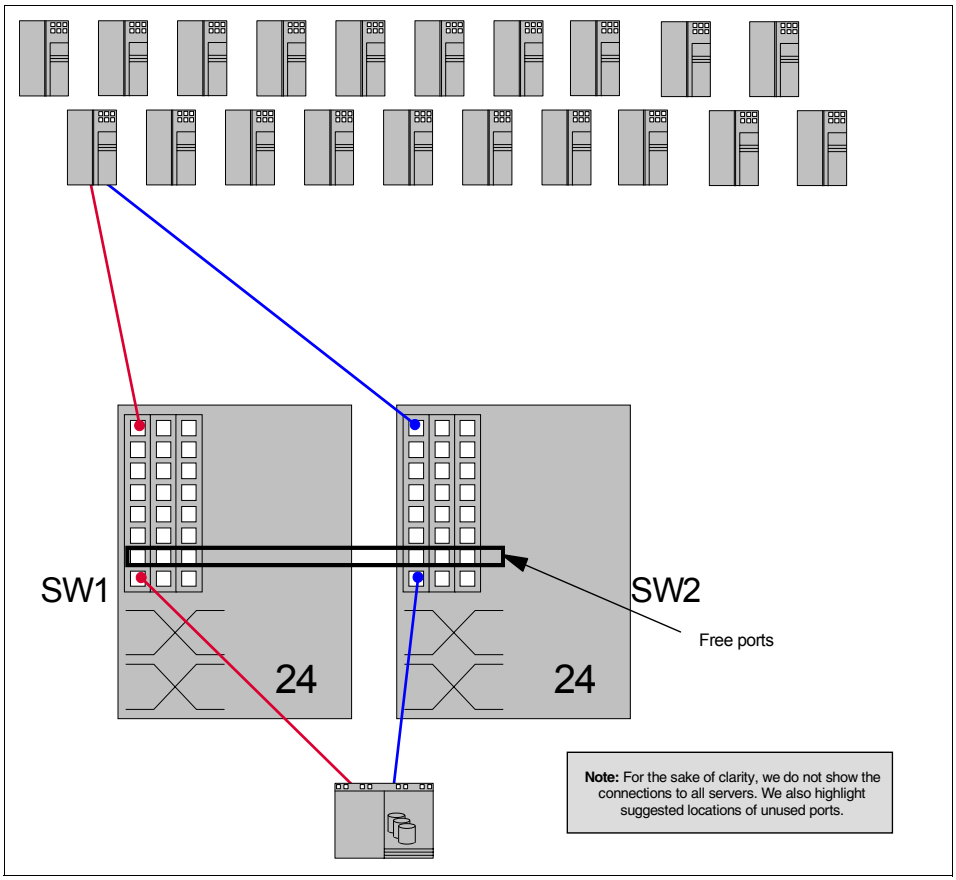


Figure 19-7 Core switch design

In Table 19-5, we show the number of ports used for the design.

Table 19-5 Number of used ports

Ports	Servers	Storage	Spare
SW1	10	1	13
SW2	10	1	13
Total	20	2	26

The product we selected can grow up to 64 ports in one box.

In Figure 19-7 you can see the design which accommodates 20 servers over a two year period. With such an approach, we satisfy all the requirements we have. Also, the characteristics of the design are the same as in the switch design.

The benefit of such a design, when compared to the switch design, is the ease of growth and management. For growth, we simply add more ports into the box, and there is no need for maintaining ISLs because all ports can communicate with full speed to each other. The reason why we used two core switches instead of one, even though they support non-disruptive upgrades and 99.999% availability, is to avoid a single point-of-failure caused by a single backplane, whether it is passive or not.

The ease of growth of such a design is shown in Figure 19-8.

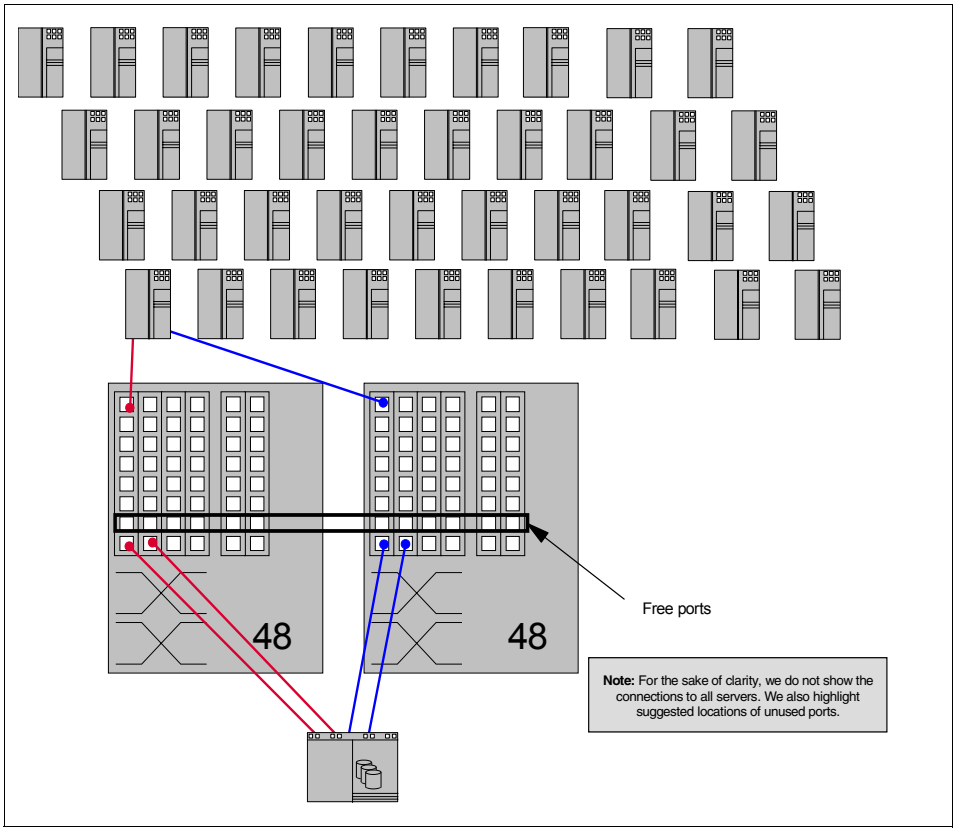


Figure 19-8 Expanding to 40 servers using core switch technology

In Table 19-6 we show the number of ports used for the design.

Table 19-6 Number of used ports

Ports	Servers	Storage	Spare
SW1	40	2	6
SW2	40	2	6
Total	80	4	12

As you can see from Figure 19-8, we simply added additional port blades inside the existing core switches. No need to allocate ISLs as in the switch design.

19.2 Case Study 2: Company Two

If we consider the company and its requirements as detailed in 18.2, “Case Study 2: Company Two” on page 689, we will propose the following designs.

19.2.1 Design

Considering the requirements as detailed in “Analysis (ports and throughput)” on page 687, we will propose two designs: In one design we use director class products, and in the other we use switches.

Switch design

In Figure 19-9 we show the initial design for the Getwell site.

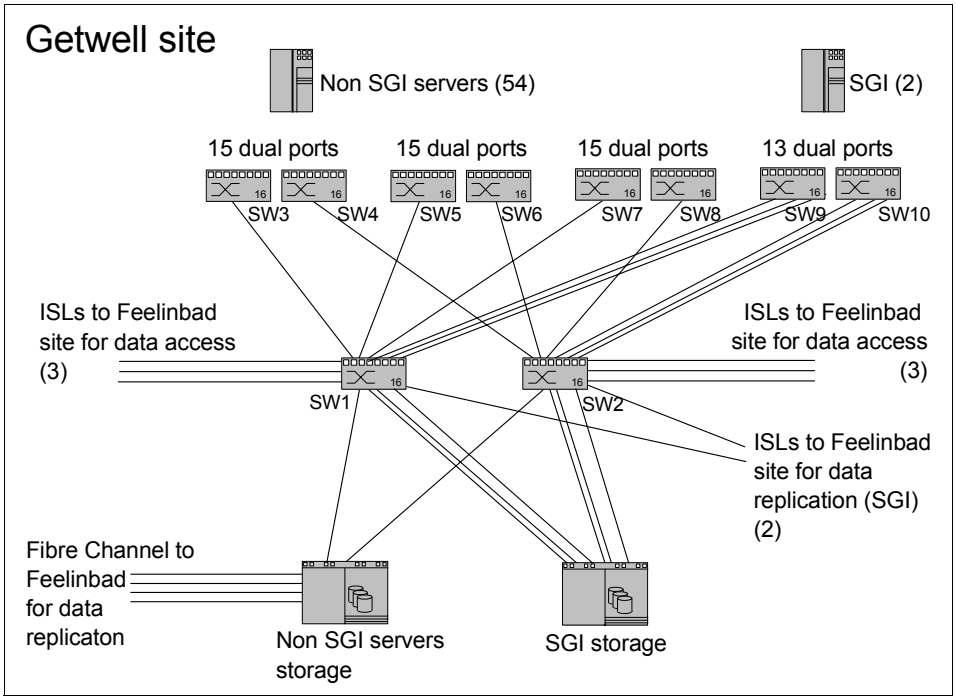


Figure 19-9 Getwell site initial design

In Table 19-7 we show the number of ports used for our solution.

Table 19-7 Number of used ports

Ports	Servers	Storage	ISLs	Spare
SW1	0	4	10	2

Ports	Servers	Storage	ISLs	Spare
SW2	0	4	10	2
SW3	15	0	1	0
SW4	15	0	1	0
SW5	15	0	1	0
SW6	15	0	1	0
SW7	15	0	1	0
SW8	15	0	1	0
SW9	11	0	3	2
SW10	11	0	3	2
Total	112	8	32	8

As you can see from the design we used a core—edge approach. We provided the following connections:

- Fifty-eight dual ports for the servers (56 needed)
- Six ISLs for connecting to Feelinbad site
- Two ISLs for SGI storage data replication
- Two ports for storage device for non-SGI servers
- Four ports for storage device for SGI servers
- Two ports for storage device for SGI data replication
- Fibre Channel connections to Feelinbad site for other servers data replication

The number of ISLs used between core and edge switches are:

- One ISL from each edge switch to core switches for non-SGI servers
- Three ISLs from the edge switch where SGI servers are connected to the core switches. One of the ISLs is used for non-SGI servers, and two for SGI servers.

The design for the Feelinbad Center is shown in Figure 19-10.

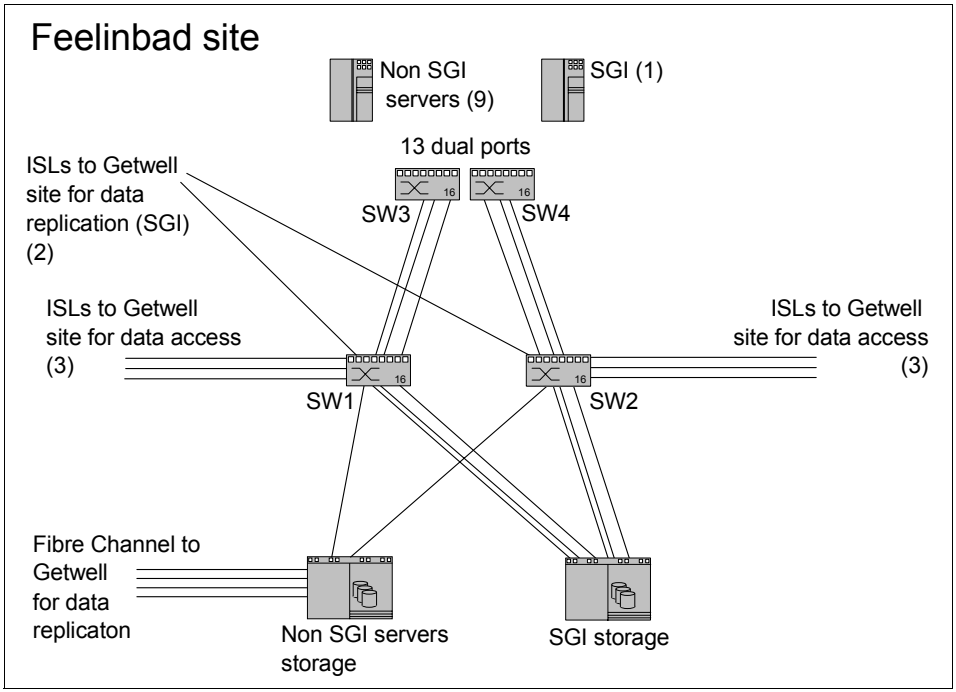


Figure 19-10 Feelinbad site initial design

In Table 19-8 we show the number of ports used for our solution.

Table 19-8 Number of used ports

Ports	Servers	Storage	ISLs	Spare
SW1	0	4	7	5
SW2	0	4	7	5
SW3	10	0	3	3
SW4	10	0	3	3
Total	20	8	20	16

As you can see from the design we used a core—edge approach. We provided the following connections:

- Thirteen dual ports for the servers (10 needed)
- Six ISLs for connecting to the Feelinbad site

- Two ISLs for SGI storage data replication
- Two ports for the storage device for non-SGI servers
- Four ports for the storage device for SGI servers
- Two ports for the storage device for SGI data replication
- Four Fibre Channel connections to Feelinbad site for non-SGI servers data replication

The number of ISLs used between core and edge switches are:

- Three ISLs from the edge servers where one of the ISLs is used for non-SGI servers, and two for SGI servers

With such a design we are fulfilling these requirements:

- No single point-of-failure — redundant SAN.
- All bandwidth requirements are met.
- SAN components can be upgraded without impact on the servers. In case of upgrade or maintenance of switches, we can first upgrade one switch, while paths across the second one are still available. After the traffic is established back on the upgraded switch, we can upgrade the second switch.
- Possible growth can be accommodated without impact on the production system. Because we are using a redundant SAN, we can introduce additional switches without the downtime on the existing servers.

We also provided enough ports for planned growth on both sides:

- Four SAN storage ports for SGI storage
- Four SAN server ports for SGI servers (2 dual server ports)
- Four SAN port for ISLs between core and edge switches for SGI servers

In the following sections we will outline some aspects of the design.

19.2.2 Performance

As we can see from the design we have an initial 54:1 fan-out ratio for all non-SGI servers accessing the storage device port. This means that 54 servers are accessing the same storage pool. As we need 75.3 MB/s for all the servers we have enough bandwidth on the storage port for all servers. The fan-out ratio for SGI servers is 1:1. The fan-out ratio will not change in the future, because we will only increase the number of ports on SGI servers. When we increase the number of ports on the SGI servers we will also increase the SGI storage ports by the same number. As you can see from the design, the number of ISLs covers all bandwidth requirements.

Note: With reference to the 54:1 fan-out ratio; if it is too high you can add an additional storage port for non-SGI servers. We will still have two free storage ports on the core switches. If you add more storage ports for non-SGI servers, you will lose expansion ports for SGI storage in the future. To solve this problem we would accommodate an additional pair of 8-port switches and use them for non-SGI traffic. This solution is shown in Figure 19-11.

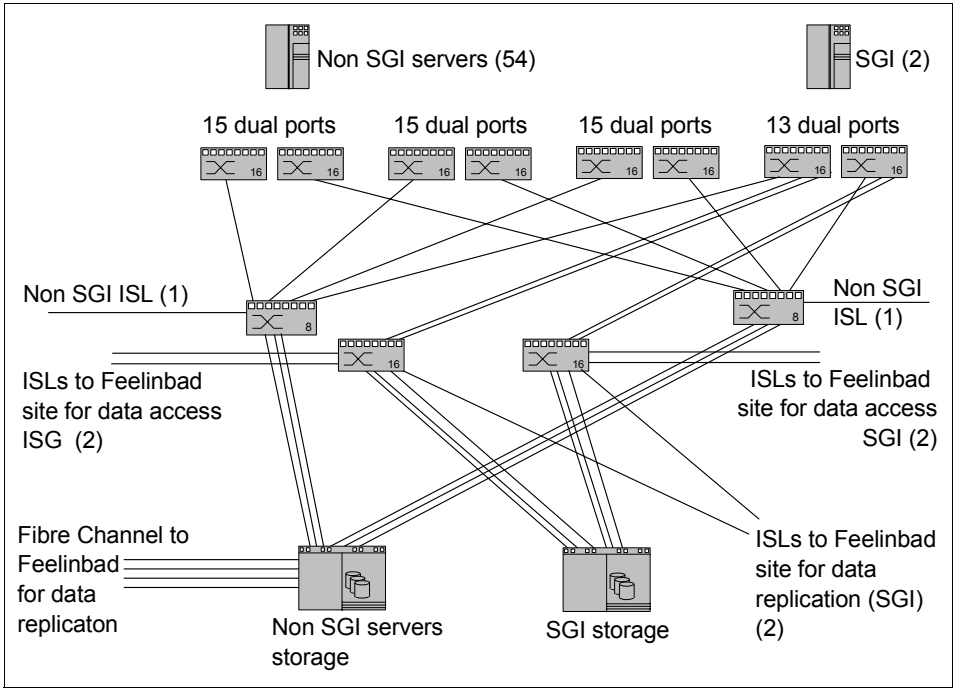


Figure 19-11 Adding additional storage ports for non-SGI servers

By adding two additional SAN storage ports on each core switch for non-SGI servers we will decrease the fan-out ratio to 18:1. Using two separate 8-port switches for data replication of SGI storage we freed two additional ports on the core switches for future SGI expansion. As you can see we have two core areas: One for SGI traffic, using 16-port switches, and one for the non-SGI server using 8-port switches.

The one hop we use for the ISL connection will not greatly increase the latency, because those switches will be put together over a short distance.

The latency between the sites will be around 208 microseconds. We will use synchronous copying of the storage data and 208 microseconds should not cause significant time delays for applications.

Note: The rule of thumb is that the latency per 1 Km of fibre cable is 5 microseconds.

Trunking

As you can see we are using multiple ISLs for the same traffic between switches to the remote site and we will use ISL trunking so that the ISLs appear as one link. With this function, traffic can actually be balanced across the ISLs. If trunking is not used, then multiple ISLs will present multiple paths to the destination.

As we know from the FSPF definition, only one path is used by one port and it could happen that traffic congestion occurs even if some of the ISLs are not in use.

In our example we have decided to use trunking on all multiple ISLs. In Figure 19-12 we show where we applied trunking.

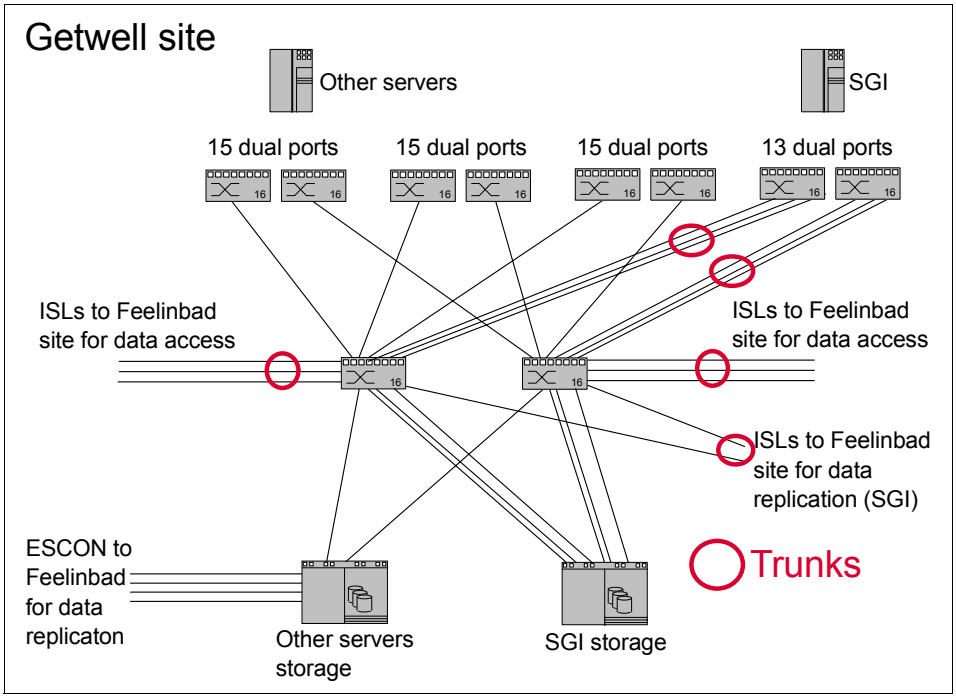


Figure 19-12 Trunking in Getwell Center

Trunking will give us better utilization of the ISLs.

19.2.3 Availability

The redundant SAN design, with two paths from each server to the storage device, will give us the opportunity to perform maintenance on it, without downtime of the servers. We will also be able to perform upgrades or replacements without affecting the production servers.

As we have two sites, our SAN is also designed to be redundant in connecting those two sites. We are providing redundant paths for data replication, and also for data access in a case that the storage device in the primary site fails.

19.2.4 Security

When implementing the switches, we need to accommodate some security related items:

- **Switch security**
  - Change the default passwords to access the switches.

- Put switches in separate management network if one is already in place for non-SGI functions.

- **Zoning**

  In our case, we have a multi-platform environment. Because we have only one storage port for all servers (except SGI) there is no need to implement any zoning. If in the future we expand the number of storage ports we can group the servers by the storage ports and separate them into zones for performance reasons. This can be implemented non-disruptively using software zoning.

  We will have the following zones:

  - Zone for all SGI servers in both sites. This will be implemented using software zoning.
  - Zone for SGI data replication, which will include SGI storage ports for data replication from both sites and ISLs used. This zone will be implemented using hard zoning. In this zone we will include all the participating ports including the ports for ISLs. With this we will assure that data replication traffic is completely separated from other traffic.

  Note: There are no data integrity issues if you do not implement zoning. In our example, we would only use zoning for performance reasons.

19.2.5 Distance

As you can see from the designs we are using a maximum of one hop between the switches in each site. This means that there is no practical delay in the performance. For all the local connections, we will use shortwave SFPs as the servers are in a radius of 500 m (300 m for 2 Gbps SFP). It is possible to move some servers further away by using longwave SFPs for ISLs. Even with the fabric expansion, as our designs show, we should have no problems with the delays in the fabric OS for name server and FSPF changes.

For connections from site to site, we will use longwave SFPs which will be connected over DWDM products. With the use of a DWDM product we will solve the degradation of laser signal, but we still need to adjust the buffers needed on switch ports for this type of connection. For this, we need to enable the Extended Fabrics feature in the switches, which in fact assigns more buffer credits to those E_Ports being used by long distance ISLs.

19.2.6 Scalability

Within our designs, we have accommodated three years growth. For the predicted growth we will only accommodate new ports for SGI servers and storage. . If in the future we need more ports, the design is ready to be expanded with new switches.

19.2.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: The clustering solution will failover to the passive server dynamically.
- HBA: If one of the HBA fails, multipath software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multipath software will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: Another redundant power supply may be added to the switch, and should one fails, other will take over automatically.
- Switch port: If one of the ports fails, you may replace using a hot-pluggable SFP, without outage or performance problem.
- Switch: If a switch fails, the server will use the alternate switch to connect to the storage, without outage or performance problem.

19.2.8 Manageability and management software

In addition to what has been discussed in the previous design, the only important factor here is that because we have a rather large number of switches, it is recommended to use some kind of management software which supports topology management.

Core switch design

We show the core switch design for the Getwell site in Figure 19-13.

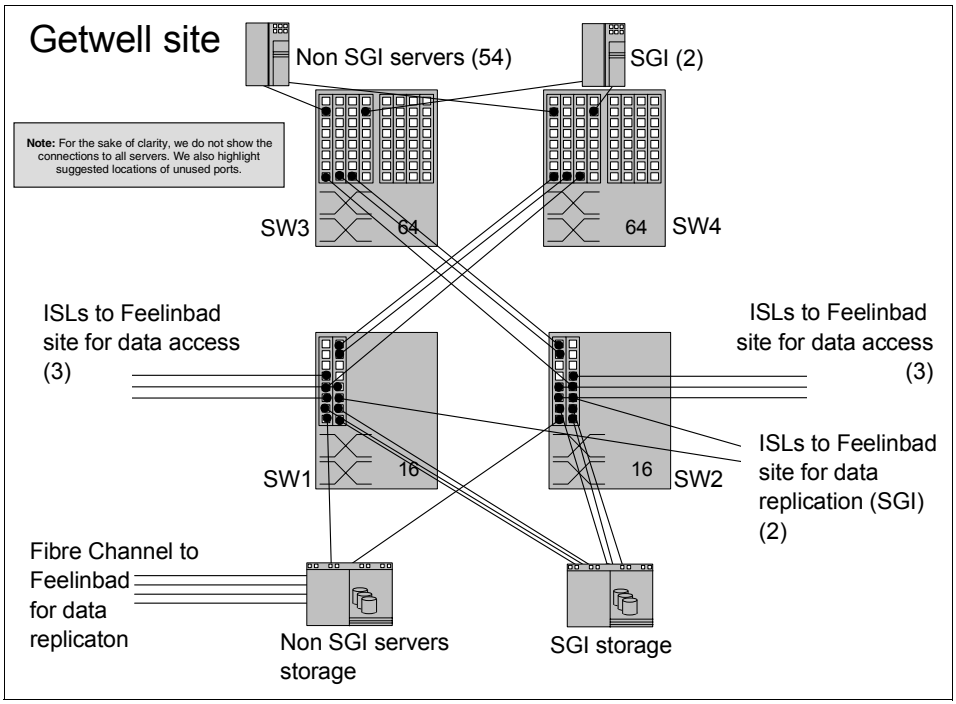


Figure 19-13 Core switch design for Getwell site

In Table 19-9 we show the number of ports used for solution.

Table 19-9 Number of used ports

Ports	Servers	Storage	ISLs	Spare
SW1	0	4	7	5
SW2	0	4	7	5
SW3	56	0	3	5
SW4	56	0	3	5
Total	112	8	20	20

We used a core-edge design with four switches to accommodate the requirements. On the bottom core switches we have accommodated all the storage connections and ISLs to the Feelinbad site. We used three ISLs to the edge switches. Two of them were used for SGI traffic and one of them for non-SGI traffic.

The design for Feelinbad is shown in Figure 19-14.

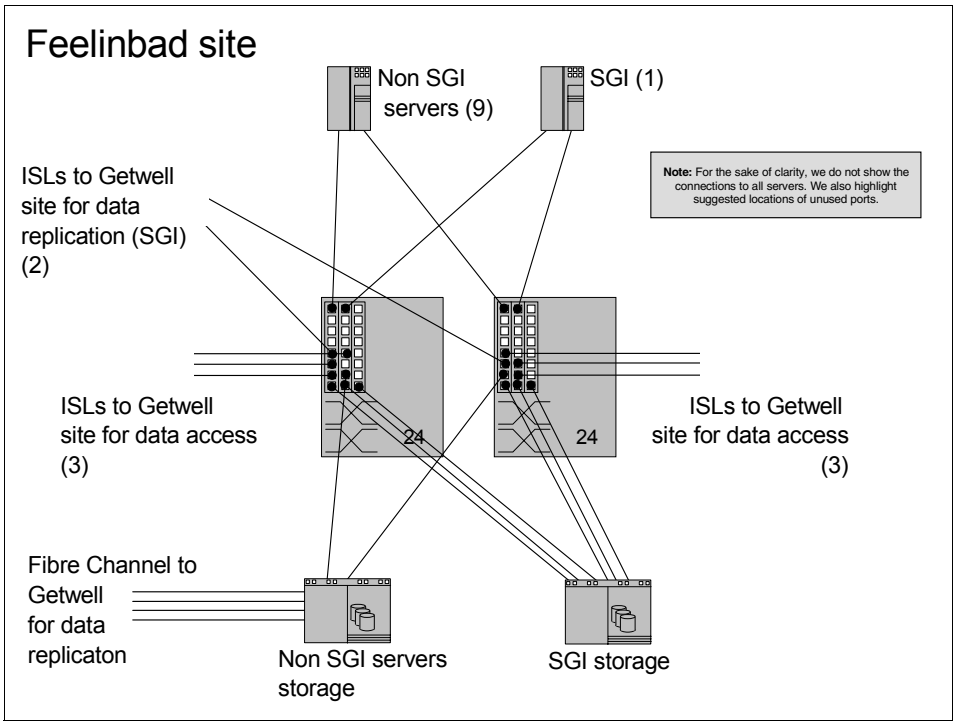


Figure 19-14 Core switch design for Feelinbad site

In Table 19-10 we show the number of ports used for our solution.

Table 19-10 Number of used ports

Ports	Servers	Storage	ISLs	Spare
SW1	10	4	4	6
SW2	10	4	4	6
Total	20	8	8	12

For the Feelinbad site we used two core switches with 24 ports in each switch. The total used ports are 18 per switch.

With such a design we are fulfilling all the requirements and we are easing the possible expansion of it in the future, because we have a lot of space for additional ports on both sites.

We could also introduce more than one SAN storage port in the Getwell Center to reduce the fan-out ratio from 54:1 to 18:1 as we did in Figure 19-11 on page 733. With this we would use an additional two ports in each core switch.

19.3 Case Study 3: Company Three

If we consider the company and its requirements as detailed in 18.3, “Case Study 3: Company Three” on page 697, we will propose this design.

19.3.1 Design

In considering the requirements as detailed in 18.3.5, “Analysis (ports and throughput)” on page 699, we decided to use two 16-port Fibre Channel switches for our design.

We show the proposed design in Figure 19-15.

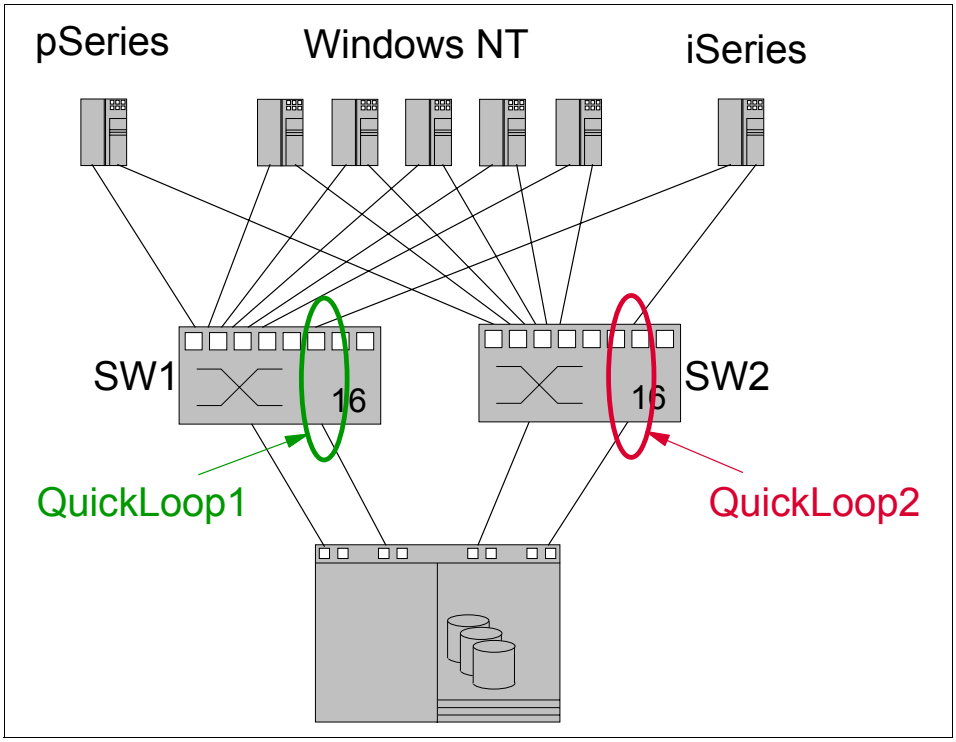


Figure 19-15 Proposed design for Company Three

In Table 19-11 we show the number of ports used for our solution.

Table 19-11 Number of used ports

Ports	Servers	Storage	ISLs	Spare
SW1	7	4	0	5
SW2	7	4	0	5
Total	14	8	0	10

With such a design we are fulfilling the requirements:

- No single point-of-failure — redundant SAN.
- All bandwidth requirements are met.
- SAN components can be upgraded without impact on the servers. In case of upgrade or maintenance of switches, we can first upgrade one switch, while paths across the second one are still available. After the traffic is established back on the upgraded switch, we can upgrade the second switch.

- Possible growth without impact on production. Because we are using a redundant SAN, we can introduce additional switches without downtime on the existing servers.
- We have provided the option to group the iSeries related ports and isolate them from other traffic as this is a requirement for this platform.

In the following sections, we outline some aspects of the design.

19.3.2 Performance

In the initial design, we have a 6:1 fan-out ratio for all servers except for iSeries. This means that six servers are accessing one storage port. The fan-out ratio for iSeries is 1:1. There are no great latency issues, because the traffic is only going through the switches, where the latency is no greater than two microseconds.

QuickLoop

For the iSeries platform, we need to provide private FC-AL communication between the iSeries server and his storage port. This means that we need to give the iSeries server the impression that it is on the private loop with its storage port.

The other requirement is that only one storage port can be on that loop. QuickLoop is a feature of the IBM 2109 family of switches (except the Hxx and Mxx models) and also the managed hub. This feature emulates a private loop on up to two switches connected together in the fabric. You can have only one QuickLoop per switch.

19.3.3 Availability

The redundant SAN design, with two paths from each server to the storage device, will give us the opportunity to perform maintenance of it, without downtime of the servers. We will also be able to perform upgrades or replacements without affecting the production system.

We are also providing dual paths for iSeries. Failover of the traffic will be managed by iSeries operating system (OS/400). IBM TotalStorage Switch solution supports OS/400 V5R2 or later.

19.3.4 Security

When implementing the switches, we need to consider some security matters:

- **Switch security**
  - Change the default passwords to access the switches.
  - Put switches in separate management network if one is already in place for other functions.
- **Zoning**

  In our example, we have the requirement to separate iSeries traffic from other servers. We will introduce the following zones on both switches:
  - Zone for iSeries server and its storage port
  - Zone for all other servers and their storage port

19.3.5 Distance

As you can see from the designs, we are using only one switch between the servers and storage. For all the connections, we use shortwave SFPs.

19.3.6 Scalability

Within the designs we accommodated three years growth. We have enough bandwidth allocated for all requirements in the next three years. The only expansion will probably be adding additional storage ports, and we have accommodated capacity for this. If in the future, we will need more ports the design is ready to be expanded with new switches.

19.3.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: Application will not be available. Server has to be replaced.
- HBA: If one of the HBA fails, multipath software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multipath software will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: Another redundant power supply may be added to the switch, and should one fail, the other will take over automatically.
- Switch port: If one of the ports fails, you may replace using a hot-pluggable SFP, without outage or performance problem.

- Switch: If a switch fails, the server will use the alternate switch to connect to the storage, without outage or performance problem.

Note: HBA failover of iSeries will be managed by its operating system.

19.3.8 Manageability and management software

The management techniques are described in 19.1.8, “Manageability and management software” on page 725.

19.4 Case Study 4 - Company Four

If we consider the company and its requirements as detailed in Figure 18-4, “Case Study 4: Server schematic” on page 702, we will propose the following solution.

19.4.1 Design

Considering the requirements as detailed in 18.4.5, “Analysis (ports and throughput)” on page 703, we have proposed in this section a switch based solution for our design.

Both the open systems and zSeries platforms will share the same storage device which will provide a much higher availability.

As you can see from our proposed design in Figure 19-16, we have replaced the ESCON director with IBM TotalStorage SAN Switch M14 for both sites. IBM TotalStorage SAN Switch M14 is designed to support FICON and Fibre Channel intermix which is best suited for our scenario with two different platforms. This director-class switch shown has been configured with thirty-two ports initially, a full configuration can support up to 128 ports.

.

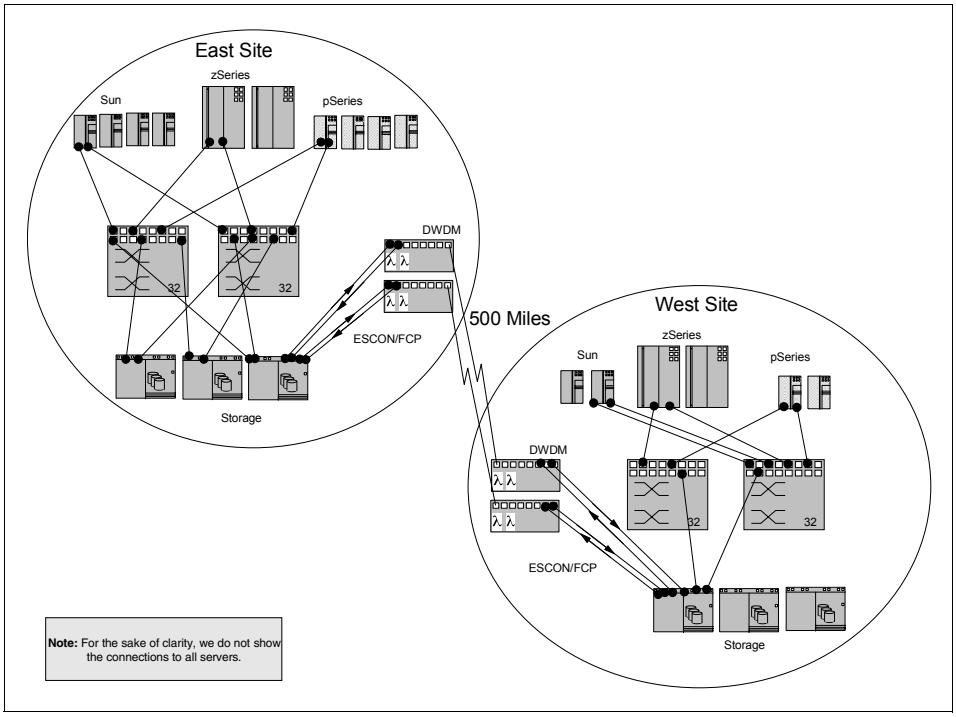


Figure 19-16 Initial design using IBM TotalStorage SAN Switch M14

As we are using redundant paths from each server and to the storage we are providing high availability. To utilize the redundant connections we will use multipathing software (such as IBM's Subsystem Device Driver) on each of the open systems servers, while for our mainframe servers we have z/OS built in failover support for path redundancy. The solution shown has been configured with two 32-ports initially. With this design, we are partially fulfilling the requirements:

- Redundant SAN with two switches and all servers dual connected.
- All bandwidth requirements are met.
- By incorporating some of the Director class functionality, firmware upgrades and maintenance can be achieved non-disruptively. That is, there is no requirement to redirect all traffic through one switch while the other is being upgraded.
- Future growth by adding more switches can be accomplished without service interruption.

In the following sections, we outline some of the aspects of this solution.

For storage replication we are using ESCON links over DWDM. Depending on the implementation time frame, we may be able to substitute these with FCP using PPRC V2. The peak bandwidth used by the servers is 11.58 MB/s. As only 30% of these are writes then we need to accommodate 3.48 MB/s. Two pairs of ESCON channels will accomplish this and provide redundancy. The same DWDM devices used for open systems can be used for the zSeries requirement.

In Table 19-12 we show the ports at the East site.

Table 19-12 Ports: East site

Storage	Server	ISL	Spare
10	20	0	34

In Table 19-13 we show the ports at the West site.

Table 19-13 Ports: West Site

Storage	Server	ISL	Spare
10	12	0	42

19.4.2 Performance

In the design, we have a fan-out ratio of 20:6 for all servers in the East site. The fan-out ratio in the West site is 12:6. Latency through the switches is negligible (~2 micro seconds).

The latency between the sites will be around 4 milliseconds.

19.4.3 Availability

This SAN design, with two paths from each server to the storage device, will give us the ability for each server to experience an HBA failure without impacting performance and without downtime of the servers.

IBM TotalStorage SAN Switch M14 supports functionalities such as non-disruptive firmware upgrades and maintenance. That is, there is no requirement to redirect all traffic through one director while the other is being upgraded.

19.4.4 Security

When implementing the directors we need to take care about some security related things:

- **Director security**

- Change the default passwords to access the switches.
- Put directors in a separate management network if one is already in place for other functions.

- **Zoning**

We will be implementing zoning so we can group the servers by storage ports (FICON/FCP) and separate then into zones.

19.4.5 Distance

In all the connections, we will use shortwave SFPs as the servers are within a radius of 500 m (300 m for 2 Gbps SFP). It is possible to move some servers further away by using longwave SFPs for ISLs.

19.4.6 Scalability

The design accommodates three years growth. The growth can be achieved without any interruptions in the production environment.

19.4.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: The clustering solution will failover to the passive server dynamically.
- HBA (open systems): If one of the HBA fails, multi-path software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- HBA (zSeries): If one HBA fails, z/OS will handle failover.
- Cable: If a cable between a server and the switch fails, multi-path software or z/OS will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: IBM TotalStorage SAN Switch M14 solution has dual power supplies as a standard feature and no disruption to service will occur should one fail. The replacement power supply can be installed with no outage incurred.
- Director port: If one of the ports fails, you may replace the hot-pluggable optic. No outage will occur as the server will use its alternate path.
- Director: Should the director fail, the server will use its alternate path through another switch without an outage or performance degradation.
- DWDM device: Should the DWDM device fail, the alternate one will be used.

19.4.8 Manageability and management software

The management techniques are described in 19.1.8, “Manageability and management software” on page 725.

19.5 Case Study 5: Company Five

If we consider the company and its requirements as detailed in 18.5, “Case Study 5: Company Five” on page 705, we will propose the following design.

19.5.1 Design

In considering the requirements as detailed in “Analysis (ports and throughput)” on page 708, we have decided to use two 8-port Fibre Channel switches for our design.

We show the proposed design in Figure 19-17.

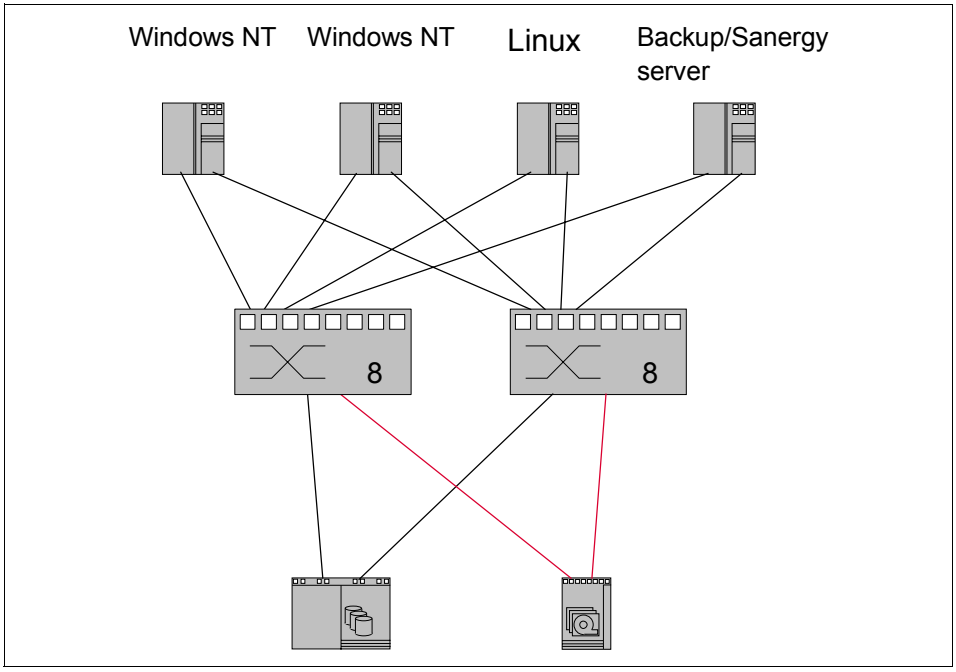


Figure 19-17 Proposed design for Company Five

In Table 19-14 we show the number of ports used for our solution.

Table 19-14 Number of used ports

Ports	Servers	Storage + Tape	ISLs	Spare
SW1	4	4	0	0
SW2	4	4	0	0
Total	8	8	0	0

With such a design we are fulfilling these requirements:

- No single point-of-failure — redundant SAN.
- All bandwidth requirements are met.
- SAN components can be upgraded without impact on the servers. In the case of upgrade or maintenance of the switches we can first upgrade one switch, while paths across the second one are still available. After the traffic is established back on the upgraded switch, we can upgrade the second switch.
- Possible growth without impact on the production. Because we are using a redundant SAN we can introduce additional switches without downtime on the existing servers.
- Because the switches we have used support FC-AL connectivity we can connect tape devices directly to them.

Note: We chose 8-port switches as a low cost solution. If we need to allow for future expansion, 16-port switches can be substituted.

In the following sections, we will outline some aspects of the design.

19.5.2 Performance

In the initial design we have a 4:1 fan-out ratio for all servers accessing the storage. This means that four servers are accessing one storage port. All servers can also access the tape device, which was the requirement for LAN-free backup implementation.

Attention: Host software (such as Tivoli Storage Manager) must be present to allow functions that enable tape sharing and/or tape library sharing.

There are no latency issues, because the traffic is only going through the switches, where the latency is up to 2 microseconds.

19.5.3 Availability

The redundant SAN design, with two paths from each server to the storage device, will give us the opportunity to perform maintenance of it without downtime of the servers. We will also be able to perform upgrades or replacements without affecting the production system.

19.5.4 Security

When implementing the switches we need to consider security:

- **Switch security**
  - Change the default passwords to access the switches.
  - Put switches in separate management network if one is already in place for other functions.

- **Zoning**

  Because we have only one storage port for all servers there is no requirement to implement any zoning. If in the future we would expand the number of storage ports we can group the servers and the storage ports into separate zones for performance reasons. This can be implemented non-disruptively using software zoning.

  Note: There are no data integrity issues if you do not implement zoning. In our example we would only use zoning for performance reasons.

19.5.5 Distance

As you can see from the designs, we are using only one switch between the servers and storage. All the connections we will use are shortwave SFPs.

19.5.6 Scalability

Within the design, we have accommodated three years growth. We have enough bandwidth allocated for all requirements in the next three years. The only expansion will probably be adding additional storage ports, and we have accommodated space for this. If in the future we would need more ports, the design is ready to be expanded with new switches.

19.5.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: Application will not be available. Server has to be replaced.
- HBA: If one of the HBA fails, multipath software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multipath software will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: Another redundant power supply may be added to the switch, and should one fail, the other will take over automatically.
- Switch port: If one of the ports fails, you may replace using a hot-pluggable SFP, without outage or performance problem.
- Switch: If a switch fails, the server will use the alternate switch to connect to the storage, without outage or performance problem.

19.5.8 Manageability and management software

The management techniques are the same as in Case Study 1, and are described in 19.1.8, “Manageability and management software” on page 725.

19.6 Case Study 6: Company Six

In considering the company and its requirements as detailed in 18.6, “Case Study 6: Company Six” on page 710, we will propose this design.

19.6.1 Design

Taking into account the requirements as detailed in 18.6.5, “Analysis (ports and throughput)” on page 712, we have decided to use two 16-port Fibre Channel switches for our design. You can see the proposed design for the Primary site in Figure 19-18.

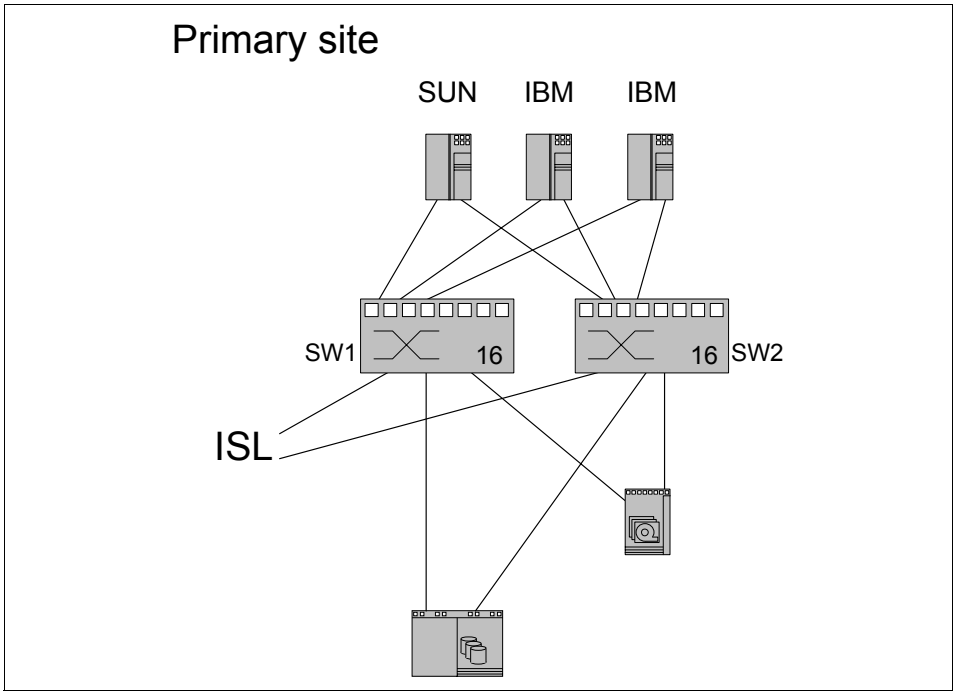


Figure 19-18 Proposed design for Primary site

In Table 19-15 we show the number of ports used for our solution.

Table 19-15 Number of used ports

Ports	Servers	Storage + Tape	ISLs	Spare
SW1	3	4	1	8
SW2	3	4	1	8
Total	6	8	2	16

In the secondary site, we already have a SAN, using Compaq StorageWorks SAN Switch 16. These are Brocade OEM switches of the SilkWorm 2800.

We show the design of the Secondary site in Figure 19-19.

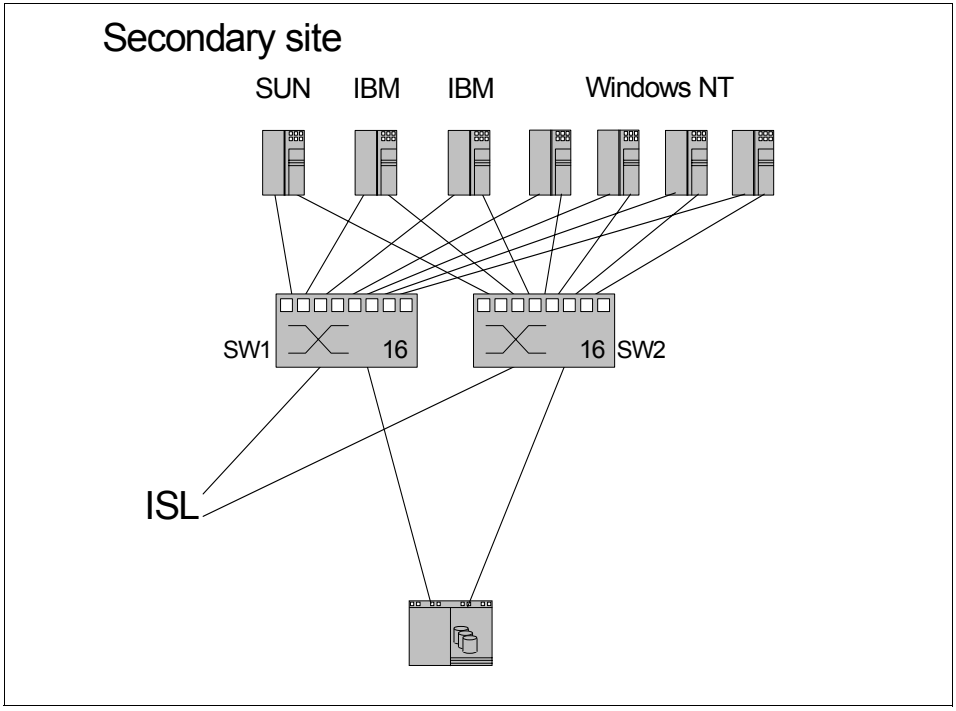


Figure 19-19 Proposed design for Secondary site

In Table 19-16 we show the number of ports used for our solution.

Table 19-16 Number of used ports

Ports	Servers	Storage + Tape	ISLs	Spare
SW1	7	2	1	6
SW2	7	2	1	6
Total	14	4	2	12

Both sites are connected with two ISLs. Because the distance is 900 miles, we will use DWDM products to connect them.

We show the design in Figure 19-20.

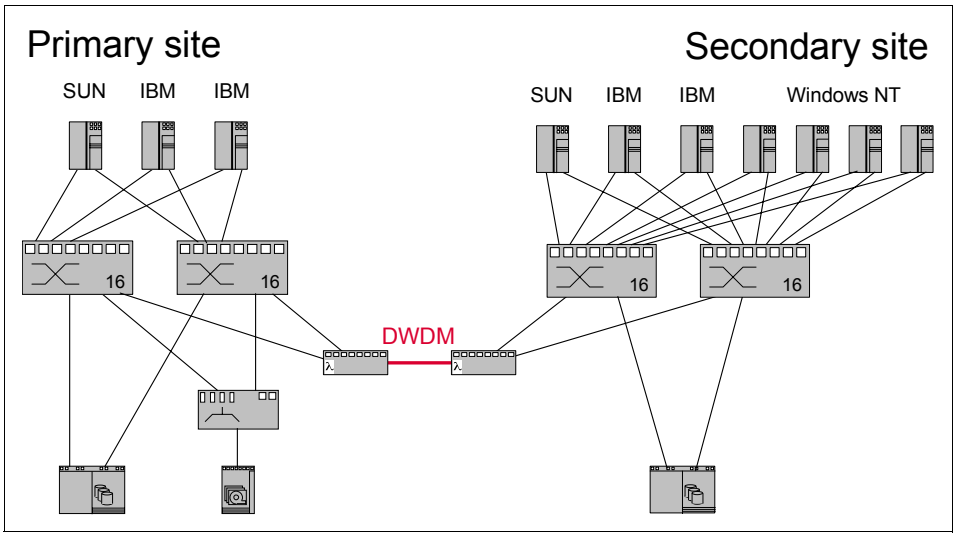


Figure 19-20 DWDM connection between sites

Note: Be sure that the firmware of the IBM switches matches firmware of Compaq switches used in the secondary site. There is no need to replace them, because they are OEM'd from the same manufacturer.

With such a design we are fulfilling these requirements:

- No single point-of-failure — redundant SAN.
- All bandwidth requirements are met.
- SAN components can be upgraded without impact on the servers. In the case of upgrade or maintenance of switches, we can first upgrade one switch, while

paths across the second one are still available. After the traffic is established back on the upgraded switch, we can upgrade the second switch.

- Possible growth without impact on the production. Because we are using a redundant SAN we can introduce additional switches without downtime on the existing servers.
- We are reusing existing equipment (switches and tape).
- We are providing storage replication to remote site for disaster recovery.
- We are improving backup performance by providing an infrastructure for LAN-free backup.

Attention: Host software (such as Tivoli Storage Manager) must be present to allow functions that enable tape sharing and/or tape library sharing.

In the following sections, we outline some aspects of the design.

19.6.2 Performance

In the initial design we have a 3:1 fan-out ratio for all servers accessing the storage on the Primary site. This means that three servers are accessing one storage port. All servers can also access the tape device, which was the requirement for the LAN-free backup implementation. There are no latency issues, because the traffic is only going through the switches, where the latency is up to 2 microseconds. On the Secondary site we have a 7:1 fan-out ratio. As with the Primary site, we do not have any latency issues.

19.6.3 Availability

The redundant SAN design, with two paths from each server to the storage device, will give us the opportunity to perform maintenance of it, without downtime of the servers. We will also be able to perform upgrades or replacements without affecting the production.

As we have two sites, our SAN is designed to be redundant also in connecting those two sites. We are providing redundant paths for data replication. It is recommended that the connections between two DWDM boxes are also redundant.

19.6.4 Security

When implementing the switches, we need to take into account security:

- **Switch security**

- Change the default passwords to access the switches.
- Put switches in separate management network if one is already in place for other functions.

- **Zoning**

  In our case we have a heterogeneous platform accessing the storage. We will require to group platforms including servers accessing tape library for greater performance.

19.6.5 Distance

As you can see from the designs we are only using one switch between the server and the storage in each site. This means that there is no practical delay in the performance. For all the local connections, we will use shortwave SFPs.

For connections from site-to-site, we use Longwave SFPs which will be connected over DWDM products. With the use of a DWDM product we will solve degradation of the laser signal, but we still need to adjust the buffers needed on switch ports for this type of connection. For this we need to enable the Extended Fabrics feature in the switches, which in fact assigns more buffer credits to those E_Ports being used by long distance ISLs.

19.6.6 Scalability

Within the designs we have accommodated three years growth. We have enough bandwidth allocated for all our requirements in the next three years. The only expansion will probably be adding additional storage ports, and we have accommodated space for this. If in the future we need more ports, the design is ready to be expanded with new switches.

19.6.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: Application will not be available. Server has to be replaced.
- HBA: If one of the HBA fails, multipath software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multipath software will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: Another redundant power supply may be added to the switch, and should one fail, the other will take over automatically.

- Switch port: If one of the ports fails, you may replace using a hot-pluggable SFP, without outage or performance problem.
- Switch: If a switch fails, the server will use the alternate switch to connect to the storage, without outage or performance problem.

19.6.8 Manageability and management software

The management techniques are the same as in Case Study 1, and we describe them in 19.1.8, “Manageability and management software” on page 725.

You will also need to consider to integrate the management of DWDM products into your management infrastructure.

20

IBM TotalStorage SAN m-type family solutions

In this chapter, we will show solutions to the case studies based on the products in the IBM McDATA portfolio.

© Copyright IBM Corp. 2004. All rights reserved.

20.1 Case Study 1: Company One

If we consider the company and its requirements as detailed in 18.1, “Case Study 1: Company One” on page 686, we will propose the following solution.

20.1.1 Design using Directors

Considering the requirements as detailed in 18.1.5, “Analysis (ports and throughput)” on page 687, we have proposed in this section a Director-based solution for our design.

We show the proposed design in Figure 20-1 using a McDATA ED-6064 with 28 ports (a fully configured McDATA ED-6064 has 64 ports).

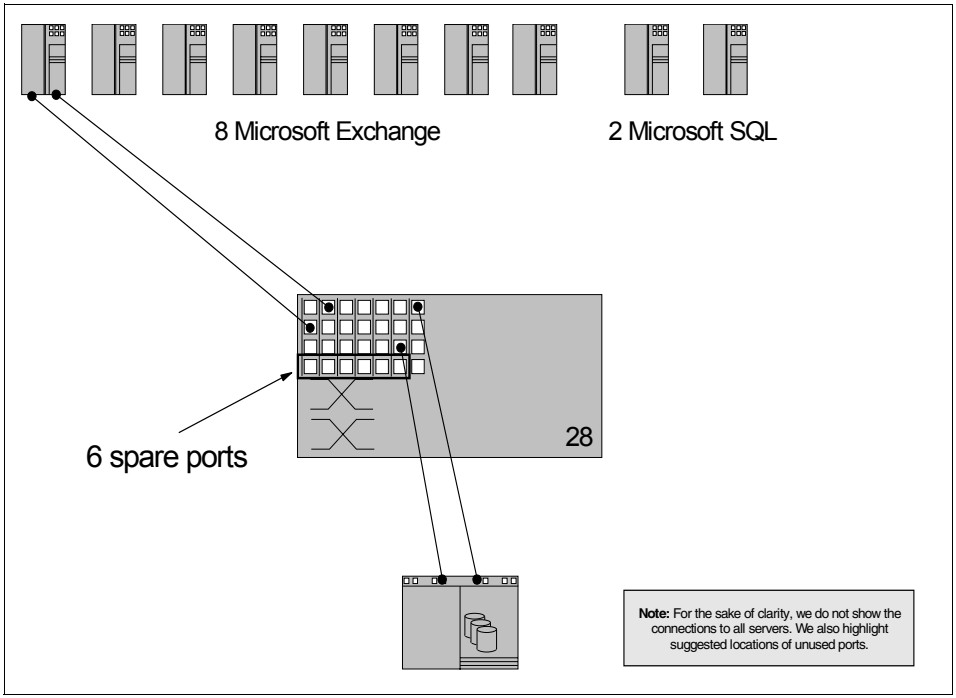


Figure 20-1 Core SAN design using a McDATA ED-6064 Director

As we are using two paths from each of the ten servers to the storage, we are providing redundancy and high availability. To utilize the dual connections, we will use multipathing software (such as IBM’s Subsystem Device Driver- SDD) on each of the servers. The solution shown has been configured with 28 ports in the Director initially, additional ports can be added in groups of four. With such a design, we are partially fulfilling the requirements:

- Fault resilient SAN with 99.999% availability.
- All bandwidth requirements are met (40KB/s and 4MB/s from servers and 8.32MB/s to storage).
- The Director can be upgraded (firmware and additional port cards) without impact on the servers. There is no requirement to bring the Director down for firmware upgrades, applications will continue to perform I/O during this operation, but may experience a minimal delay. No re-routing and/or cabling changes need to be made to accommodate this.
- Possible growth without impact on the production. Due to the nature of the Director, we can add additional 4 port cards (blades) concurrently. We can also introduce additional Director(s) for a fully redundant SAN.

With this design we have six SAN ports per Director free for future expansion. This means that you could potentially connect an additional three servers with redundant connections.

Care should be taken to ensure that primary and secondary connections from one server are not attached to the same blade. If a blade does fail, the alternate path should be available through another blade.

Although the solution presented provides 99.999% availability and is fault tolerant, it is not fully redundant as there is a single passive backplane. To overcome this we can introduce another Director. This will give us the capability of expanding to 128 usable ports in two devices. During the re-cabling process, be sure to identify the secondary paths (cables) from each server and re-connect to the second Director. During this process, there will be single points of failure for each server.

In the following pictures we will show how you can expand your fabric without impact on the production servers. As this was the design requirement, we will show that our design is capable of handling this. In the second year, we should expect to accommodate 20 servers, with the same I/O characteristics.

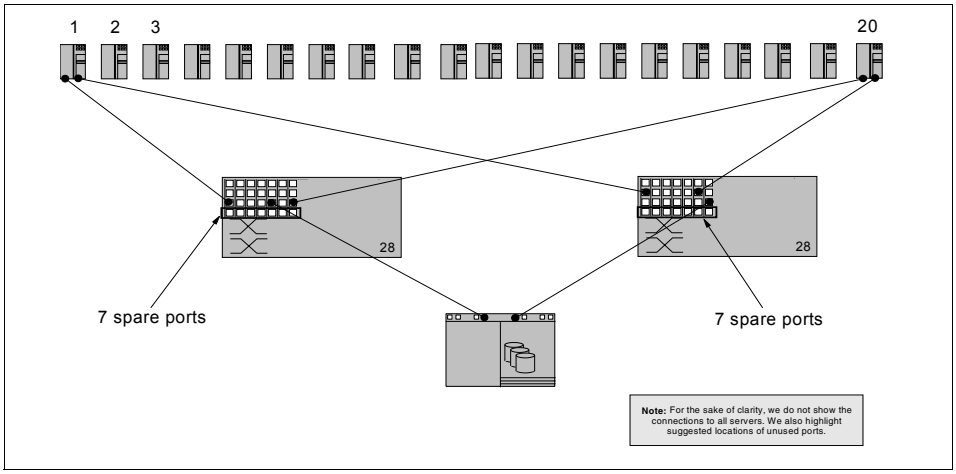


Figure 20-2 Fully redundant McDATA ED-6064 Director solution

Each blade has one free port, this will accommodate the failure of a complete blade, and also allow for a port to be used for maintenance. It also allows for a possible expansion of seven additional servers.

With the uncertainty surrounding the growth of the complex into the third year, should the additional 20 servers come to fruition, then the total number of ports required could increase to 84. The solution proposed above will accommodate this with the simple non-disruptive approach of adding port blades, as shown Figure 20-3.

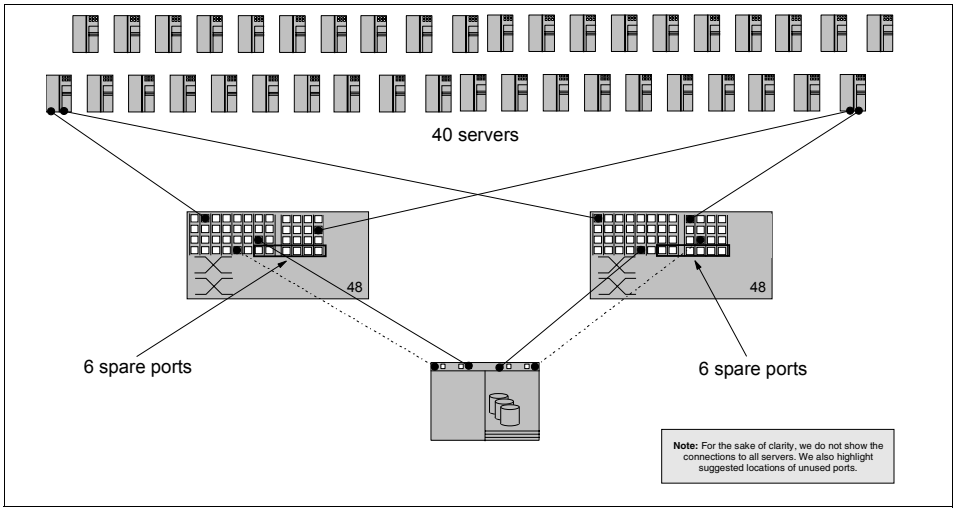


Figure 20-3 McDATA ED-6064 solution with all potential servers

Spare ports are still available for maintenance functions. It is also worth noting that the port blade cards can be moved from one Director to another, should you wish to use one Director initially, populate it, and then introduce the second Director. Additional port expansion capabilities still exist with a maximum port count of 64 per Director.

Note: We have doubled the number of connections to the storage device to reduce the fan-out ratio to 20:1. This is not due to any bandwidth requirement, but solely due to the number of connections that must be handled.

We assume the same bandwidth requirements for the new servers as the original servers.

In Table 20-1 we show Year 1 requirements.

Table 20-1 Year 1 ports: 10 servers

Storage	Server	ISL	Spare
2	20	0	6

In Table 20-2 we show Year 2 requirements.

Table 20-2 Year 2 ports: 20 servers

Storage	Server	ISL	Spare
2	40	0	14

In Table 20-3 we show Year 3 requirements.

Table 20-3 Year 3 ports: 40 servers

Storage	Server	ISL	Spare
4	80	0	12

In the following sections, we will outline some aspects of the design.

20.1.2 Performance

As we can see from the design, we have an initial fan-out ratio of 10:1 for each server accessing a storage device port. This means that 10 servers are accessing the same storage pool. As we only need 8.32 MB/s for the servers we have enough bandwidth on the storage port for all servers. By adding the next 10 servers, we will get a fan-out ratio of 20:1. The bandwidth requirements will be 16.64 MB/s. The same ratio will be maintained if we double the number of servers by doubling the number of storage ports used.

20.1.3 Availability

The McDATA ED-6064 single Director solution provides 99.999% availability. This number will be effectively somewhat higher in a dual Director configuration. This SAN design, with two paths from each server to the storage device, will give us the ability for each server to experience an HBA failure without impacting performance and without downtime of the servers.

As we are using a Director solution, we will also be able to perform upgrades (hardware and software) or replacements concurrently, that is without affecting the production servers. This is the case, whether we use one Director or two.

20.1.4 Security

When implementing the Directors we need to take into account some security matters:

- **Director security**
  - Change the default passwords to access the Director(s).
  - Put Director(s) in a separate management network if one is already in place for other functions.
- **Zoning**

  In our case, we have only one platform (Windows NT/2000/2003). Because we have only one storage port for all servers there is no need to implement any zoning. If in the future we would expand the number of storage ports, we can group the servers by the storage ports separately, then into zones for performance reasons.

  Note: There are no data integrity issues if you do not implement zoning. In our example we would only use zoning for performance reasons.

20.1.5 Distance

In all the connections, we will use shortwave SFPs as the servers are within a radius of 300 m.

20.1.6 Scalability

The designs accommodate three years growth. The growth can be achieved without any interruptions in the production environment. If there is additional growth, we can simply add more port and blades until we hit the maximum of 128 ports of the two Directors, at which point we can introduce a third Director.

20.1.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: The clustering solution will failover to the passive server dynamically.
- HBA: If one of the HBA fails, multi-path software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the Director fails, multi-path software will automatically failover workload to the alternate path. If a cable between the Director and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: Director class solutions have dual power supplies as a standard feature and no disruption to service will occur should one fail. The replacement power supply can be installed with no outage incurred.
- Director port: If one of the ports fails, you may replace using a hot-pluggable SFP.
- Director blade: If a Director port blade fails, the alternate path will be used. The port on the failed blade can be dynamically plugged into the spare ports in other blades until the blade can be replaced.
- Director: The Director is a 99.999% type solution, it is therefore improbable that it will fail. Should it fail there will be no impact if a dual Director solution has been implemented.

20.1.8 Manageability and management software

The Directors have built-in management capabilities:

- Serial interface to the OS, typically only used to configure the network address
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use

To fully utilize the management features, the Directors have to be connected to the network (Ethernet). Although inband management is supported, in our solution we are not using ISLs, and therefore, inband management is not an option.

We show an example of setting up the management network in Figure 20-4.

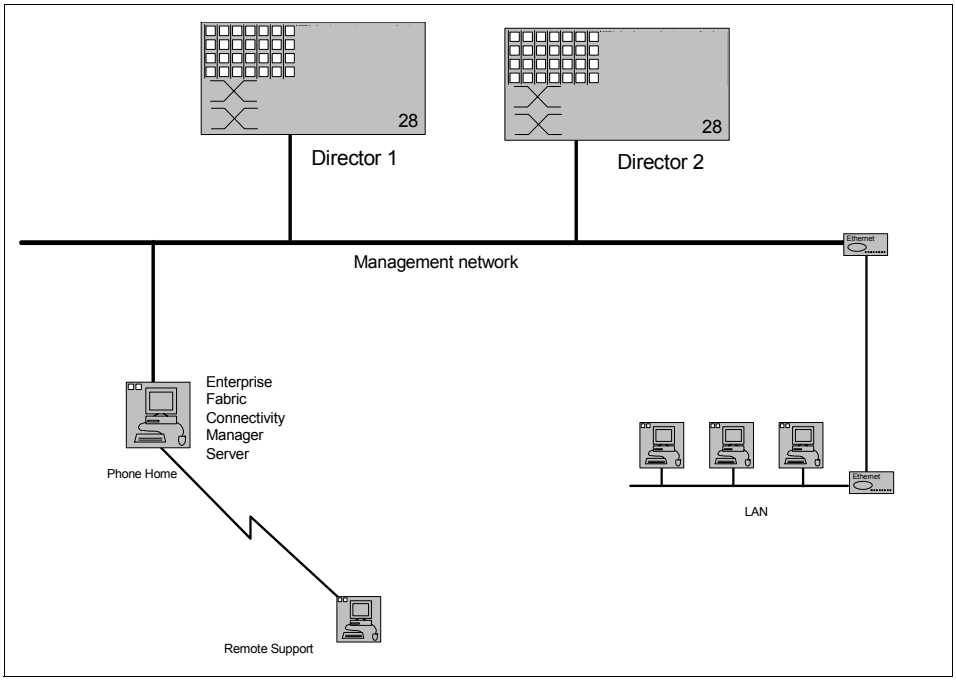


Figure 20-4 Management network

Note: This network can be part of your existing network infrastructure.

Full functional management is available through EFC Manager software.

20.1.9 Design using switches

Considering the requirements as detailed in 18.1.5, “Analysis (ports and throughput)” on page 687, we have proposed in this section a switch based solution for our design.

We show the proposed design in Figure 20-5.

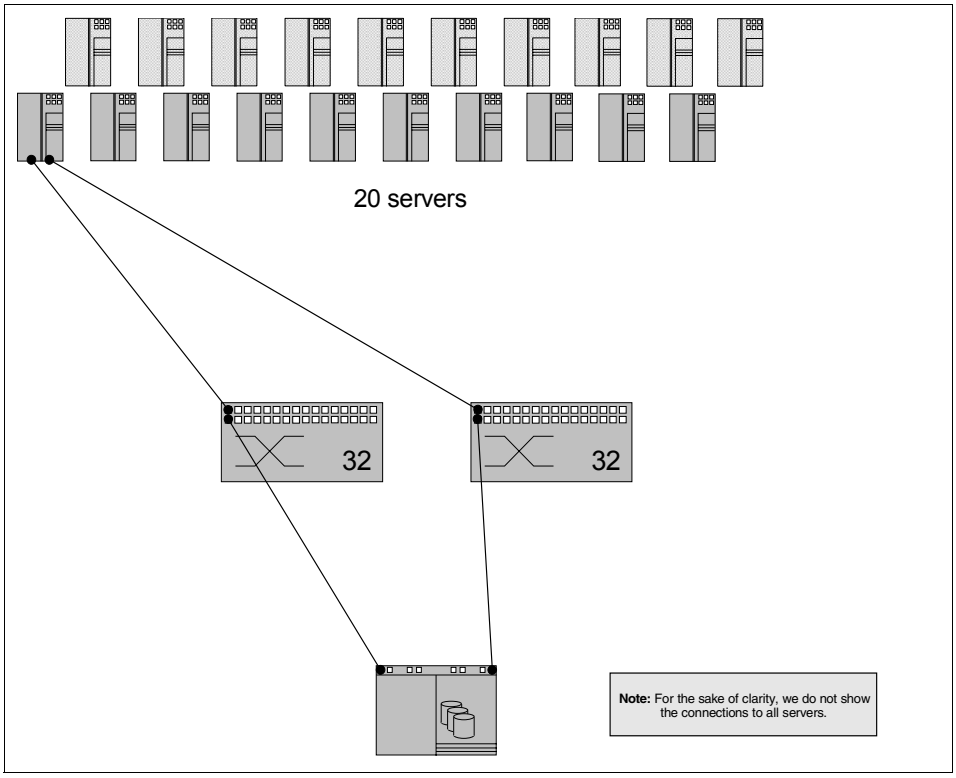


Figure 20-5 Initial design using McDATA 2027-232 Switches

As we are using two paths from each server to the storage, we are providing redundancy and high availability. To utilize the dual connections, we will use multipathing software (such as IBM's Subsystem Device Driver) on each of the servers. The solution shown has been configured with two 32-port switches. With this design we are fulfilling these requirements:

- Redundant SAN with two switches and all servers dual connected
- All bandwidth requirements are met (40 KB/s and 4 MB/s from servers and 8.32 MB/s to storage).
- With McDATA switches, firmware upgrades and maintenance can be achieved non-disruptively. That is, there is no requirement to redirect all traffic through one switch while the other is being upgraded.
- Future growth by adding more switches can be accomplished without service interruption.

With this design, we have eleven SAN ports free for future expansion. This means that you could potentially connect an additional eleven servers with redundant connections.

To expand the fabric to accommodate all 40 servers we must introduce another 32-port switch. This will give us a total of 96 ports, of which three will be used for connectivity to storage and 80 will be used for servers (dual connected).

Special consideration needs to take place when introducing the third switch to avoid bottlenecks and evenly distribute the connections from the servers to the switches. The outcome required should represent 40 servers dual connections spread across the three switches, this equates to 27:27:26 as the distribution.

As you can see, this design will only be truly balanced if the number of servers is divisible by three. In this case study, the performance requirements are quite low and we do not expect there to be an issue. However, as the number of servers increases, the relative fan-out ratio should negate any ill effect this may have on performance.

As we left the two switch implementation there were 20 connections per switch. Introducing the third switch will accommodate all 20 primary connections for the new servers. The new servers will need a total of 20 secondary connections across the original two switches which can be achieved as we are only using 21 of the 32-ports per switch (that is (32-20-1)*2=22).

However, this will leave the used port configuration as 30:30:20 which will leave an unbalanced switch-storage link utilization.

To resolve this we need to take three secondary links from each of the original switches/servers to the new switch. Although the cables will need to be physically unplugged and then plugged, no outage should occur as we are using multi-path software that can use the existing primary path. It may be wise to perform the migration one cable at a time.

The solution is highlighted in Figure 20-6.

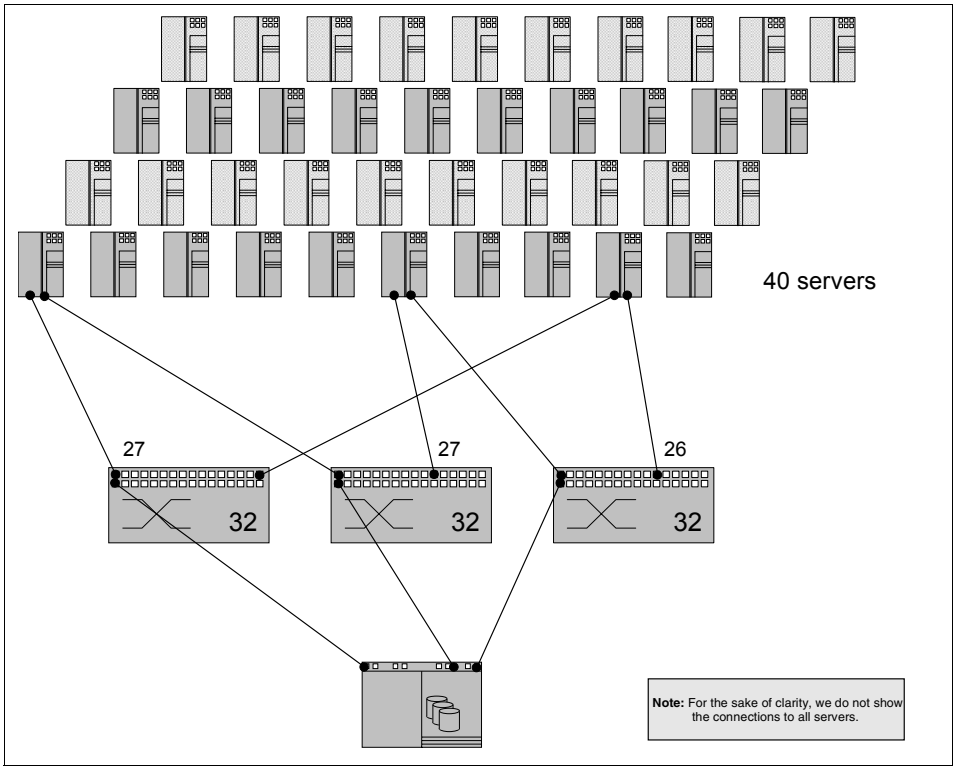


Figure 20-6 Final design to accommodate all potential servers

Important: Ensure that only secondary connections are identified and unplugged. If a primary and secondary link from the same server are removed, then that server will no longer be able to perform I/O across the SAN.

The process described above uses the least effort to achieve the desired results, other considerations and preferences for switch locations and server groups could require a more complex re-cabling effort. The use of a fiber patch panel may aid with cable relocation.

The final configuration, assuming all potential servers are added to the SAN will leave a total of 13 ports free for expansion.

In Table 20-4 we show Year 1 ports.

Table 20-4 Year 1 ports: 10 servers

Storage	Server	ISL	Spare
2	20	0	42

In Table 20-5 we show Year 2 ports.

Table 20-5 Year 2 ports: 20 servers

Storage	Server	ISL	Spare
2	40	0	22

In Table 20-6 we show Year 3 ports.

Table 20-6 Year 3 ports: 40 servers

Storage	Server	ISL	Spare
3	80	0	13

We also took these aspects into account.

20.1.10 Performance

As we can see from the design we have an initial fan-out ratio of 10:1 for each server accessing a storage device port. This means that 10 servers are accessing the same storage pool. As we only need 8.32 MB/s for the servers we have enough bandwidth on the storage port for all servers. By adding the next 10 servers we will get a fan-out ratio of 20:1. The bandwidth requirements will be 16.64 MB/s. As we introduce the third switch and servers the fan-out ratio changes somewhat. As we mentioned earlier, the end result will mean that we have fan-out ratios of 26:1 and 27:1, depending on which server is connected to which switch. Only when the number of servers is exactly divisible by three will we achieve equal fan-out ratios. In this scenario, we do not believe this to be of concern as the I/O throughput requirements are low.

20.1.11 Availability

This SAN design, with two paths from each server to the storage device, will give us the ability for each server to experience an HBA failure without impacting performance, and without downtime of the servers.

The McDATA 2027-232 switch incorporates some of the Director class functionality; firmware upgrades and maintenance can be achieved non-disruptively. That is, there is no requirement to redirect all traffic through one switch, while the other is being upgraded.

It should be noted that when introducing the third switch and performing the re-cabling functions, there will temporarily be single points of failure for six servers as the secondary link is migrated to the new switch.

20.1.12 Security

When implementing the switches we need to take into account security matters:

- **Switch security**
  - Change the default passwords to access the switches.
  - Put switches in a separate management network if one is already in place for other functions.
- **Zoning**

  In our case, we have only one platform (Windows NT/2000/2003). Because we have only one storage port for all servers there is no need to implement any zoning.

20.1.13 Distance

In all the connections, we use shortwave optics as the servers are within a radius of 300 m. It is possible to move some servers further away by using longwave optics for extending the fabric.

20.1.14 Scalability

The designs accommodate three years growth. The growth can be achieved without any interruptions in the production environment. Six additional servers with dual connections can be added to the existing environment. If there are additional growth requirements, we can introduce a fourth switch. At this time, we would recommend that we re-design the cable connections to ensure even distribution.

20.1.15 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: The clustering solution will failover to the passive server dynamically.
- HBA: If one of the HBA fails, multi-path software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multi-path software will automatically failover workload to the alternate path. If a cable between

the switch and the disk storage fails, an alternate route will be used. There is no performance loss.

- Power supply: The McDATA 2027-232 solution has dual power supplies as a standard feature and no disruption to service will occur should one fail. The replacement power supply can be installed with no outage incurred.
- Switch port: If one of the ports fails, you may replace the hot-pluggable optic. No outage will occur as the server will use its alternate path.
- Switch: Should the switch fail, the server will use its alternate path through another switch without an outage or performance degradation.

20.1.16 Manageability and management software

The switches have built-in management capabilities:

- Serial interface to the OS, typically only used to configure the network address
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use

To fully utilize the management features, the switches have to be connected to the network (Ethernet). Although inband management is supported, in our solution, we are not using ISLs, and therefore, inband management is not an option.

We show an example of setting up the management network in Figure 20-7.

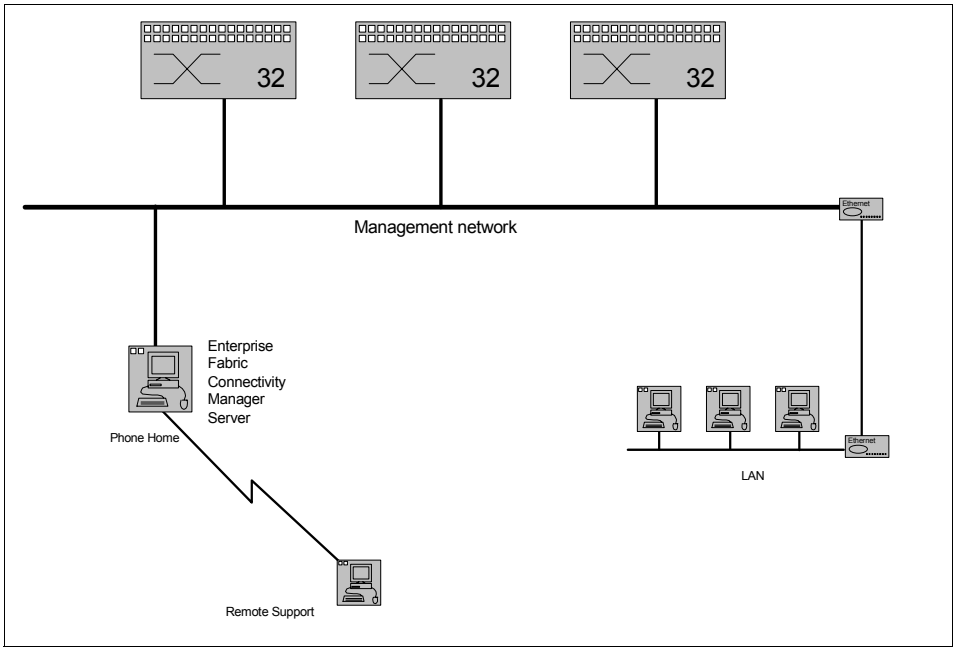


Figure 20-7 Management network

Note: This network can be part of your existing network infrastructure.

Full functional management is available through EFC Manager software.

20.2 Case Study 2: Company Two

If we consider the company and its requirements, as detailed in 18.2, “Case Study 2: Company Two” on page 689, we will propose the following solution.

20.2.1 Design

Considering the requirements as detailed in 18.2.5, “Analysis (ports and throughput)” on page 692, we have proposed in this section a Director based solution for our design.

The configuration for the Getwell facility can be seen in the proposed design in Figure 20-8. In this solution, we are using two McDATA ED-6064s which have been configured with 28 ports each initially (additional ports can be added in groups of four) and four McDATA 2027-232s. The McDATA ED-6064s can have a maximum of 64-ports.

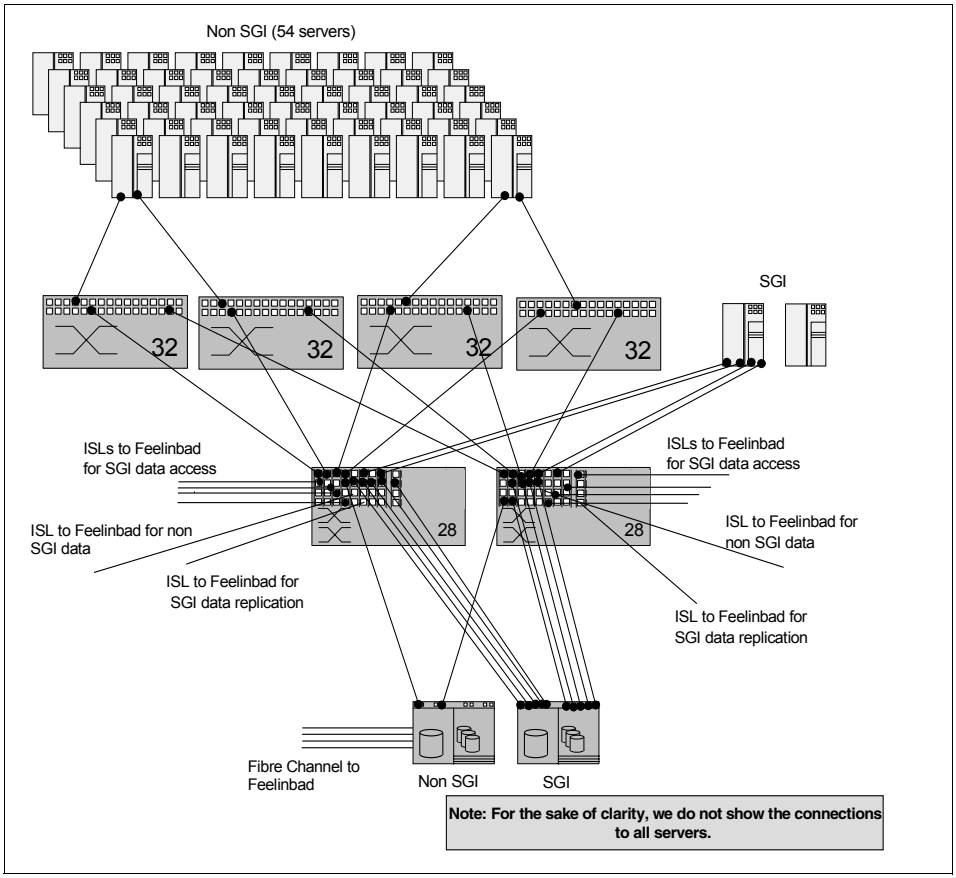


Figure 20-8 Getwell SAN design using McDATA ED-6064 Director and 2027-232 switches

To summarize the port allocation:

- One hundred and eight edge ports for non-SGI servers (dual connect)
- Two ports for non-SGI Storage
- Eight ports for SGI Servers (quad connect)
- Eight ports for SGI Storage
- Two ports for SGI replication (storage - switch)
- Four Fibre Channel connections from non-SGI storage device to Feelinbad
- Sixteen ports for core-edge (2 x 8) connectivity for non-SGI servers.

- Eight ports for ISLs for connecting to Feelinbad
- Two ports for ISLs for SGI replication
- Two ports for ISLs for non-SGI

Which makes for a total of 156 Fibre Channel connections.

Table 20-7 Initial ports

Storage	Server	ISL	Spare
12	116	28	28

The configuration for the Feelinbad facility can be seen in the proposed design in Figure 20-9

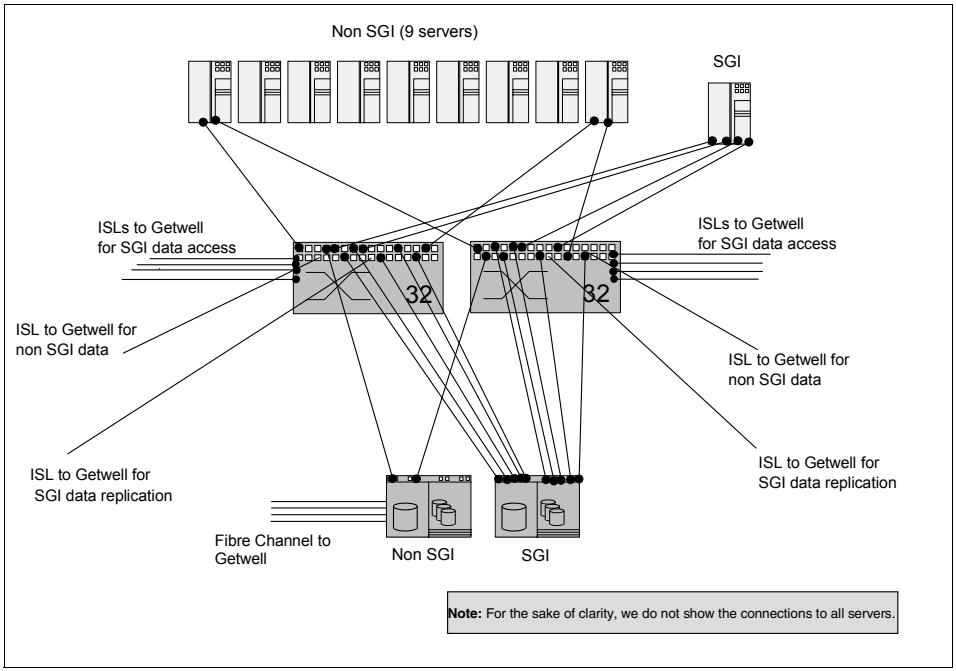


Figure 20-9 Feelinbad SAN design using McDATA ES-3032 switches

To summarize the port allocation:

- Eighteen ports for non-SGI servers (dual connect)
- Two port for non-SGI storage
- Four ports for SGI servers
- Eight ports for SGI storage
- Two ports for SGI replication (storage-switch)
- Four fibre channel connections from the storage device to Getwell
- Eight ports for ISLs for connecting to Getwell

- Two ports for ISLs for SGI replication
- Two ports for ISLs for non-SGI

Which makes for a total of 46 Fibre Channel connections.

Table 20-8 Initial ports

Storage	Server	ISL	Spare
12	22	12	18

With such a design we are fulfilling the requirements:

- No single point of failure, redundant SAN
- All bandwidth requirements are met.
- SAN components can be upgraded without impact on the servers. In case of upgrade or maintenance of switches, we can first upgrade one switch, while paths across the second one are still available. After the traffic is established back on the upgraded switch, we can upgrade the second switch.
- Possible growth without impact on the production servers. Because we are using a redundant SAN, we can introduce additional switches without downtime on the existing servers.

We also provided enough ports for the planned growth on both sides:

- Four SAN storage ports for SGI storage
- Four SAN server ports for SGI servers (Two dual server ports)
- Four SAN ports for ISLs between for SGI servers

In the following sections, we outline some aspects of the design.

20.2.2 Performance

As we can see from the design, we have an initial 54:1 fan-out ratio for all non-SGI servers accessing storage device port. This means that 54 servers are accessing the same storage pool. As we need 75.3 MB/s for all the servers we have enough bandwidth on the storage port for all servers. The fan-out ratio for SGI servers is 1:1. The fan-out ratio will not change in the future, because we will only increase the number of ports on SGI servers. When we increase the number of ports on the SGI servers, we also increase the SGI storage ports by the same number. As we can see from the design, the number of ISLs covers all bandwidth requirements.

Note: Refer to your storage device information to determine if a 54:1 fan-out ratio is too high.

By adding two additional SAN storage ports on each core switch for non-SGI servers, we will decrease our fan-out ratio to 18:1.

The one hop we used for the ISL connection of additional switches will not increase the latency significantly, because those switches will be within a short distance of each other.

The latency between the sites will be around 208 microseconds. We will use synchronous copying of the storage data and 208 microseconds should not cause significant time delays for applications.

Note: The rule of thumb is that the latency per 1 km of fibre cable is 5 microseconds.

20.2.3 Availability

The McDATA ED-6064 Director solution provides 99.999% availability. This number will effectively be somewhat higher in a dual Director configuration. This SAN design, with two paths from each server to the storage device, will give us the ability for each server to experience an HBA failure without impacting performance and without downtime of the servers.

The McDATA 2027-232 switch incorporates some of the Director class functionality; firmware upgrades and maintenance can be achieved non-disruptively. That is, there is no requirement to redirect all traffic through one switch while the other is being upgraded.

As we have two sites, our SAN is designed to be redundant also in the area of connecting those two sites. We provide redundant paths for data replication and also for data access in the case that storage in the primary site will fail.

20.2.4 Security

When implementing the switches and Directors, we need to accommodate some security related items:

- **Switch and Director security**
  - Change the default passwords to access the switches.
  - Put directors/switches in separate management network if one is already in place for non-SGI functions.

- **Zoning**

  In our case, we have a multi-platform environment. Because we have only one storage port for all servers except SGI, there is no need to implement any zoning. If in the future we expand the number of storage ports, we can group the servers by the storage ports, and separate them into zones for performance reasons. This can be implemented non-disruptively using software zoning.

  This would result in having the following zone:

  – Zone for all SGI servers in both sites.

  Note: There are no data integrity issues if you do not implement zoning. In our example, we only use zoning for performance reasons.

20.2.5 Distance

As you can see from the designs, we are using a maximum of one hop between the switches in each site. This means that there is no practical delay in the performance. All the local connections will use shortwave optics as the servers are within a radius of 500 m. It is possible to move some servers further away by using longwave optics for ISLs. Even with the fabric expansion, as shown in designs, we should have no problems with the delays in the fabric OS for name server and FSPF changes.

For connections from site to site we will use shortwave optics which will be connected over DWDM products. With the use of a DWDM product we will solve the degradation of laser signal, but we still need to adjust the buffers needed on switch and Director ports for this type of connection.

20.2.6 Scalability

From the designs we accommodated three years growth. For the predicted growth we will only accommodate new ports for SGI servers and storage. In case that in the future we would need more ports the design is ready to be expanded with new switches and/or Director port blades.

20.2.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: The clustering solution will failover to the passive server dynamically.
- HBA: If one of the HBA fails, multipath software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.

- Cable: If a cable between a server and the switch fails, multipath software will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: Dual power supplies are supported and no disruption to service will occur if one fails. The replacement power supply can be installed concurrently.
- Switch port: If one of the ports fails, you may replace using a hot-pluggable optics, without outage or performance problem.
- Switch: If a switch fails, the server will use the alternate switch to connect to the storage, without outage or performance problem.
- Director port: If one of the ports fails, you may replace it using a hot-pluggable optics.
- Director blade: If a Director port blade fails, the alternate path will be used. The port on the failed blade can be dynamically plugged into the spare ports in other blades until the blade can be replaced.
- Director: The Director is a 99.999% type solution, it is therefore, improbable that it will fail. Should it fail, there will be no impact if a dual Director solution has been implemented.

20.2.8 Manageability and management software

The Directors and switches have built-in management capabilities:

- Serial interface to the OS, typically, only used to configure the network address.
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use.

To fully utilize the management features, the Directors have to be connected to the network (Ethernet). Although inband management is supported, we are using outband management.

An example of the management network can be seen in Figure 20-10.

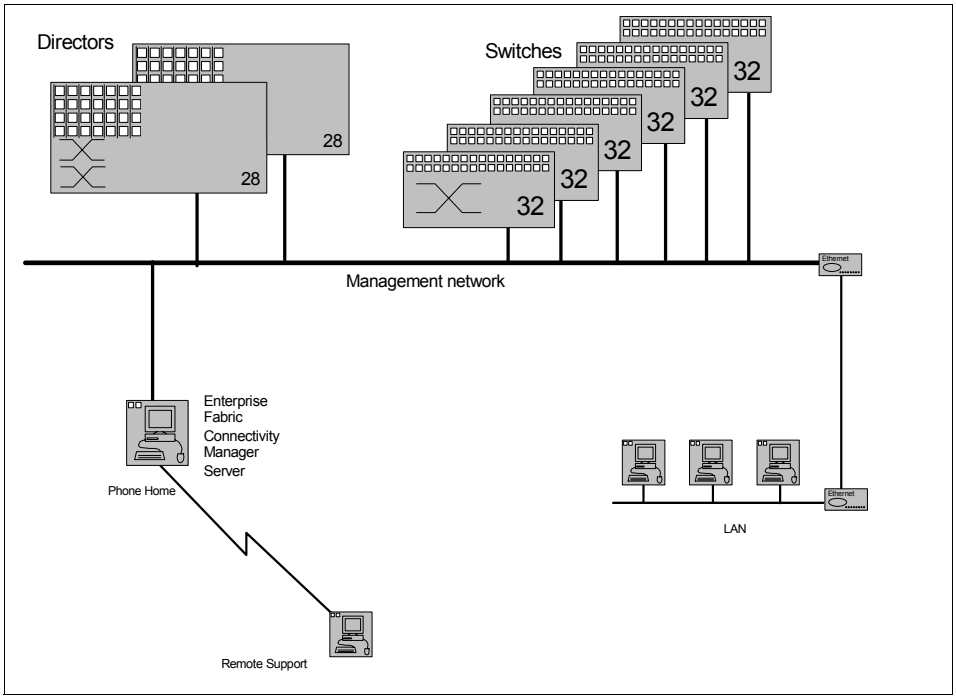


Figure 20-10 Management network

Note: This network can be part of your existing network infrastructure.

Full functional management is available through EFC manager software.

20.3 Case Study 3: Company Three

If we consider the company and its requirements, as detailed in 18.3, “Case Study 3: Company Three” on page 697, we will propose the following solution.

20.3.1 Design

Considering the requirements as detailed in 18.3.5, “Analysis (ports and throughput)” on page 699, we have proposed in this section a switch based solution for our design.

We show our proposed design in Figure 20-11.

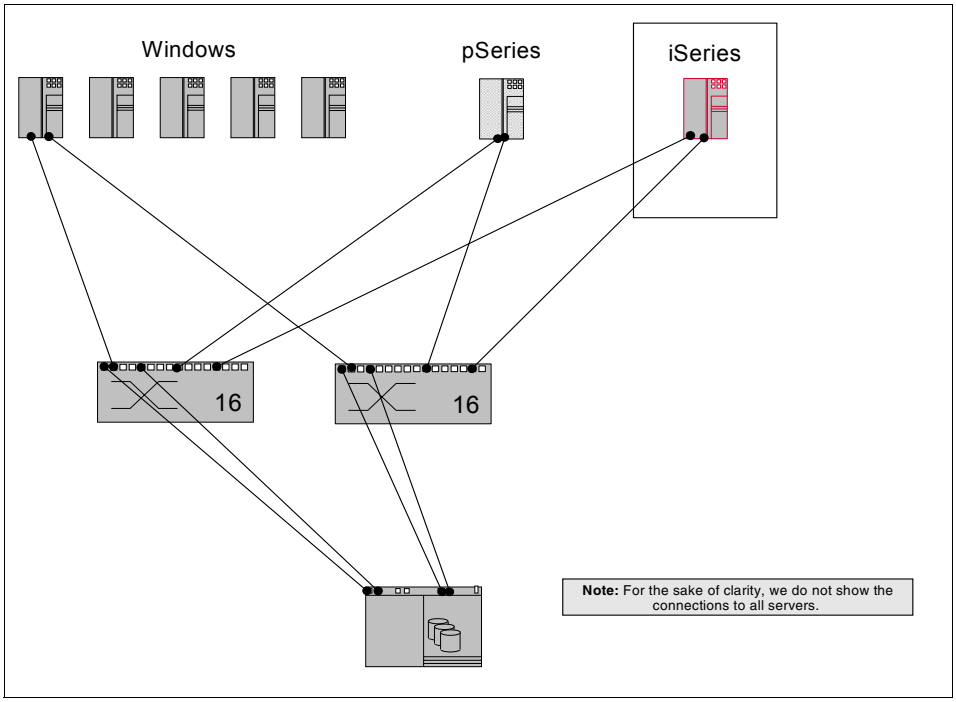


Figure 20-11 Initial design using McDATA 2027-232 Switches

In this scenario, we are using two McDATA 2027-232 switches with a 16 port configuration initially, additional ports can be added in groups of eight. Maximum port configuration of the McDATA 2027-232 is thirty-two.

It is important to note that the McDATA solutions do not support iSeries with code level lower than OS/400 V5.2.

For interoperability matrix, please refer to this website:

http://www.storage.ibm.com/index.htm

As we are using two paths from each Windows and pSeries server to the storage, and two paths from the iSeries server to the storage, we are providing redundancy and high availability. To utilize the dual connections we will use multipathing software (such as IBM's Subsystem Device Driver-SDD) on each of the servers. For iSeries server, OS/400 has its own failover support. The solution shown has been configured with two 16-port switches. With this design we are partially fulfilling the requirements:

- Redundant SAN with two switches and all servers dual connected.
- All bandwidth requirements are met.

- By incorporating some of the Director class functionality, firmware upgrades and maintenance can be achieved non-disruptively. That is, there is no requirement to redirect all traffic through one switch while the other is being upgraded.
- Future growth by adding more switches can be accomplished without service interruption.

In Table 20-9 we summarize the port usage.

Table 20-9 Port usage

Storage	Server	ISL	Spare
4	14	0	14

In the following sections, we outline some of the aspects of this solution.

20.3.2 Performance

In the design we have a fan-out ratio of 6:1 for all servers except iSeries, and 1:1 for iSeries. Latency through the switches is negligible (~2 micro seconds).

20.3.3 Availability

This SAN design, with two paths from each server to the storage device, will give us the ability for each server to experience an HBA failure without impacting performance and without downtime of the servers.

The McDATA 2027-232 switch incorporates some of the Director class functionality; firmware upgrades and maintenance can be achieved non-disruptively. That is, there is no requirement to redirect all traffic through one switch while the other is being upgraded.

20.3.4 Security

When implementing the switches we need to take care about some security related things:

- **Switch security**
  - Change the default passwords to access the switches.
  - Put switches in a separate management network if one is already in place for other functions.
- **Zoning**

Zoning will be implemented to separate the three platforms (Windows, pSeries and iSeries), and isolate the iSeries I/O to its own storage ports.

20.3.5 Distance

In all the connections we will use shortwave optics as the servers are within a radius of 300 m. It is possible to move some server further away by using longwave optics for ISLs.

20.3.6 Scalability

As you can see from the design, there are spare ports for growth. The growth can be achieved without any interruptions in the production environment.

20.3.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Open Systems Server: The clustering solution will failover to the passive server dynamically.
- iSeries server: Application is not available. Server has to be replaced or repaired.
- HBA (Open Systems): If one of the HBA fails, multi-path software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- HBA (iSeries): OS/400 failover support will handle HBA redundancy. There will be no impact of the performance.
- Cable: If a cable between a server and the switch fails, multi-path software will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: The McDATA 2027-232 solution has dual power supplies as a standard feature and no disruption to service will occur should one fail. The replacement power supply can be installed with no outage incurred.
- Switch port: If one of the ports fail, you may replace the hot-pluggable optic. No outage will occur as the server will use its alternate path.
- Switch: Should the switch fail, the server will use its alternate path through another switch without an outage or performance degradation.

20.3.8 Manageability and management software

The switches have built-in management capabilities:

- Serial interface to the OS, typically only used to configure the network address
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use.

To fully utilize the management features, the switches have to be connected to the network (Ethernet). Although inband management is supported, in our solution we are not using ISLs, and therefore, inband management is not an option.

We show an example of setting up the management network in Figure 20-4.

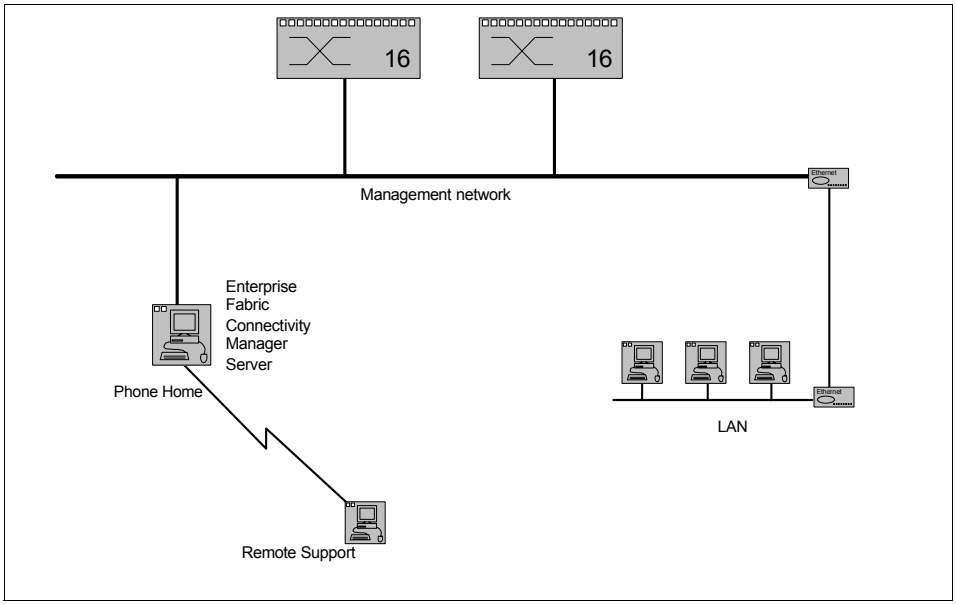


Figure 20-12 Management network

Note: This network can be part of your existing network infrastructure.

Full functional management is available through EFC manager software.

20.4 Case Study 4 - Company Four

If we consider the company and its requirements as detailed in 18.4, “Case Study 4: Company Four” on page 700, we will propose the following solution.

20.4.1 Design

Considering the requirements as detailed in 18.4.5, “Analysis (ports and throughput)” on page 703, we have proposed in this section a switch based solution for our design.

Both the open systems and zSeries platforms will share the same storage device which will provide a much higher availability.

As you can see from our proposed design in Figure 20-13, we have replaced the ESCON director with the McDATA ED-6064 director for both sites. McDATA ED-6064 is designed to support FICON and Fibre Channel intermix which is best suited for our scenario with two different platforms. The McDATA ED-6064 shown has been configured with thirty-two ports initially, additional ports can be added in groups of four.

.

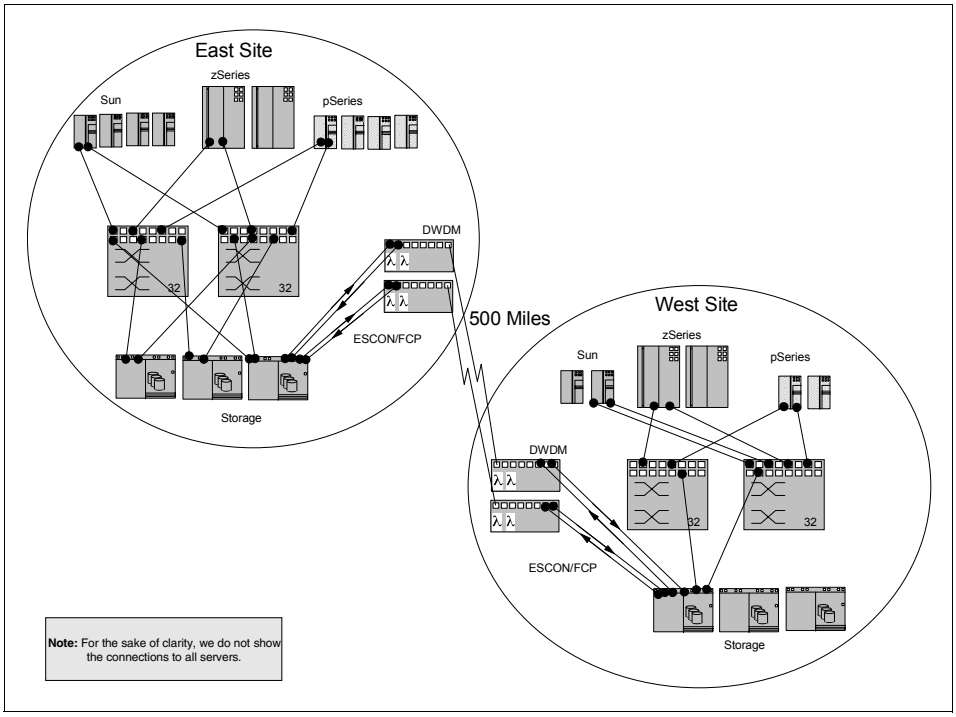


Figure 20-13 Initial design using McDATA ED-6064 Switches.

As we are using redundant paths from each server and to the storage we are providing high availability. To utilize the redundant connections we will use multipathing software (such as IBM's Subsystem Device Driver) on each of the open systems servers, while for our mainframe servers we have z/OS built in failover support for path redundancy. The solution shown has been configured with two 32-ports initially. With this design, we are partially fulfilling the requirements:

- Redundant SAN with two switches and all servers dual connected.
- All bandwidth requirements are met.
- By incorporating some of the Director class functionality, firmware upgrades and maintenance can be achieved non-disruptively. That is, there is no requirement to redirect all traffic through one switch while the other is being upgraded.
- Future growth by adding more switches can be accomplished without service interruption.

In the following sections, we outline some of the aspects of this solution.

For storage replication we are using ESCON links over DWDM. Depending on the implementation time frame, we may be able to substitute these with FCP using PPRC V2. The peak bandwidth used by the servers is 11.58 MB/s. As only 30% of these are writes then we need to accommodate 3.48 MB/s. Two pairs of ESCON channels will accomplish this and provide redundancy. The same DWDM devices used for open systems can be used for the OS/390 requirement.

In Table 20-10 we show the ports at the East site.

Table 20-10 Ports: East site

Storage	Server	ISL	Spare
10	20	0	34

In Table 20-11 we show the ports at the West site.

Table 20-11 Ports: West Site

Storage	Server	ISL	Spare
10	12	0	42

20.4.2 Performance

In the design, we have a fan-out ratio of 20:6 for all servers in the East site. The fan-out ratio in the West site is 12:6. Latency through the switches is negligible (~2 micro seconds).

The latency between the sites will be around 4 milliseconds.

20.4.3 Availability

This SAN design, with two paths from each server to the storage device, will give us the ability for each server to experience an HBA failure without impacting performance and without downtime of the servers.

The McDATA ED-6064 supports functionalities such as non-disruptive firmware upgrades and maintenance. That is, there is no requirement to redirect all traffic through one director while the other is being upgraded.

20.4.4 Security

When implementing the directors we need to take care about some security related things:

- **Director security**
  - Change the default passwords to access the switches.

  - Put directors in a separate management network if one is already in place for other functions.

- **Zoning**

We will be implementing zoning so we can group the servers by the storage ports (FICON/FCP) and separate then into zones.

20.4.5 Distance

In all the connections, we will use shortwave SFPs as the servers are within a radius of 500 m. It is possible to move some servers further away by using longwave SFPs for ISLs.

20.4.6 Scalability

The design accommodates three years growth. The growth can be achieved without any interruptions in the production environment.

20.4.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: The clustering solution will failover to the passive server dynamically.
- HBA: If one of the HBA fails, multi-path software or z/OS will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multi-path software or z/OS will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: The McDATA ED-6064 solution has dual power supplies as a standard feature and no disruption to service will occur should one fail. The replacement power supply can be installed with no outage incurred.
- Director port: If one of the ports fails, you may replace the hot-pluggable optic. No outage will occur as the server will use its alternate path.
- Director: Should the director fail, the server will use its alternate path through another switch without an outage or performance degradation.
- DWDM device: Should the DWDM device fail, the alternate one will be used.

20.4.8 Manageability and management software

The switches have built-in management capabilities:

- Serial interface to the OS, typically only used to configure the network address.
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use.

To fully utilize the management features, the switches have to be connected to the network (Ethernet). Although inband management is supported, in our solution we are not using ISLs, and therefore, inband management is not an option.

We show an example of setting up the management network in Figure 20-14.

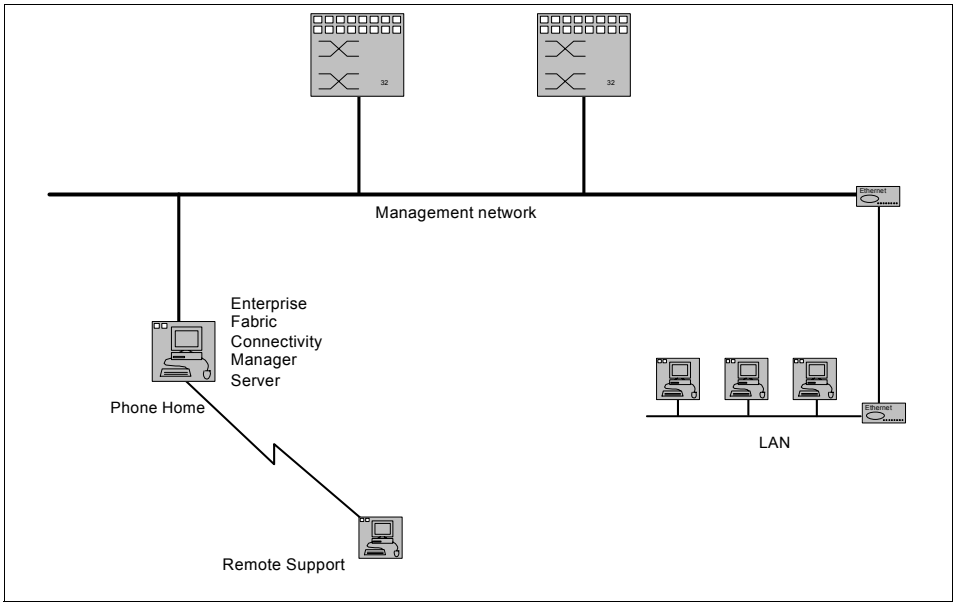


Figure 20-14 Management network

Note: This network can be part of your existing network infrastructure.

Full functional management is available through EFC Manager software.

20.5 Case Study 5 - Company Five

If we consider the company and its requirements as detailed in 18.5, “Case Study 5: Company Five” on page 705, we will propose the following solution.

20.5.1 Design

Considering the requirements as detailed in 18.5.5, “Analysis (ports and throughput)” on page 708, we selected a solution based around the McDATA 2026-224 since it supports FC-AL which is required for accessing tape drives.

The proposed design is shown in Figure 20-15.

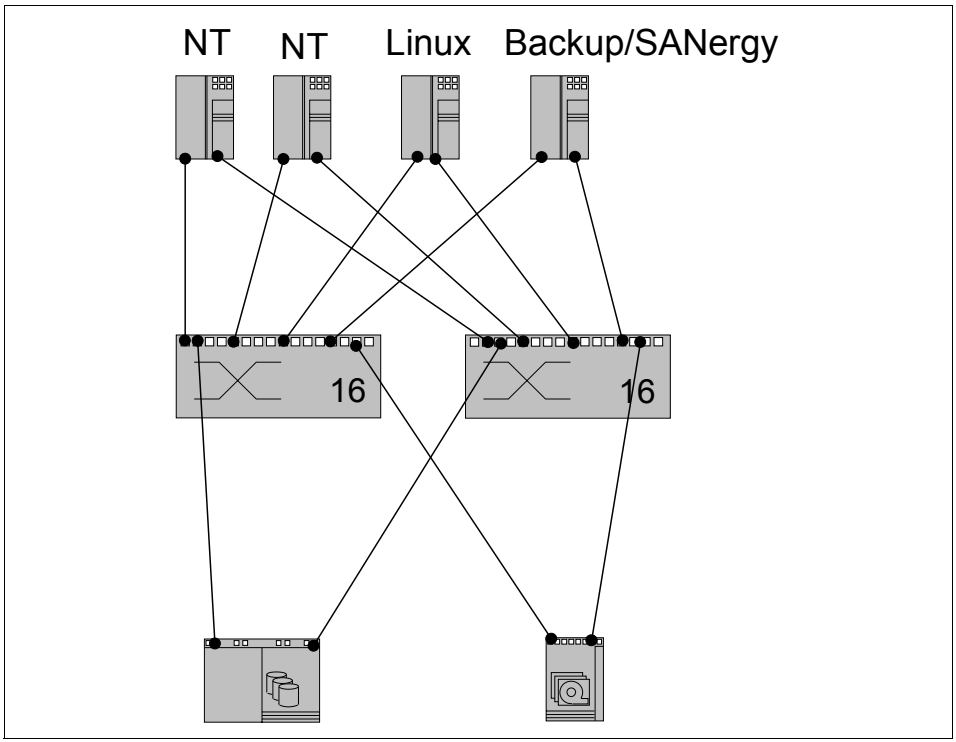


Figure 20-15 The proposed design for Company Five

As we are using two paths from each server to the storage we are providing redundancy and high availability. To utilize the dual connections, we will use multipathing software (such as IBM’s Subsystem Device Driver) on each of the servers. The solution shown has been configured with two 16-port switches (additional ports can be added in groups of eight, to a maximum of 24). With this design we are partially fulfilling the requirements:

- Redundant SAN with two switches. All servers are dual connected to the switches. Tape drive is connected to the same SAN via FC-AL.
- All bandwidth requirements are met.
- By incorporating some of the Director class functionality, firmware upgrades and maintenance can be achieved non-disruptively. That is, there is no requirement to redirect all traffic through one switch while the other is being upgraded.
- Future growth by adding more switches can be accomplished without service interruption.

You will notice that we have introduced a fourth server to accommodate heterogeneous file sharing (using Tivoli SANergy) and also backup/restore software (using Tivoli Storage Manager). After the migration from NT to Linux is complete, we could consider using the vacant NT server, thus reducing the number of servers to three.

In Table 20-12 we show the initial ports used.

Table 20-12 Initial ports

Storage	Server	ISL	Spare
4	8	0	20

In the following sections, we outline some of the aspects of this solution.

20.5.2 Performance

In the initial design, we have a 4:1 fan-out ratio for all servers accessing the storage. This means that four servers are accessing one storage port. All servers can also access the tape devices.

Attention: Host software (such as Tivoli Storage Manager) must be present to allow functions that enable tape sharing and/or tape library sharing.

McDATA 2026-224 supports FC-AL which enables direct access of the tape drives.

20.5.3 Availability

All servers and the storage have multiple paths giving resilience through redundancy.

The McDATA 2026-224 has these functionalities: firmware upgrades and maintenance can be achieved non-disruptively. That is, there is no requirement to redirect all traffic through one switch while the other is being upgraded.

20.5.4 Security

When implementing the devices, we need to address some security issues:

- **Director security**
  - Change the default passwords.
  - Put the directors into a separate management network if one is already in place for other functions.
- **Zoning**

  We have only one disk storage port per switch for all servers, therefore, there is no requirement to implement any zoning. In the future, if we expand the number of storage ports, we can group the servers to the storage ports in separate zones for performance reasons. This can be implemented non-disruptively using software zoning.

  Note: There are no data integrity issues if you do not implement zoning. In our example we would only use zoning for performance reasons.

20.5.5 Distance

All devices (servers, switches, and storage) will use shortwave optics and cables as they are within a radius of 500 m.

20.5.6 Scalability

In the design, we accommodated three years growth. We have enough bandwidth allocated for all requirements in the next three years. If, in the future, we need more ports, then the 2026-224 can accommodate another 10 servers or tape drives. McDATA 2026-224 is upgradeable and supports a maximum number of 24 ports.

20.5.7 “What if” failure scenarios

- Server: Application will not be available. Server has to be replaced or repaired.
- HBA: If one of the HBA fails, multi-path software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.

- Cable: If a cable between a server and the switch fails, multi-path software will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: McDATA 2026-224 has redundant power supplies.
- Switch Port: If one of the ports fails, you may replace it using a hot-pluggable optic, without outage or performance problem.
- Switch: If a switch fails the alternate switch will accommodate the workload.

20.5.8 Manageability and management software

The switches have built-in management capabilities:

- Serial interface to the OS, typically only used to configure the network address
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use

To fully utilize the management features, switches have to be connected to the network (Ethernet). Although inband management is supported, in our solution we are not using ISLs, and therefore, inband management is not an option.

We show an example of setting up the management network in Figure 20-16.

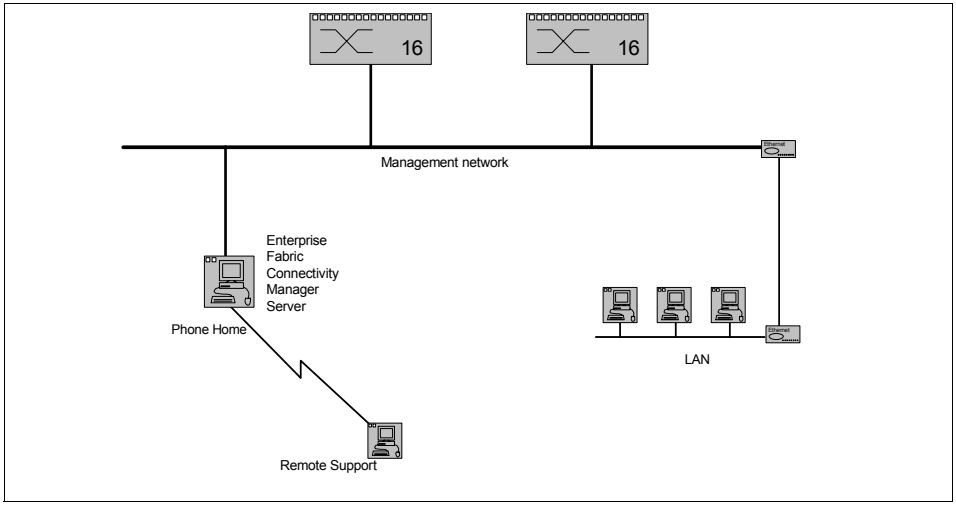


Figure 20-16 Management network

Note: This network can be part of your existing network infrastructure.

Full functional management is available through EFC Manager software.

20.6 Case Study 6: Company Six

If we consider the company and its requirements, as detailed in 18.6, “Case Study 6: Company Six” on page 710, we will propose the following solution.

20.6.1 Design

Considering the requirements as detailed in 18.6.5, “Analysis (ports and throughput)” on page 712, we selected a solution based around the McDATA 2026-224.

We show the proposed design for Primary site in Figure 20-17.

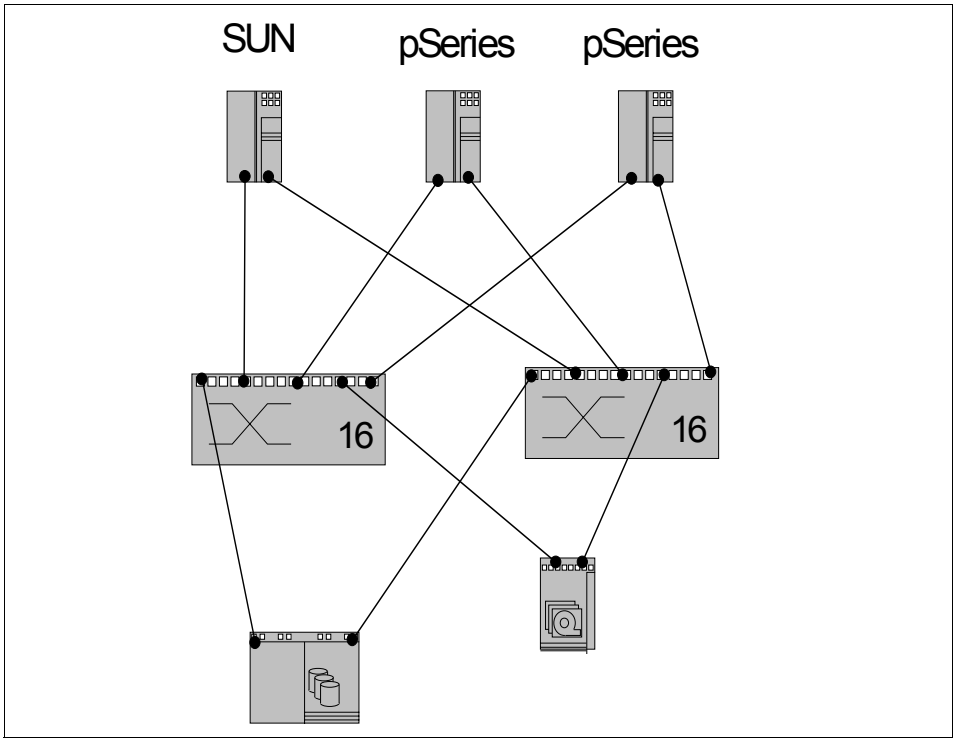


Figure 20-17 The proposed design for the Primary site

In the secondary site, we already have a SAN, using Compaq StorageWorks SAN Switch 16. These are Brocade OEM switches of the SilkWorm 2800.

You can see, the proposed design of the Secondary site in Figure 20-18.

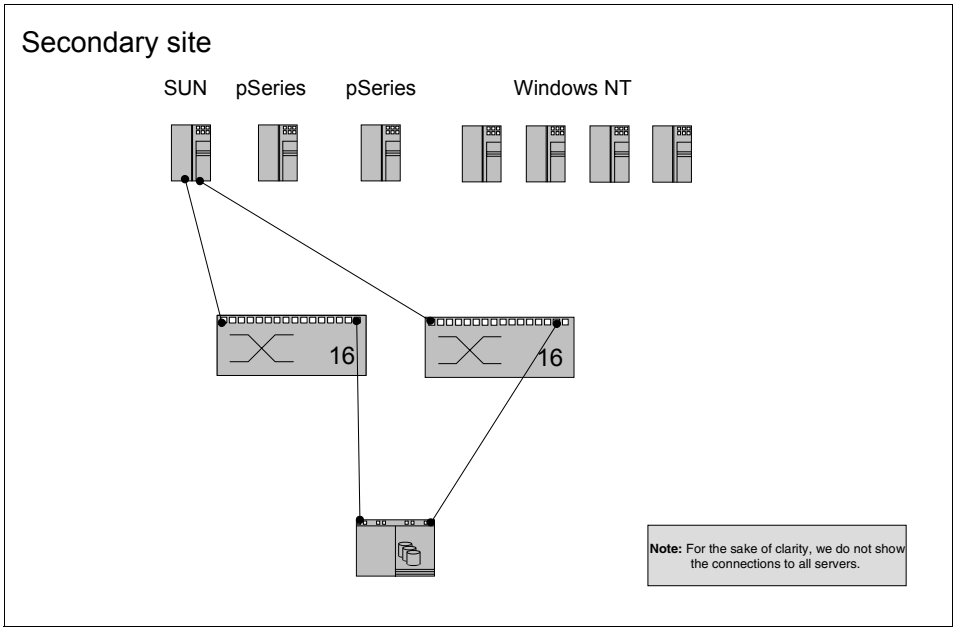


Figure 20-18 The proposed design for the Secondary site

Both sites are connected with two ISLs. Because the distance is 900 miles, we will use DWDM products to connect them. You can see the connection setup in Figure 20-19.

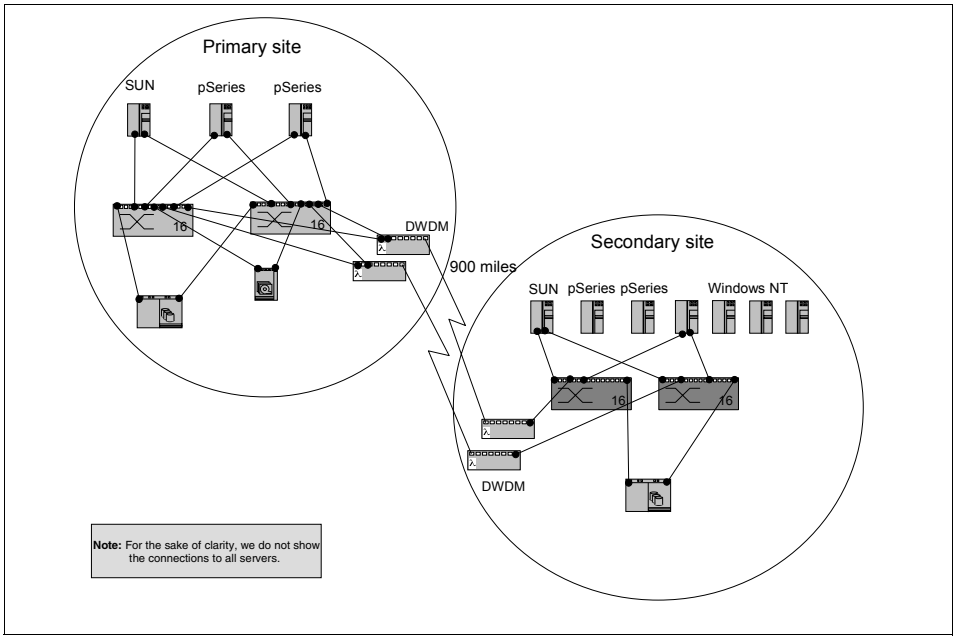


Figure 20-19 The complete solution

Note: To the best of our knowledge, this solution (with the McDATA 2026-224 and the Compaq) has not been tested, and is therefore not certified.

As you can see we used a McDATA 2026-224 with 16 ports configured initially, additional ports can be added in groups of eight. With such a design, we are fulfilling these requirements:

- Redundant SAN with two switches and all servers dual connected
- All bandwidth requirements are met.
- These directors supports functionalities such as a non-disruptive firmware upgrade. That is, there is no requirement to redirect all traffic through one switch while the other is being upgraded.
- Future growth by adding more directors can be accomplished without service interruption.
- We are re-using existing equipment (directors and tape).
- We are providing storage replication to remote site for disaster recovery.
- We are improving backup performance by providing infrastructure for LAN-free backup.

Attention: Host software (such as Tivoli Storage Manager) must be present to allow functions that enable tape sharing and/or tape library sharing.

In Table 20-13 we show the ports at the Local site.

Table 20-13 Ports at Local site

Storage	Server	ISL	Spare
4	6	4	20

There are ample ports left at the Remote site, the final port allocation for the remote site is represented in Table 20-14.

Table 20-14 Ports at Remote site

Storage	Server	ISL	Spare
2	14	2	14

In the following sections, we will outline some aspects of the design.

20.6.2 Performance

In the initial design, we have a 3:1 fan-out ratio for all servers accessing the storage on the Primary site. This means that three servers are accessing one storage port. All servers can also access the tape device with greater performance and availability, which was the requirement for LAN-free backup implementation. There are no latency issues, because the traffic is only going through the Director, where the latency in the low microseconds.

At the Secondary site we have 7:1 fan-out ratio. In this case, the traffic is only going through the switches and again the latency is so low as to be of no significance.

20.6.3 Availability

All servers and the storage have multiple paths giving resilience through redundancy.

As we have two sites our SAN is designed to be redundant also in the area of connecting those two sites. We are providing redundant paths for data replication. It recommended that the connections between two DWDM boxes are also redundant. This however, could be a considerable expense.

20.6.4 Security

When implementing the directors, we need to address some security issues:

- **Director security**
  - Change the default passwords.
  - Either put the Directors into an existing separate management network if one is already in place for other functions, or implement a new network.
- **Zoning**

  In our case, we will implement zoning to group each paltforms including servers accessing tape drives for performance reasons.

20.6.5 Distance

As can be seen from the designs, we are only using one switch between the server and the storage in each site. This means that there is no practical delay in the performance. All the local connections will use shortwave optics.

With the use of a DWDM product we will solve the degradation of laser signal, but we still need to adjust the buffers needed on switch ports for this type of connection. On the switches at both sites, we need to enable the Extended Distance option for the ISL ports.This assigns more buffer credits to those E_Ports being used by long distance ISLs.

20.6.6 Scalability

From the designs we have accommodated three years growth. We have enough bandwidth allocated for all requirements in the next three years. The only expansion will be probably by adding additional storage ports, and we accommodated the space for this. In case we would need more ports, the design is ready to be expanded with new directors.

20.6.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: Application will not be available. Server has to be replaced or repaired.
- HBA: If one of the HBA fails, multi-path software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multi-path software will automatically failover workload to the alternate path. If a cable between

the switch and the disk storage fails, an alternate route will be used. There is no performance loss.

- Power supply: McDATA 2026-224 has redundant power supplies.
- Switch: If one of the ports fails, you may replace it using a hot-pluggable optic, without outage or performance problem.
- Switch: If a switch fails, the server will use the alternate switch to connect to the storage, without outage or performance problem.

20.6.8 Manageability and management software

The switches have built-in management capabilities:

- Serial interface to the OS, typically, only used to configure the network address.
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use.

To fully utilize the management features, the switches have to be connected to the network (Ethernet). Although inband management is supported, in our solution we are not using ISLs and therefore inband management is not an option.

We show an example of setting up the management network in Figure 20-4.

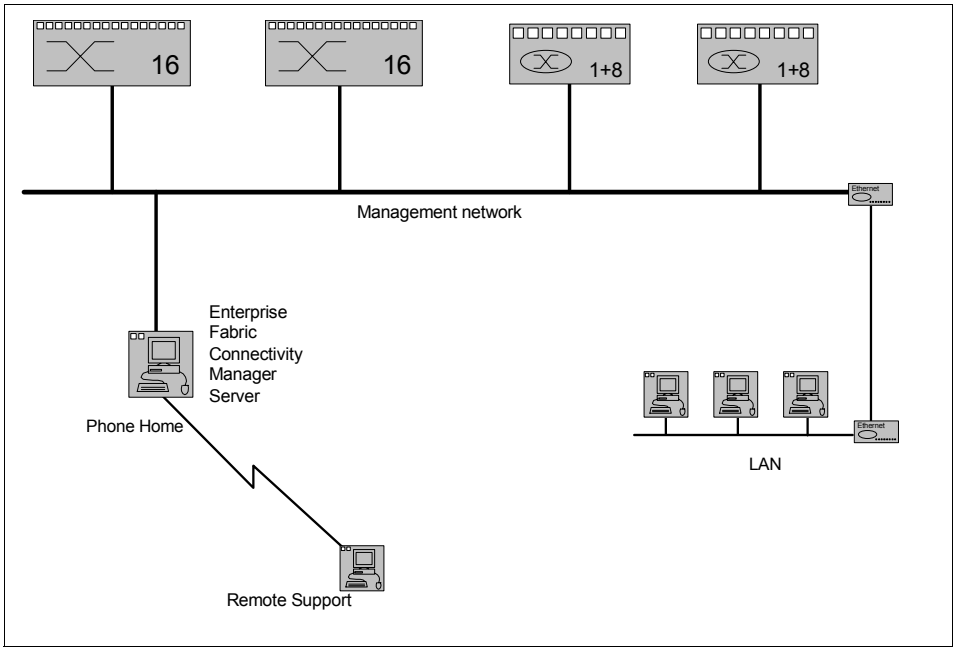


Figure 20-20 Management network

Note: This network can be part of your existing network infrastructure.

Full functional management is available through EFC manager software.

The introduction of an Enterprise SAN Management tool (such as Tivoli Storage Network Manager) should be effective the full operational status of all switches).

The DWDM products would also need to be integrated into the management infrastructure.

21

Chapter 21.

IBM TotalStorage SAN n-type family solutions

In this chapter we will describe solutions that are based upon the ICNT FC/9000 Director.

© Copyright IBM Corp. 2004. All rights reserved.

21.1 Case Study 1: Company One

If we consider the company and its requirements, as detailed in 18.1, “Case Study 1: Company One” on page 686, we will propose the following solution.

21.1.1 Design

When considering the requirements as detailed in 18.1.5, “Analysis (ports and throughput)” on page 687, we selected a solution based around the CNT FC/9000 Director Model 128. As the size of the environment grows, the SAN can grow to accommodate this and ultimately will be expanded to two fabrics.

A decision as to when to migrate from a single director to two directors would need to be made based on identifying your business and technological goals. Indeed, the entire solution could be satisfied using a single director, if for instance real estate is a problem, or the SAN could be implemented with a dual fabric right from the outset.

The initial solution is shown in Figure 21-1.

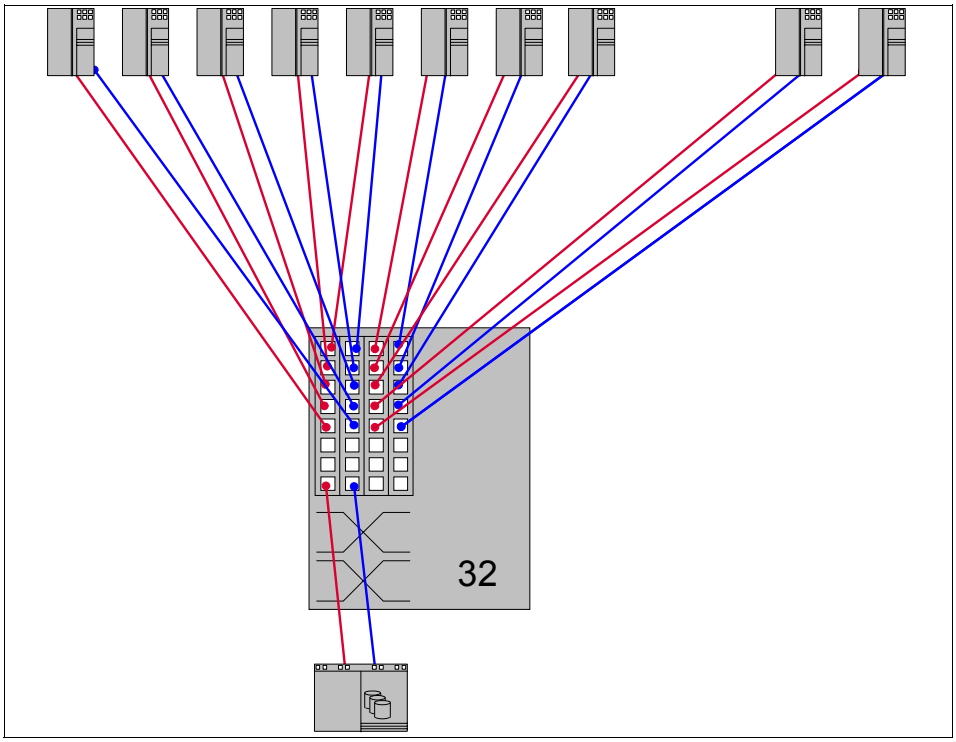


Figure 21-1 Single director solution

There is a summary of the ports and their usage in Table 21-1.

Table 21-1 Port usage at Company One in the first year

Director	Servers	Storage	ISLs	Spare
Director 1	20	2	0	10

Note: It would be possible to have only two spare ports by using a 24-port director, but as discussed below the SAN is more flexible with a 32-port device.

As we are using two paths from each server to the director, we are providing redundancy and high availability. For utilizing this physical setup, we will use multipathing software on each of the servers. With such a design, we are fulfilling these requirements:

- All bandwidth requirements are met (40 KB/s and 4 MB/s from servers and 8.32MB/s to storage).
- SAN components can be upgraded without impact on the servers. The CNT FC/9000 Director Model 128 allows us to upgrade firmware non disruptively.
- Possible growth without impact on the production. The CNT FC/9000 director allows us to add FIO blades non disruptively.
- The director class product offers 99.999% uptime. (If we absolutely must avoid a single point-of-failure then we can implement the solution from the outset using two directors.)
- Each server is connected into two different blades for added resilience.
- The solution could have been implemented using three blades (24 ports) but the suggested design uses four blades (32 ports). This means that even in the event of an entire blade failing, all of the connections on that blade could be moved to another blade, while the failed blade is replaced. If cost is of higher concern than resilience, then the cost of a blade could be saved. If that is the case, then the ports from that blade would be distributed to other blades, making sure that all nodes were connected to two blades.
- With this design, we have 10 SAN ports free for future expansion. This means that five additional servers can be added, each with redundant connections. Beyond that, it would be necessary to add another one or more I/O blades.

So, let us discuss the steps that we need to go through to add more servers. The expected growth is to 20 servers in the first year.

Note: None of this expansion requires any interruption to the SAN.

As servers are added, they will follow the same connection scenario as the solution in place. In other words:

- The new servers will be connected into the director by two links.
- The two links will be into different blades.
- A maximum of five additional servers can be added before the need for another FIO blade.
- To add the sixth server will involve the installation of another FIO blade.
- The additional blade gives an extra 8 ports, enough for 4 extra servers.
  - Some connections should be moved at this point to ensure that no server has both of its connections going into one FIO blade.
- To add the 20th server we will need a 6th blade.

We can see the 20 server, single director solution in Figure 21-2.

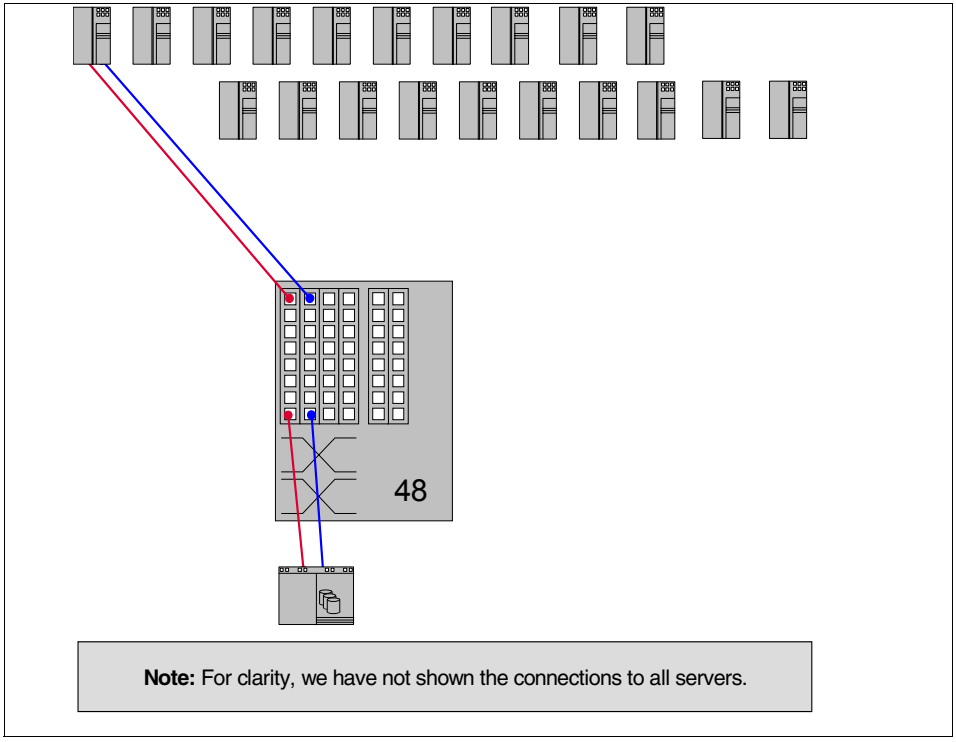


Figure 21-2 The 20-server solution

There is a summary of the ports and their usage in Table 21-2.

Table 21-2 Port usage at Company One in the second year

Director	Servers	Storage	ISLs	Spare
Director 1	40	2	0	6

The projected growth over the following year is to 40 servers. In order to accommodate this, we will again need more FIO blades.

The growth could, again follow the pattern already in place. That is, the new servers would be connected into two blades in the single director. As an alternative, a second director could be introduced. When the port count in the SAN rises above 64, it becomes an attractive option to use two directors as we shall see.

The single director approach

The steps that are needed for this expansion are as follows:

- Up to three additional servers can be added before needing more FIO blades.
- Another FIO blade can be added, allowing another four servers to be connected (27 in total).
- Another FIO blade can be added, allowing another four servers to be connected (31 in total).
- The port count is now at 64, so to add another FIO, we need to extend the backplane.

Attention: This is a disruptive upgrade.

- A further FIO can be added allowing another four servers to be connected (35 in total).
- A further FIO can be added allowing another four servers to be connected (39 in total).
- A further FIO can be added allowing connection of the 40th server (leaving 6 spare ports).

The final configuration is shown below in Figure 21-3.

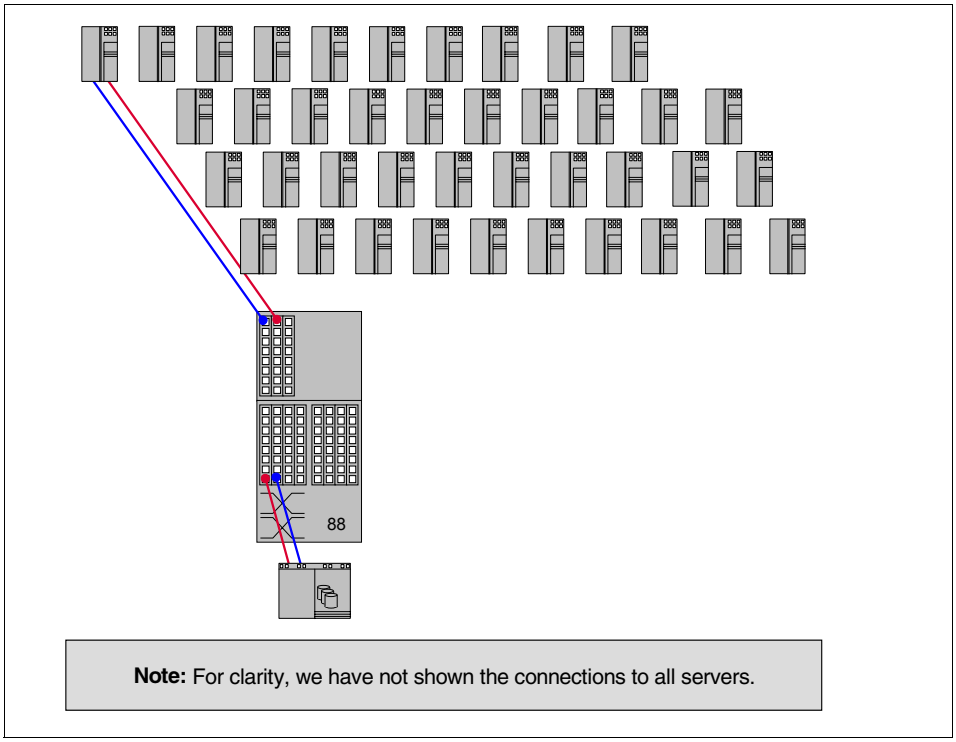


Figure 21-3 The 40 server, single director solution

There is a summary of the ports and their usage in Table 21-3.

Table 21-3 Port usage at Company One in the third year, single director

Director	Servers	Storage	ISLs	Spare
Director 1	80	2	0	6

As we saw, taking the CNT FC/9000 over 64 ports required some down time (CNT specify two to six hours for this). There are two ways around this problem:

- Purchase the backplane extension at the time of original purchase.
- Rather than extending the first director, use a second.

The second option above is the preferred route because the:

- Initial outlay is kept as low as possible.
- Initial solution is based around director class product.
- Final (40 server) solution has dual fabric leading to added resilience.
- Costs of the two final solutions are quite similar.

The dual fabric solution is shown in Figure 21-4.

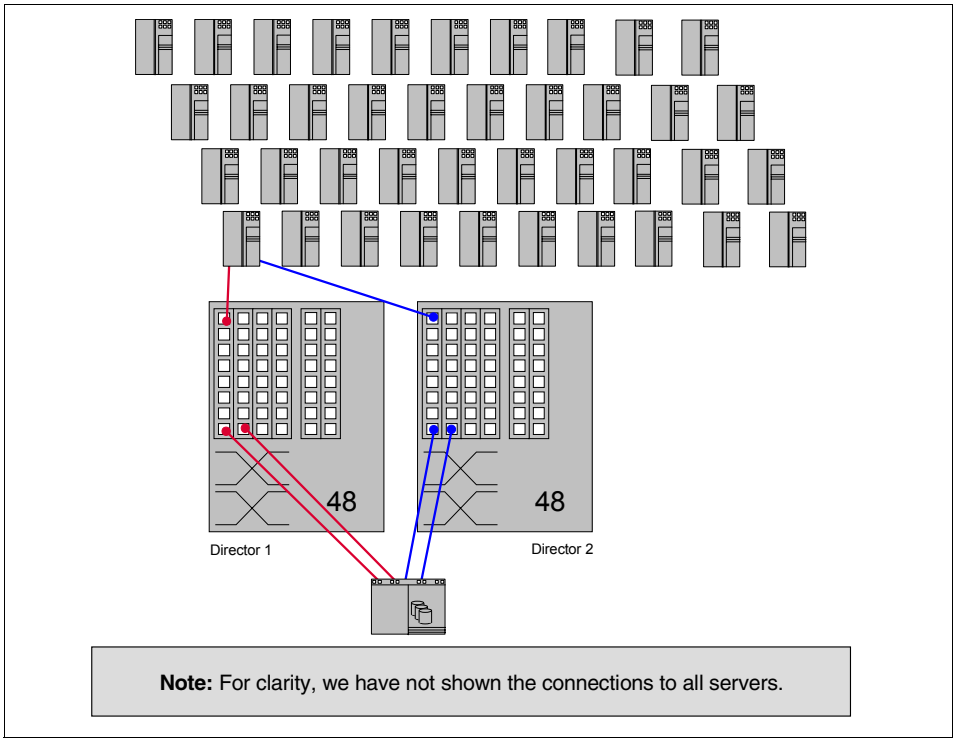


Figure 21-4 The 40-server, dual director solution

There is a summary of the ports and their usage in Table 21-4.

Table 21-4 Port usage at Company One in the third year, dual director

Director	Servers	Storage	ISLs	Spare
Director 1	40	2	0	6
Director 2	40	2	0	6
Total	80	4	0	12

The steps to migrate from the single fabric solution in Figure 21-2 to the dual fabric solution in Figure 21-4 are outlined here:

Note: This migration is non disruptive to the SAN.

- Install the second director.

- Connect the storage to the new director using two links (two for added resilience at low outlay).
- Migrate one half of the connections from the existing servers to the new director.
- Add the new servers, connecting each server to both directors.

21.1.2 Performance

The initial bandwidth requirement is 8.32 MB/s and this is well within the bandwidth capabilities of the links to the storage.

At the final stage, with 40 servers, we can assume a requirement of 32 MB/s on the servers. The four links to the storage will easily carry this load.

There are no concerns over latency in the design.

21.1.3 Availability

All servers and the storage have multiple paths giving resilience through redundancy. The director class FC/9000 should give 99.999% uptime. In the event that the single backplane is of concern, then a dual fabric SAN could easily be implemented by using a second director.

The dual director solution, if implemented at the outset, would appear very similar to the final solution, but with fewer FIO blades.

The dual fabric solution could be implemented easily at any stage by adding the second director. It should be borne in mind that blades from the first director could be used in the second one.

21.1.4 Security

When implementing the director, we need to address some security issues:

- **Director security**
  - Change the default passwords.
  - Either put the directors into an existing separate management network if one is already in place for other functions, or implement a new network.
- **Zoning**

  Because we have multiple storage ports which are accessible by all servers, we might consider implementing software zoning for performance reasons, but in fact, we have no need to because of the bandwidths discussed above.

21.1.5 Distance

In this case, there are no requirements for long distances. Shortwave lasers and multimode cable will allow 300m which is ample for this scenario.

21.1.6 Scalability

The scalability for the first few years has been discussed already. On going scalability is as follows:

- Up to two more FIO blades per director
  - That gives an extra 16 servers.
- To add more than 16 more servers
  - A CNT FC/9000 Director Model 128 can be shutdown and upgraded to as many as 128 ports.
  - It can be brought back on line and the other director upgraded similarly.
  - That will allow for a further 64 dual pathed connections.
- Further scalability is possible, but by that stage the entire situation should be reconsidered.

21.1.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: The clustering solution will failover to the passive server dynamically.
- HBA: If one of the HBA fails, multipath software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the director fails, multipath software will automatically failover workload to the alternate path. If a cable between the director and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: The directors have redundant power supplies.
- Director port: If one of the ports fails, you may replace it using a hot-pluggable SFP.
- Director: In the highly unlikely (though technically possible) event that a backplane fails, then that entire director *could* stop working. In that case, considerable downtime would occur. If the 99.999% uptime is not enough then a second director should be utilized from the outset! If a director in a dual fabric design was to have such a problem, then the data would flow through the other fabric.

21.1.8 Manageability and management software

The management of the SAN solution will be carried out using IN-VSN. The Enterprise Manager Server (EM) should be on a private network with the directors for security. The EM can be used to manage the SAN, and so can workstations either on the corporate LAN/WAN by running the Enterprise Manager Client software.

The topology of the management subsystem is shown in Figure 21-5.

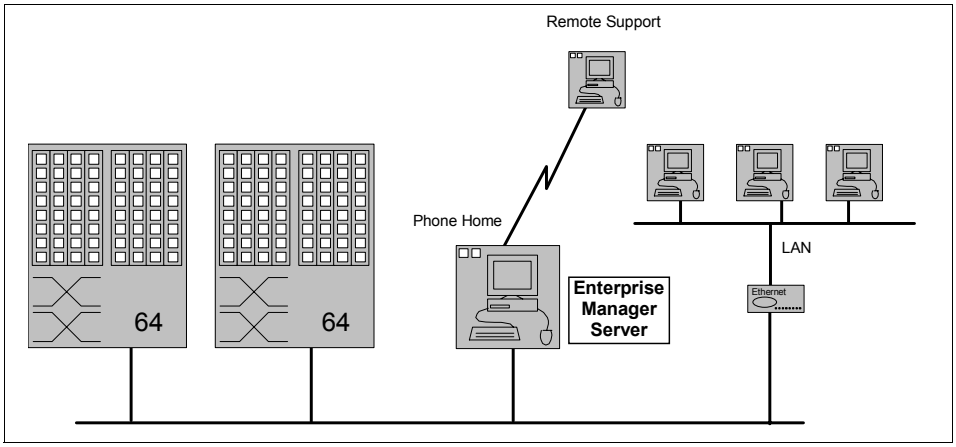


Figure 21-5 Management of a CNT SAN solution

All directors should be connected to the same LAN in this case, because there are no ISLs; we cannot use in-band management to the second director

21.2 Case Study 2: Company Two

If we consider the company and its requirements as detailed in 18.2, “Case Study 2: Company Two” on page 689, we will propose the following solution.

21.2.1 Design

In considering the requirements as detailed in 18.2.5, “Analysis (ports and throughput)” on page 692, we selected a solution based on the CNT FC/9000 Model 128 director.

The present port requirements for Company Two’s SAN ports are tabulated in Table 21-5.

Table 21-5 Summary of port requirements

Getwell					Feelinbad				
1Gb/s		2Gb/s			1Gb/s		2Gb/s		
Srv	Stor	Srv	Stor	ISL	Srv	Stor	Srv	Stor	ISL
108	2	4	6	8	18	2	2	6	8
110		18			20		16		
128					36				

In our design, we will use the CNT SAN256N with 80 ports configured initially. This gives added resilience and provides a solution with no single points of failure. Company Two are obviously sensitive to the need for resilience, bearing in mind the potential $80,000/hour cost of downtime. For this reason, we recommend dual directors at each site.

Getwell site

The design for Getwell is shown in Figure 21-6.

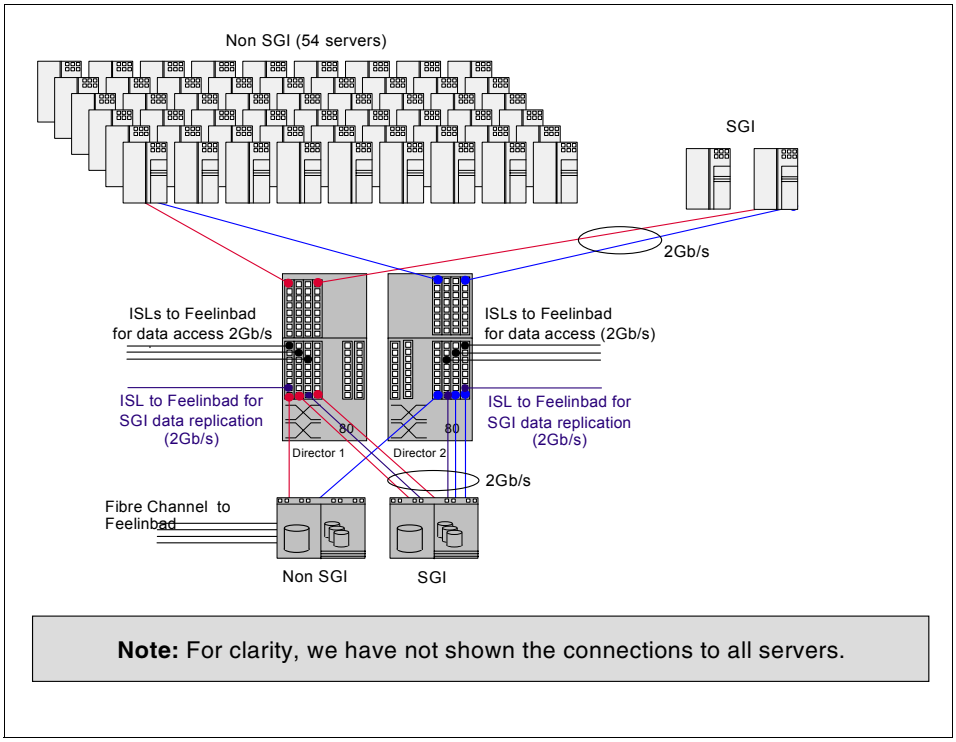


Figure 21-6 The Getwell configuration

We provided the following connections:

- Six ISLs to Feelinbad site for data access, three per fabric
- Two ISLs for SGI storage data replication, one per fabric
- Two ports for storage device for non-SGI servers, one per fabric
- Four ports for storage device for SGI servers, two per fabric
- Two ports for storage device for SGI data replication, one per fabric
- One hundred-eight non-SGI server ports, 54 per fabric
- Four SGI server ports, two per fabric
- Thirty-two free SAN ports
- Four Fibre Channel connections for non-SGI servers data replication

The SAN is fully redundant with no single point-of-failure.

There is a summary of the ports and their usage in Table 21-6.

Table 21-6 Port usage at Getwell

Director	Servers	Storage	ISLs	Spare
Director 1	56	4	4	16
Director 2	56	4	4	16
Total	112	8	8	32

Feelinbad site

The design for Feelinbad is shown in Figure 21-7.

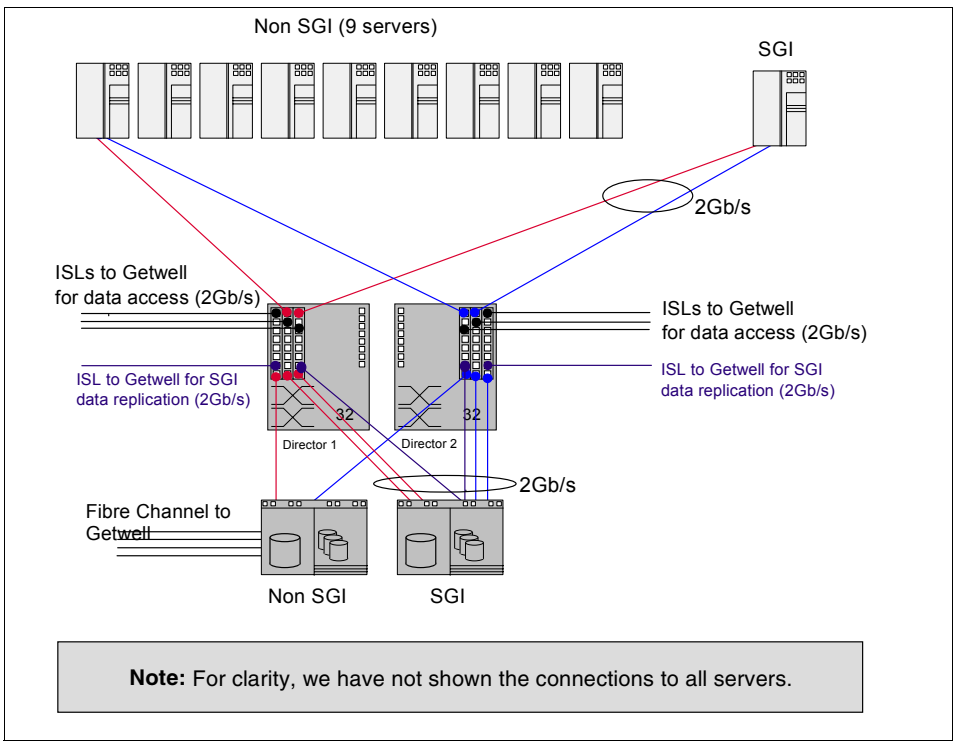


Figure 21-7 The Feelinbad configuration

We provided the following connections:

- Six ISLs to Feelinbad site for data access, three per fabric
- Two ISLs for SGI storage data replication, one per fabric
- Two ports for storage device for non-SGI servers, one per fabric
- Four ports for storage device for SGI servers, two per fabric
- Two ports for storage device for SGI data replication, one per fabric
- Eighteen non-SGI server ports, nine per fabric

- Two SGI server ports, one per fabric
- Twenty-eight free SAN ports
- Four Fibre Channel connections for non-SGI servers data replication

The SAN is fully redundant with no single point-of-failure.

Note: As an alternative for the Feelinbad configuration, we could also consider offering one SAN256N and create two logical domains; one for SGI and one for non-SGI. This would still allow the customer to keep the platforms completely separate (including the ability to perform firmware upgrades for only one domain) while reducing costs for the director (one chassis instead of two!). The logical domain feature is licensed by specifying IBM Feature Code # 7303 (SAN services bundle). The number of ports and blades for each domain would preferably remain unchanged compared to the two-director scenario.

There is a summary of the ports and their usage in Table 21-7.

Table 21-7 Port usage at Feelinbad

Director	Servers	Storage	ISLs	Spare
Director 1	10	4	4	14
Director 2	10	4	4	14
Total	20	8	8	28

Combined solution

With such a design we are fulfilling the requirements:

- No single point-of-failure — redundant SAN.
- All bandwidth requirements are met.
- SAN components can be upgraded without impact on the servers. CNT FC/9000 director allows us to upgrade firmware non disruptively.
- Possible growth without impact on the production. The CNT FC/9000 director allows us to add FIO blades non disruptively.
- The director class product offers 99.999% uptime.

We also provided enough ports for the planned growth on both sides:

- Four SAN storage ports for SGI storage
- Four SAN server ports for SGI servers (2 dual server ports)
- Four SAN ports for ISLs

21.2.2 Performance

As we can see from the design we have an initial 54:1 fan-out ratio for all non-SGI servers accessing each storage device port.

Note: Refer to the storage device documentation to determine if a 54:1 fan-out ratio is too high for it. We can add additional storage ports for non-SGI servers.

This means that 54 servers are accessing the same storage pool. As we need 75.3 MB/s total for all the servers, we have enough bandwidth on the storage ports.

The fan-out ratio for the SGI servers is 1:1. The fan-out ratio will not change in the future. When we increase the number of ports on the SGI servers we will also increase the SGI storage ports by the same number. As we can see from the design, the number of ISLs covers all bandwidth requirements.

The latency within the directors is so trivial (a few microseconds) that it can safely be ignored. Between the sites, the latency in the ISLs will be around 208 microseconds. We will use synchronous copying of the storage data and 208 microseconds should not cause significant time delays for applications.

Note: The rule of thumb is that the latency per 1 Km of fibre cable is 5 microseconds.

21.2.3 Availability

The redundant SAN design, with two paths from each server to the storage device, will give us the opportunity to perform maintenance on it, without downtime of the servers. We will also be able to perform upgrades or replacements without affecting the production servers.

As we have two sites, our SAN is also designed to be redundant in the area of connecting those two sites. We are providing redundant paths for data replication and also for data access in case the storage in the primary site fails.

21.2.4 Security

When implementing the director, we need to address some security issues:

- **Director security**
  - Change the default passwords.

– Either put the directors into an existing separate management network if one is already in place for other functions, or implement a new network.

- **Zoning**

  In our case we have a multi-platform environment. Because we have only one storage port for all servers except SGI, there is no need to implement any zoning. In the future, if we expand the number of storage ports, we can group the servers to the storage ports in separate zones for performance reasons. This can be implemented non-disruptively using software zoning.

  We will have the following zones:

  – A zone for all SGI servers in both sites. This will be implemented using software zoning.

  – A zone for SGI data replication, which will include SGI storage ports for data replication from both sites and ISLs used. This zone will be implemented using soft zoning. In this zone, we will include all the participating ports including the ports for ISLs. With this we will assure that data replication traffic is completely separated from other traffic.

  Note: There are no data integrity issues if you do not implement zoning. In our example, we would only use zoning for performance reasons.

21.2.5 Distance

There are no practical delays in performance introduced locally, by the SAN, as the latency of the directors is in the low microseconds. All the local connections will use shortwave SFPs as the servers are within a radius of 300 m. Even with the fabric expansion, as shown in designs, we should have no problems with the delays in the fabric OS for name server and FSPF changes.

For connections from site-to-site we will use DWDM products. With the use of a DWDM products we will solve the problem of degradation of laser signal, and we can use one pair of fibers for a large number of channels. The CNT SAN256N also allocates enough BB_Credits to all ports to cope easily with this distance.

21.2.6 Scalability

The designs accommodate three years growth.

For the predicted growth we will only accommodate new ports for SGI servers and storage. If Company Two decide to implement 2 Gb/s technology throughout the SAN from the start, then we are also covering growth on all non-SGI servers. If in the future we need more ports, the design is ready to be expanded with new FIO blades in the directors.

21.2.7 "What if" failure scenarios

Here are the "what if" scenarios we considered:

- HBA: If one of the HBA fails, multipath software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the director fails, multipath software will automatically failover workload to the alternate path. If a cable between the director and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: The directors have redundant power supplies as a standard feature.
- Director port: If one of the ports fails, you may replace using a hot-pluggable SFP.
- Director: In the highly unlikely (though technically possible) event that a backplane fails, then that entire director *could* stop working. As we have a redundant fabric at both sites, this need not cause disruption.

21.2.8 Manageability and management software

The management techniques are the same as in Case Study 1, and we described them in 21.1.8, "Manageability and management software" on page 810.

21.3 Case Study 3: Company Three

If we consider the company and its requirements as detailed in 18.3, "Case Study 3: Company Three" on page 697, we will propose the following solution.

21.3.1 Design

Considering the requirements as detailed in 18.3.5, "Analysis (ports and throughput)" on page 699, we selected a solution based around the CNT FC/9000 Director Model 128. The smallest configuration available via IBM is 24 ports.

The design can be seen in Figure 21-8.

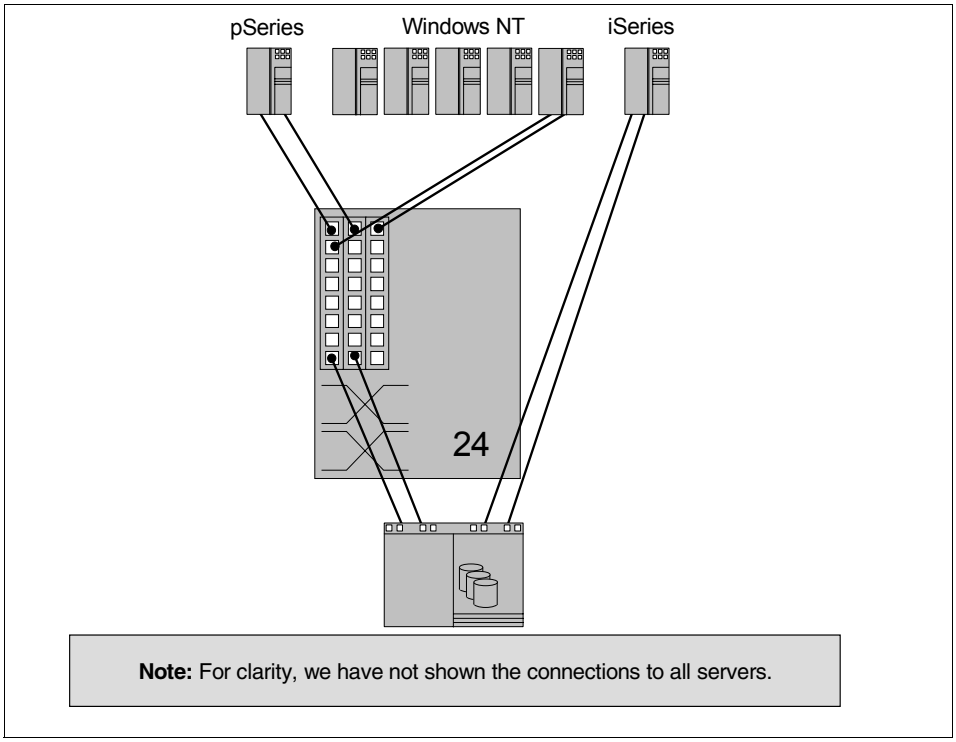


Figure 21-8 The proposed design for Company Three

Note: CNT FC/9000 Directors supports iSeries running OS/400 running V5.2 or later.

There is a summary of the ports and their usage in Table 21-8.

Table 21-8 Port usage at Company Three

Director	Servers	Storage	ISLs	Spare
Director 1	14	2	0	8

With this design we are fulfilling these requirements:

- Very high reliability: Although the director has a passive backplane which is potentially a single point-of-failure, it does offer 99.999% uptime.
- All bandwidth requirements are met.

- SAN components can be upgraded without impact on the servers. CNT FC/9000 directors allows live upgrades to firmware and swapping of FIO blades.
- Possible growth without impact on the production. We can add FIO blades to the FC/9000 without disruption.

In the following sections we will outline some aspects of the design.

21.3.2 Performance

In the initial design we have a 7:1 fan-out ratio for all servers. This means that seven servers are accessing one storage device. There are no latency issues, because the traffic is only going through the director, where the latency is less than 3 microseconds.

21.3.3 Availability

CNT FC/9000 is a director class product. All paths are redundant. These factors allows us to perform maintenance or upgrades of the SAN, without downtime of the servers.

21.3.4 Security

When implementing the director, we need to address some security issues:

- **Director security**
  - Change the default passwords.
  - Either put the directors into an existing separate management network if one is already in place for other functions, or implement a new network.
- **Zoning**

  Although we are accessing a single storage, it is much appropriate to group platforms in to different zones.

21.3.5 Distance

As you can see from the designs we are using only one director between the servers and storage. For all the connections we will use shortwave SFPs.

21.3.6 Scalability

By design, we can accommodate at least three years planned growth. We have enough bandwidth allocated for that period. The only additional ports likely to be needed are for storage and we have spares available. There is room in all the directors for additional blades to be added.

All of the above could be achieved without disruption to service.

21.3.7 “What if” failure scenarios.

- Server: Application will not be available. Server has to be replaced or repaired.
- HBA (Open Systems): If one of the HBA fails, multipath software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- HBA (iSeries): Failover support is a built-in feature of iSeries.
- Cable: If a cable between a server and the director fails, multipath software will automatically failover the workload to the alternate path. If a cable between the director and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: CNT FC/9000 director has redundant power supplies.
- Director port: If one of the ports fails, you may replace it using a hot-pluggable SFP, without outage or performance problem.
- Director blade: if an entire director blade fails, a redundant path is provided in all cases via another blade. Sufficient spare ports are available to replace a blade without requiring downtime on any servers
- Director: If the director fails then the SAN will not be operational. This is highly unlikely to happen because of the resilient design of the FC/9000. It is expected to have 99.999% uptime. If the tiny risk is deemed to be unacceptable, then a second director could be used to provide a redundant fabric.

21.3.8 Manageability and management software

The management techniques are the same as in Case Study 1, and we described them in “Manageability and management software” on page 810.

21.4 Case Study 4: Company Four

If we consider the company and its requirements as detailed in 18.4, “Case Study 4: Company Four” on page 700, we will propose the following solution.

21.4.1 Design

If we consider the requirements as detailed in 18.4.5, “Analysis (ports and throughput)” on page 703, we selected a solution based on the CNT SAN256N (or IBM 2045-N16). A feature of the CNT SAN256N is that it can support Fibre Channel and FICON ports on the same director. Not only that, but the protocols can be mixed on the same blades. This allows maximum flexibility. We are able to consolidate all storage and accommodate the storage and all the servers into a single SAN infrastructure. The SAN256N physical director can also be partitioned on a blade basis, into up to four logical domains, or switches. A logical domain consists of two or more I/O modules. Partitioning the director into logical domains can help simplify SAN consolidation by isolating different protocols, like mainframe FICON and open systems Fibre Channel, and different applications such as SAP and storage management systems. So, in our case, if a complete separation of open systems and FICON is desired (for management and security reasons), the SAN services feature (FC #7303) can be implemented.

The proposed solution for the East site is shown in Figure 21-9.

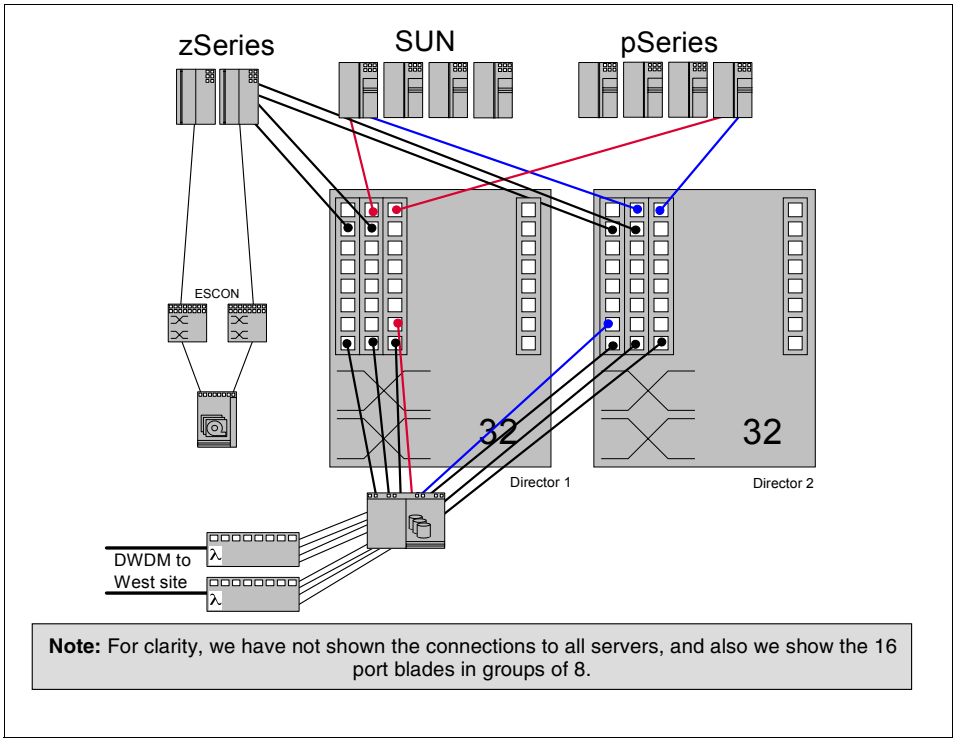


Figure 21-9 The proposed solution for the East site

There is a summary of the ports and their usage in Table 21-9.

Table 21-9 Port usage at the East site

Director	Servers	Storage	ISLs	Spare
Director 1	12	4	0	16
Director 2	12	4	0	16
Total	24	8	0	32

The proposal is to use four FICON connections, from the zSeries, two via each director. This is for absolute resilience and redundancy. The bandwidth requirement is for there to be two plus one for resilience. In order to save cost, the implementation *could* use just three connections.

It should be noted that the DWDM connections carry both FICON and ESCON traffic.

The proposed solution for the West site is shown in Figure 21-10.

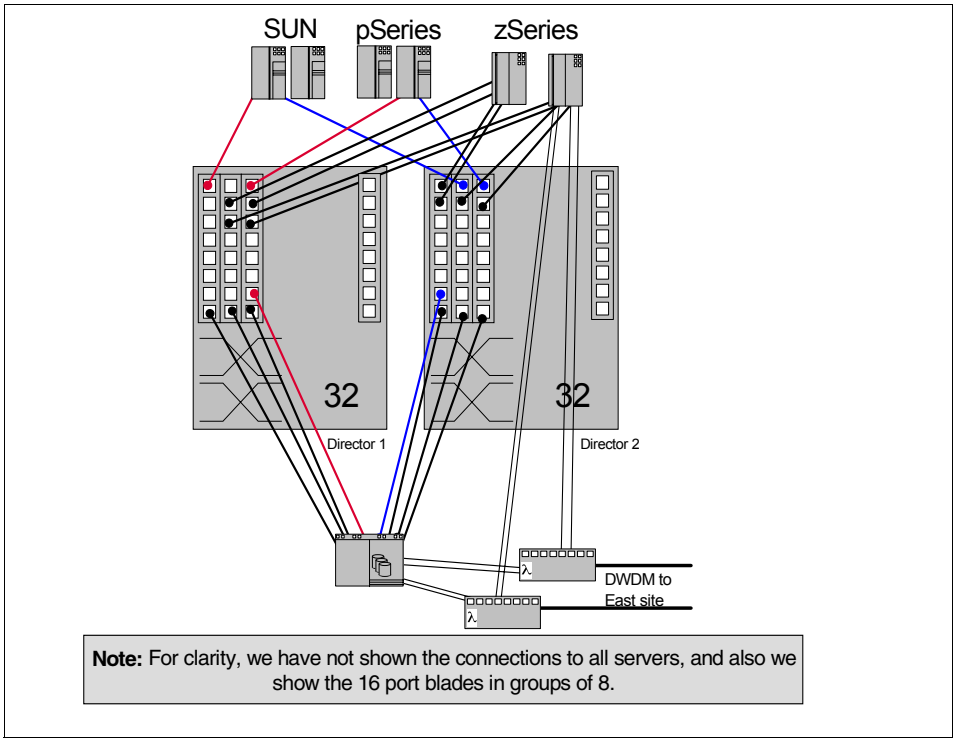


Figure 21-10 The proposed solution for the West site

There is a summary of the ports and their usage in Table 21-10.

Table 21-10 Port usage at the West Site

Director	Servers	Storage	ISLs	Spare
Director 1	12	4	0	16
Director 2	12	4	0	16
Total	24	8	0	32

The FICON connections from zSeries to the DWDM units could go via the directors, but this would involve using an extra eight director ports.

The combined solution for both sites is shown in Figure 21-11.

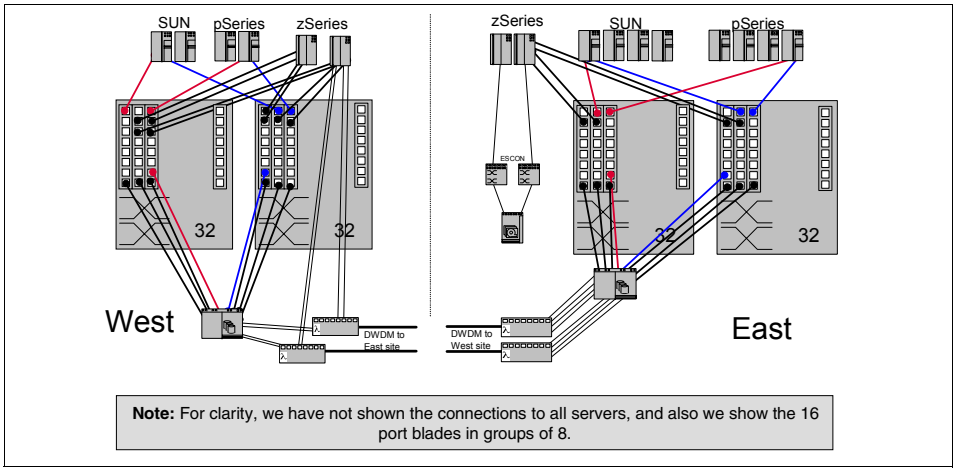


Figure 21-11 The combined solution

With such a design we are fulfilling the requirements:

- No single point-of-failure, we have a fully redundant SAN.
- All bandwidth requirements are met.
- SAN components can be upgraded without impact on the servers. The CNT FC/9000 director allows us to upgrade Firmware non disruptively.
- Possible growth without impact on the production. The CNT FC/9000 director allows us to add FIO blades non disruptively.

For the storage data replication we are using ESCON links over DWDM connection. The peak bandwidth used by the servers is 11.58 MB/s. As only 30% of this peak are writes 3.48, we will allocate two ESCON channels for the data replication.

In the following sections, we outline some aspects of the design.

21.4.2 Performance

In the initial design we have a 6:1 fan-out ratio for all servers in the East site. This means that six servers are accessing one storage port. The fan-out ratio in the West site is 4:1. There are no latency issues, because the traffic is only going trough the directors, where the latency is up to below 3 microseconds.

21.4.3 Availability

The FC/9000 offers 99.999% uptime. The redundant SAN design, with two paths from each server to the storage device, will give us the opportunity to perform maintenance, without downtime of the servers. We will also be able to perform upgrades or replacements without affecting production.

21.4.4 Security

When implementing the director, we need to address some security issues:

- **Director security**
  - Change the default passwords.
  - Either put the directors into an existing separate management network if one is already in place for other functions, or implement a new network.
- **Zoning**

  In this case we will use software zoning to isolate the FICON traffic from the FC traffic.

21.4.5 Distance

As can be seen from the designs we are only using one director between the server and the storage in each site. This means that there is no practical delay in the performance. All the local connections will use shortwave SFPs.

There are no ISLs from site-to-site. The DWDM connections between the two sites will utilize the fiber optic connections.

21.4.6 Scalability

By design, we can accommodate at least three years planned growth. We have enough bandwidth allocated for that period. The only additional ports likely to be needed are for storage and we have spares available. There is room in all the directors for additional blades to be added.

All of the above could be achieved without disruption to service.

21.4.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: The clustering solution will failover to the passive server dynamically.

- HBA: If one of the HBA fails, multipath software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and a director fails, multipath software will automatically failover workload to the alternate path. If a cable between the director and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: The FC/9000 has redundant power supplies.
- Director port: If one of the ports fails, you may replace using a hot-pluggable SFP, without outage or performance problem.
- Director blade: if an entire director blade fails, a redundant path is provided in all cases via another blade. Sufficient spare ports are available to replace a blade without requiring downtime on any servers
- Director: If a director fails then it can be replaced while the other takes the full load. This is highly unlikely to happen because of the resilient design of the FC/9000. It is expected to have 99.999% uptime.

21.4.8 Manageability and management software

The management techniques are the same as in Case Study 1, and we described them in “Manageability and management software” on page 810.

21.5 Case Study 5: Company Five

If we consider the company and its requirements as detailed in 18.5, “Case Study 5: Company Five” on page 705, we will propose the following solution.

21.5.1 Design

Considering the requirements as detailed in 18.5.5, “Analysis (ports and throughput)” on page 708, we selected a solution based on CNT FC/9000 Director Model 128. We are using a 24-port director as this is the smallest configuration available via IBM.

The proposed design is shown in Figure 21-12.

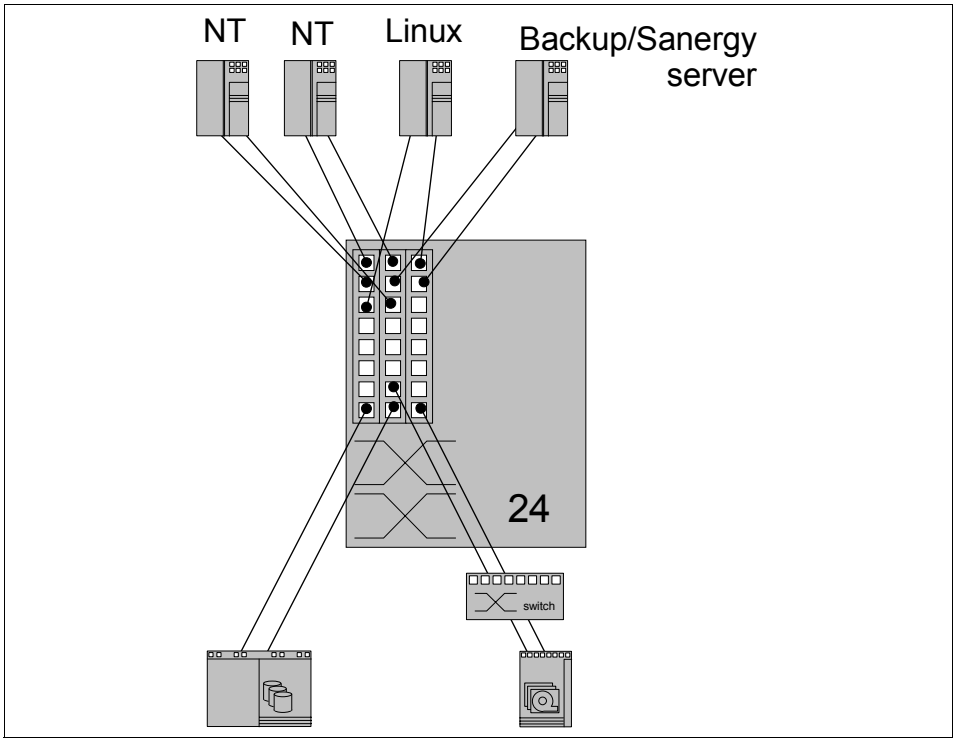


Figure 21-12 The proposed design for Company Five

There is a summary of the ports and their usage in Table 21-11.

Table 21-11 Port usage at Company Five

Director	Servers	Storage	ISLs	Spare
Director 1	8	4	0	12

With such a design we are fulfilling these requirements:

- High availability due to the director class product offering 99.999% uptime.
- All bandwidth requirements are met.
- SAN can be upgraded without impact on the servers. The FC/9000 allows live upgrades to firmware and swapping of FIO blades.
- SAN can be upgraded without impact on the servers. The FC/9000 allows FIO blades to be added on the fly.
- Because the director do not support FC-AL connectivity, we use an IBM TotalStorage SAN Switch F16 that has Quickloop feature to support FC-AL to connect tape devices to the SAN.

Note: We chose a single director solution because it offers 99.999% uptime, and the company would find it hard to justify the additional cost of a dual fabric Director solution.

We have introduced a fourth server to accommodate heterogeneous file sharing (using Tivoli SANergy) and also backup/restore software (using Tivoli Storage Manager). After the migration from NT to Linux is complete, we could consider using the vacant NT server, thus reducing the number of servers to three.

In the following sections, we will outline some aspects of the design.

21.5.2 Performance

In the initial design, we have a 4:1 fan-out ratio for all servers accessing the storage. This means that four servers are accessing one storage port. All server can also access the tape device, which was the requirement for LAN-free backup implementation. There are no latency issues, because the traffic is only going trough the director, where the latency is less than 3 microseconds.

21.5.3 Availability

All servers and the storage have multiple paths giving resilience through redundancy. The director class FC/9000 should give 99.999% uptime. In the event that the single backplane is of concern, then a dual fabric SAN could easily be implemented by using a second director, although, in this case, the additional cost might be prohibitive.

21.5.4 Security

When implementing the director, we need to address some security issues:

- **Director security**
  - Change the default passwords.
  - Either put the directors into an existing separate management network if one is already in place for other functions, or implement a new network.
- **Zoning**

  Because we have only one storage port for all servers, there is no requirement to implement any zoning. In the future, if we expand the number of storage ports, we can group the servers to the storage ports in separate zones for performance reasons. This can be implemented non-disruptively using software zoning.

Note: There are no data integrity issues if you do not implement zoning. In our example, we would only use zoning for performance reasons.

21.5.5 Distance

As can be seen from the designs, we are using only one director between the servers and storage. All the connections we will use shortwave SFPs.

21.5.6 Scalability

From the designs we accommodated three years growth. We have enough bandwidth allocated for all requirements in the next three years. The only expansion will probably be to add additional storage ports, and we have provided enough spare ports for this. If, in the future we need more ports, then the FC/9000 is ready to be expanded with new FIO blades.

This would be a non-disruptive upgrade.

21.5.7 “What if” failure scenarios

- Server: Application will not be available. Server has to be replaced or repaired.
- HBA: If one of the HBA fails, multipath software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multipath software will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: The FC/9000 has redundant power supplies.
- Director port: If one of the ports fails, you may replace it using a hot-pluggable SFP, without outage or performance problem.
- Director blade: If an entire director blade fails, a redundant path is provided in all cases via another blade. Sufficient spare ports are available to replace a blade without requiring downtime on any servers
- Director: If the director fails then the Primary site will not be operational. This is highly unlikely to happen because of the resilient design of the FC/9000. It is expected to have 99.999% uptime. If the tiny risk is deemed to be unacceptable, then a second director could be used to provide a redundant fabric at the Primary site.

21.5.8 Manageability and management software

The management techniques are the same as in Case Study 1 and we described them in “Manageability and management software” on page 810.

21.6 Case Study 6: Company Six

If we consider the company and its requirements as detailed in 18.6, “Case Study 6: Company Six” on page 710, we will propose the following solution.

21.6.1 Design

Considering the requirements as detailed in 18.6.5, “Analysis (ports and throughput)” on page 712, we selected a solution based on CNT FC/9000 director.

You can see the proposed design for Primary site in Figure 21-13.

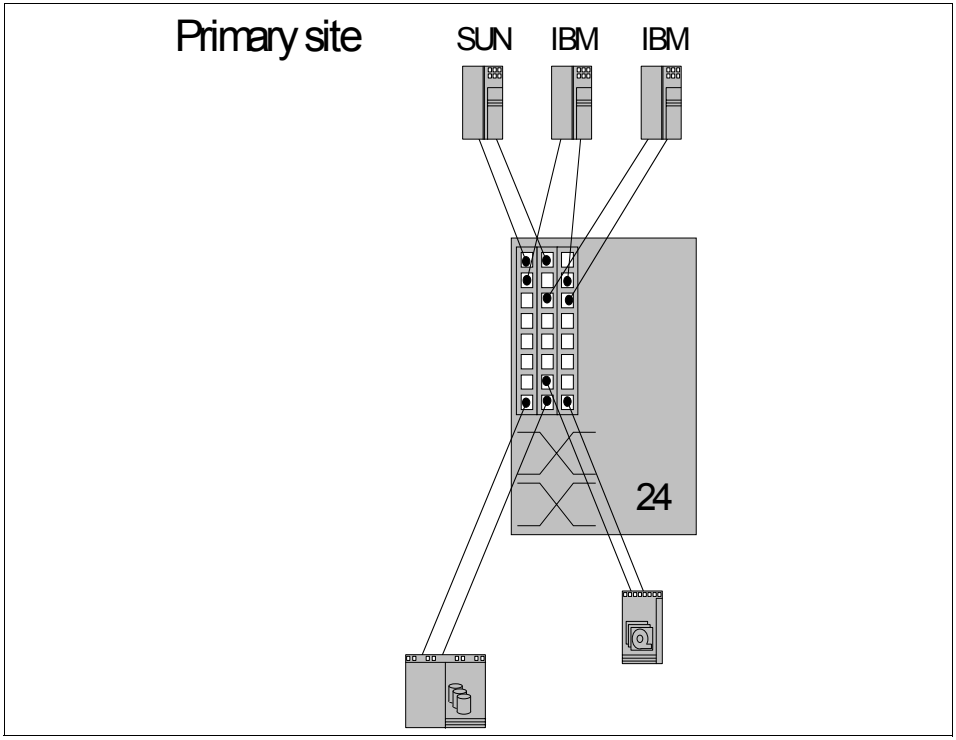


Figure 21-13 The proposed design for the Primary site

In the secondary site we already have a SAN, using Compaq StorageWorks SAN Switch 16. These are Brocade OEM switches of the SilkWorm 2800.

We show the proposed design of Secondary site in Figure 21-14.

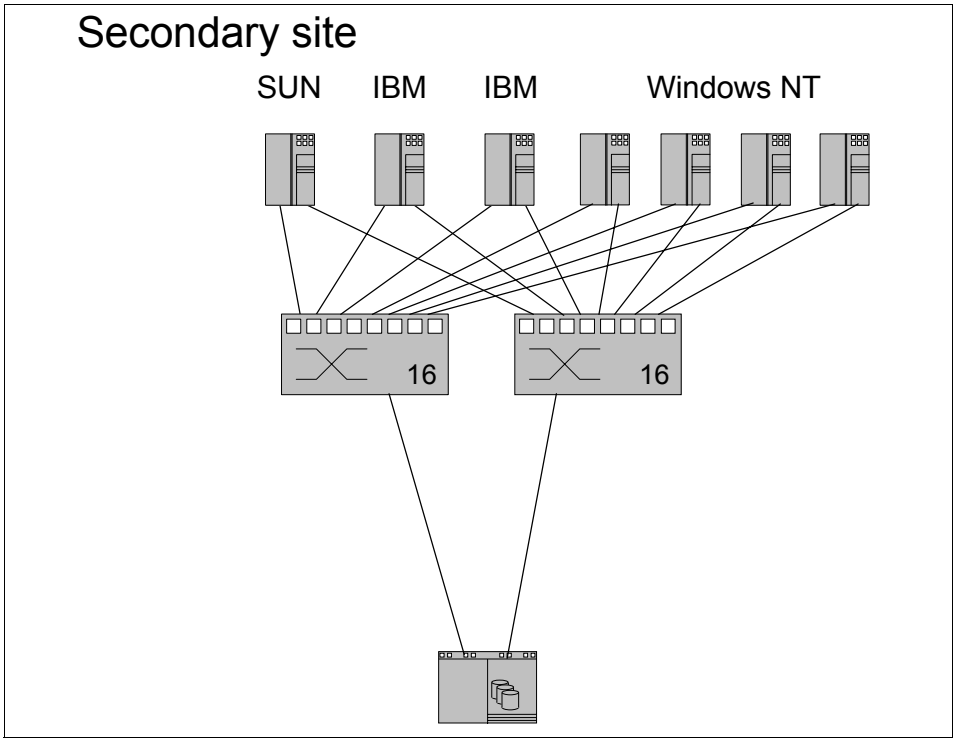


Figure 21-14 The proposed design for the Secondary site

Both sites are connected with two ISLs. Because the distance is 900 miles, we will use DWDM products to connect them.

We show the connection setup in Figure 21-15.

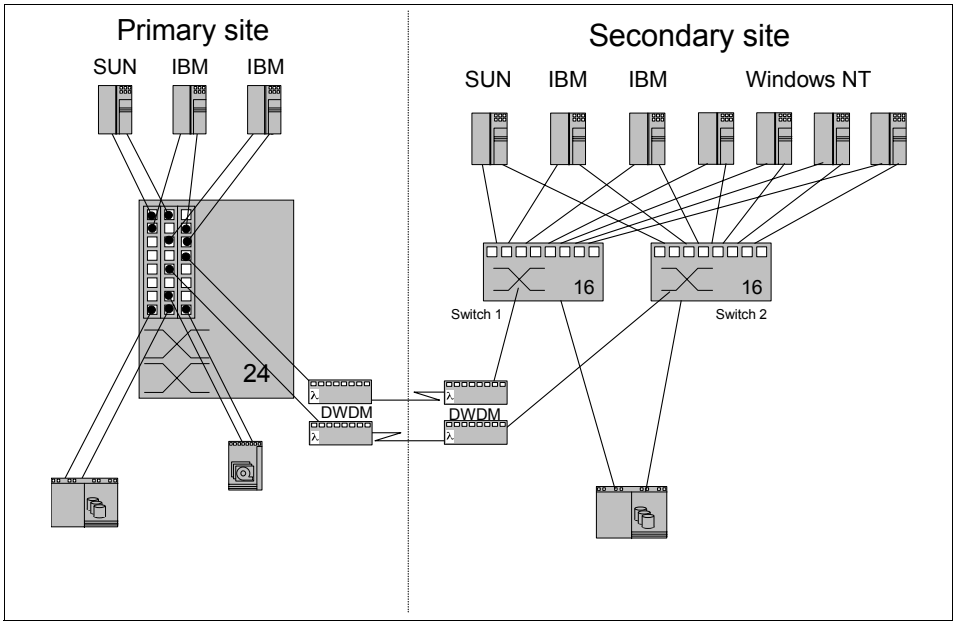


Figure 21-15 The complete solution

There is a summary of the ports and their usage in Table 21-12 and Table 21-13.

Table 21-12 Port usage at the Primary site

Director	Servers	Storage	ISLs	Spare
Director 1	6	4	2	12

Table 21-13 Port usage at the Secondary site

Director	Servers	Storage	ISLs	Spare
Switch 1	7	1	1	7
Switch 1	7	1	1	7
Total	14	2	2	14

With such a design we are fulfilling these requirements:

- Very high reliability: Director class product at Primary site offers 99.999% uptime. No single point-of-failure at Secondary site- redundant SAN.
- All bandwidth requirements are met.
- SAN components can be upgraded without impact on the servers. At the Primary site, the FC/9000 allows live upgrades to firmware and swapping of

FIO blades. At the Secondary site, we can first upgrade one switch, while paths across the second one are still available. After the traffic is established back on the upgraded switch, we can upgrade second switch.

- Possible growth without impact on the production. At the Primary site we can add FIO blades to the FC/9000. As we are using a redundant SAN at the Secondary site, we can introduce additional switches without the downtime on the existing servers.
- We are reusing existing equipment (switches and tape).
- We are providing storage replication to remote site for disaster recovery.
- We are improving backup performance by providing infrastructure for LAN-free backup.

In the following sections we outline some aspects of the design.

21.6.2 Performance

In the initial design, we have a 3:1 fan-out ratio for all servers accessing the storage on the Primary site. This means that three servers are accessing one storage port. All servers can also access the tape device, which was the requirement for LAN-free backup implementation. There are no latency issues, because the traffic is only going through the director, where the latency in the low microseconds. On the Secondary site we have 7:1 fan-out ratio. In this case, the traffic is only going through the switches and again the latency is so low as to be of no significance.

21.6.3 Availability

The FC/9000 at the Primary site, is a director class product. The SAN design, at the Secondary site is fully redundant. All paths are redundant. These factors allows us to perform maintenance or upgrades of the SAN, without downtime of the servers.

As we have two sites, our SAN is designed to be redundant also in the area of connecting those two sites. We are providing redundant paths for data replication. It recommended that the connections between two DWDM boxes are also redundant.

21.6.4 Security

When implementing the director, we need to address some security issues:

- **Director security**
  - Change the default passwords.

- Either put the directors into an existing separate management network if one is already in place for other functions, or implement a new network.

- **Zoning**

  In our case we have a multi platform environment. Because we have only one storage port for all servers there is no need to implement any zoning. In the future, if we expand the number of storage ports, we can group the servers to the storage ports in separate zones for performance reasons. This can be implemented non-disruptively using software zoning.

  Note: There are no data integrity issues if you do not implement zoning. In our example we would only use zoning for performance reasons.

21.6.5 Distance

As can be seen from the designs, we are only using one director or switch between the server and the storage in each site. This means that there is no practical delay in the performance. All the local connections will use shortwave SFPs.

For connections from site-to-site we will use Longwave SFPs, which will be connected via DWDM products. With the use of a DWDM product, we will solve the degradation of laser signal, but we still need to adjust the buffers needed on switch ports for this type of connection. On the switches at the Secondary site, we need to enable the Extended Fabrics feature in the switches.This assigns more buffer credits to those E_Ports being used by long distance ISLs.

21.6.6 Scalability

From the designs, we accommodated three years growth. We have enough bandwidth allocated for all requirements in the next three years. The only expansion will be probably adding additional storage ports, and we accommodated the space for this. If in the future we would need more ports, the design is ready to be expanded with new switches.

21.6.7 "What if" failure scenarios

Here are the "what if" scenarios we considered:

- Server: Application will not be available. Server has to be replaced or repaired.
- HBA: If one of the HBA fails, multipath software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the director/switch fails, multipath software will automatically failover workload to the alternate path. If a cable between the director/switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: Another redundant power supply may be added to the switch, and should one fails, other will take over automatically. The FC/9000 has redundant power supplies.
- Switch or director port: If one of the ports fails, you may replace using a hot-pluggable SFP, without outage or performance problem.
- Switch: If a switch fails, the server will use the alternate switch to connect to the storage, without outage or performance problem.
- Director blade: If an entire director blade fails, a redundant path is provided in all cases via another blade. Sufficient spare ports are available to replace a blade without requiring downtime on any servers
- Director: If the director fails then the Primary site will not be operational. This is highly unlikely to happen because of the resilient design of the FC/9000. It is expected to have 99.999% uptime. If the tiny risk is deemed to be unacceptable, then a second director could be used to provide a redundant fabric at the Primary site.

21.6.8 Manageability and management software

The management techniques are the same as in Case Study 1 and we described them in 21.1.8, "Manageability and management software" on page 810.

The DWDM products would also need to be integrated into the management infrastructure.

It would be possible to manage the switches using Tivoli Storage Network Manager.

22

Cisco solutions

In this chapter, we will show solutions to the case studies based on the products in the IBM Cisco portfolio.

© Copyright IBM Corp. 2004. All rights reserved.

22.1 Case Study 1: Company One

If we consider the company and its requirements as detailed in 18.1, “Case Study 1: Company One” on page 686, we will propose the following solution.

22.1.1 Design using directors

Considering the requirements as detailed in 18.1.5, “Analysis (ports and throughput)” on page 687, we have proposed in this section a director-based solution for our design.

We show the proposed design in Figure 22-1 using a Cisco MDS 9506 director with 32 ports initially configured. The Cisco MDS 9506 director can have a maximum of 128 ports.

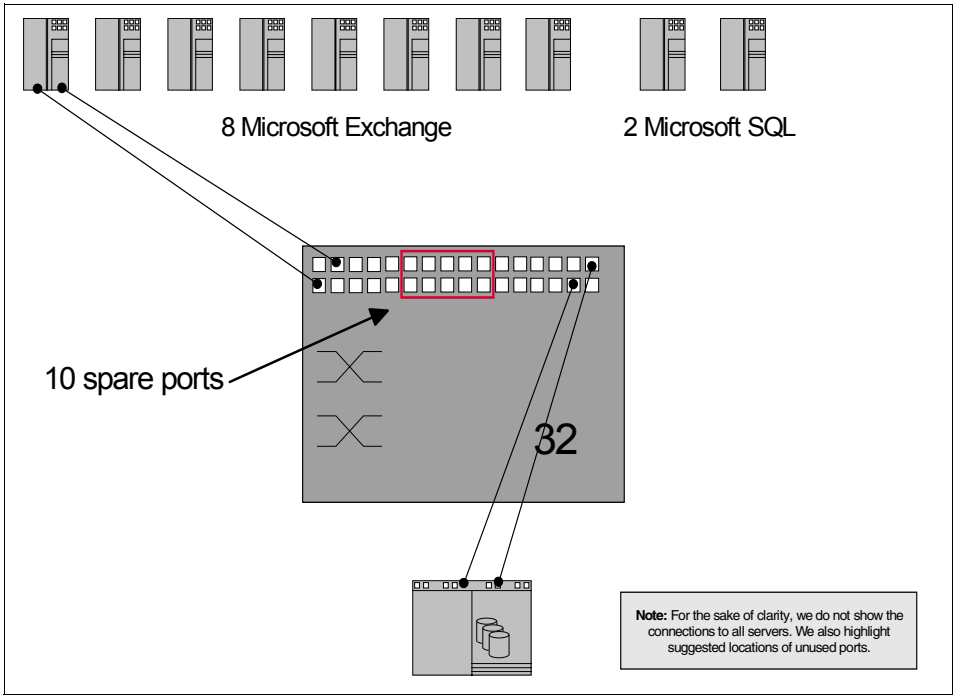


Figure 22-1 Core SAN design using a Cisco MDS 9506 Director

As we are using two paths from each of the ten servers to the storage, we are providing redundancy and high availability. To utilize the dual connections, we will use multipathing software (such as IBM’s Subsystem Device Driver- SDD) on each of the servers. The solution shown has been configured with 32 ports in the director initially. With such a design, we are partially fulfilling the requirements:

- Fault resilient SAN with 99.999% availability.
- All bandwidth requirements are met (40 KB/s and 4 MB/s from servers and 8.32MB/s to storage).
- The director can be upgraded (firmware and additional port cards) without impact on the servers. There is no requirement to bring the director down for firmware upgrades, applications will continue to perform I/O during this operation, but may experience a minimal delay. No re-routing and/or cabling changes need to be made to accommodate this.
- Possible growth without impact on the production. Due to the nature of the director, we can add additional blades concurrently. We can also introduce additional director(s) for a fully redundant SAN.

With this design we have ten SAN ports free for future expansion. This means that you could potentially connect an additional five servers with redundant connections.

Care should be taken to ensure that primary and secondary connections from one server are not attached to the same blade. If a blade does fail, the alternate path should be available through another blade.

Although the solution presented provides 99.999% availability and is fault tolerant, it is not fully redundant as there is a single passive backplane. To overcome this we can introduce another director. This will give us the capability of expanding usable ports in two devices. During the re-cabling process, be sure to identify the secondary paths (cables) from each server and re-connect to the second director. During this process, there will be single points of failure for each server.

In the following pictures we will show how you can expand your fabric without impact on the production servers. As this was the design requirement, we will show that our design is capable of handling this. In the second year, we should expect to accommodate 20 servers, with the same I/O characteristics.

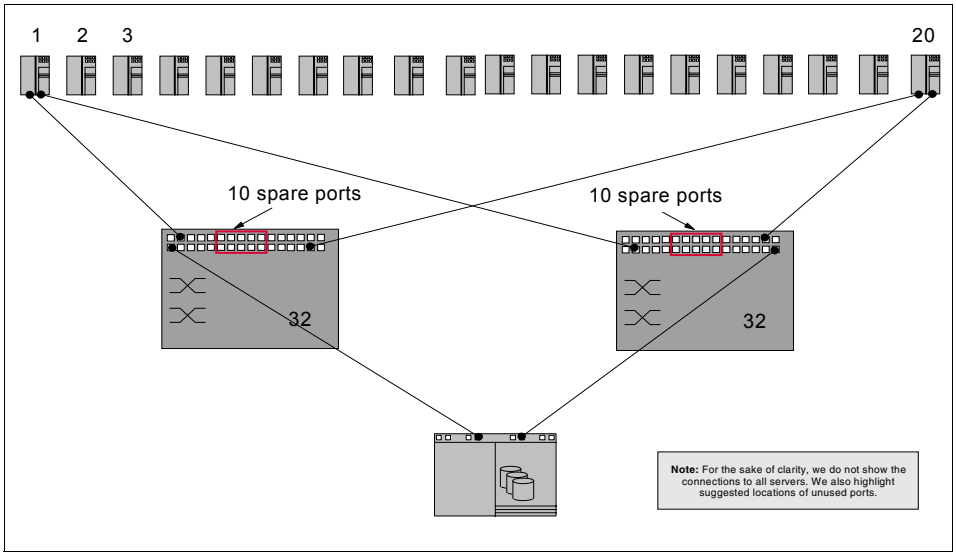


Figure 22-2 Fully redundant Cisco MDS 9506 Director solution.

In Figure 22-2, each blade has five free ports, which gives a total of ten free ports per director, This will accommodate the failure of a complete blade, and also allow for a port to be used for maintenance. It also allows for a possible expansion of ten additional servers with redundant paths.

With the uncertainty surrounding the growth of the complex into the third year, should the additional 20 servers come to fruition, then the total number of ports required could increase to 84. The solution proposed below will accommodate this with the simple non-disruptive approach of adding port blades, as shown Figure 22-3.

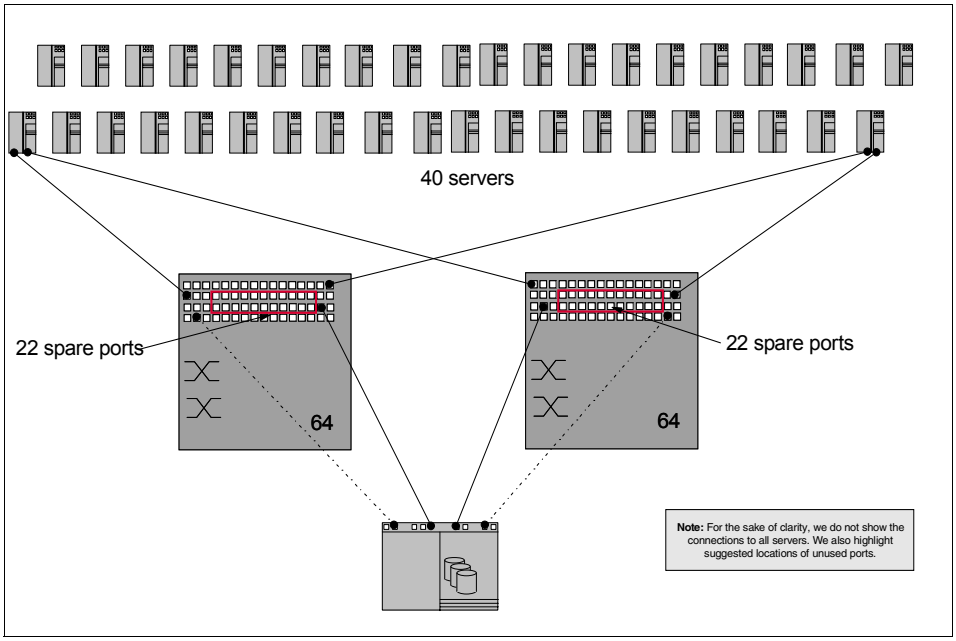


Figure 22-3 Cisco MDS 9506 solution with all potential servers

Spare ports are still available for maintenance functions. It is also worth noting that the port blade cards can be moved from one director to another, should you wish to use one director initially, populate it, and then introduce the second director. Additional port expansion capabilities still exist with a maximum port count of 128 per director.

Note: We have doubled the number of connections to the storage device to reduce the fan-out ratio to 20:1. This is not due to any bandwidth requirement, but solely due to the number of connections that must be handled.

We assume the same bandwidth requirements for the new servers as the original servers.

In Table 22-1 we show Year 1 requirements.

Table 22-1 Year 1 ports: 10 servers

Storage	Server	ISL	Spare
2	20	0	10

In Table 22-2 we show Year 2 requirements.

Table 22-2 Year 2 ports: 20 servers

Storage	Server	ISL	Spare
2	40	0	20

In Table 22-3 we show Year 3 requirements.

Table 22-3 Year 3 ports: 40 servers

Storage	Server	ISL	Spare
4	80	0	44

In the following sections, we will outline some aspects of the design.

22.1.2 Performance

As we can see from the design, we have an initial fan-out ratio of 10:1 for each server accessing a storage device port. This means that 10 servers are accessing the same storage pool. As we only need 8.32 MB/s for the servers we have enough bandwidth on the storage port for all servers. By adding the next 10 servers, we will get a fan-out ratio of 20:1. The bandwidth requirements will be 16.64 MB/s. The same ratio will be maintained if we double the number of servers by doubling the number of storage ports used.

22.1.3 Availability

The Cisco MDS 9506 single director solution provides 99.999% availability. This number will be effectively somewhat higher in a dual director configuration. This SAN design, with two paths from each server to the storage device, will give us the ability for each server to experience an HBA failure without impacting performance and without downtime of the servers.

As we are using a director solution, we will also be able to perform upgrades (hardware and software) or replacements concurrently, that is without affecting the production servers. This is the case, whether we use one director or two.

22.1.4 Security

When implementing the directors we need to take into account some security matters:

- **Director security**
  - Change the default passwords to access the director(s).

- Put director(s) in a separate management network if one is already in place for other functions.

- **Zoning**

  In our case, we have only one platform (Windows). Because we have only one storage port for all servers there is no need to implement any zoning. If in the future we would expand the number of storage ports, we can group the servers by the storage ports separately, then into zones for performance reasons.

  Note: There are no data integrity issues if you do not implement zoning. In our example we would only use zoning for performance reasons.

22.1.5 Distance

In all the connections, we will use shortwave SFPs as the servers are within a radius of 300 m.

22.1.6 Scalability

The designs accommodate three years growth. The growth can be achieved without any interruptions in the production environment. If there is additional growth, we can simply add more port and blades until we hit the maximum of 256 ports of the two directors, at which point we can introduce a third director.

22.1.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: The clustering solution will failover to the passive server dynamically.
- HBA: If one of the HBA fails, multi-path software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the director fails, multi-path software will automatically failover workload to the alternate path. If a cable between the director and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: director class solutions have dual power supplies as a standard feature and no disruption to service will occur should one fail. The replacement power supply can be installed with no outage incurred.
- Director port: If one of the ports fails, you may replace it using a hot-pluggable SFP.

- Director blade: If a director port blade fails, the alternate path will be used. The port on the failed blade can be dynamically plugged into the spare ports in other blades until the blade can be replaced.
- Director: The director is a 99.999% type solution, it is therefore improbable that it will fail. Should it fail there will be no impact if a dual director solution has been implemented.

22.1.8 Manageability and management software

The directors have built-in management capabilities:

- Serial interface to the OS, typically only used to configure the network address
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use

To fully utilize the management features, the directors have to be connected to the network (Ethernet). Although inband management is supported, in our solution we are not using ISLs, and therefore, inband management is not an option.

We show an example of setting up the management network in Figure 22-4.

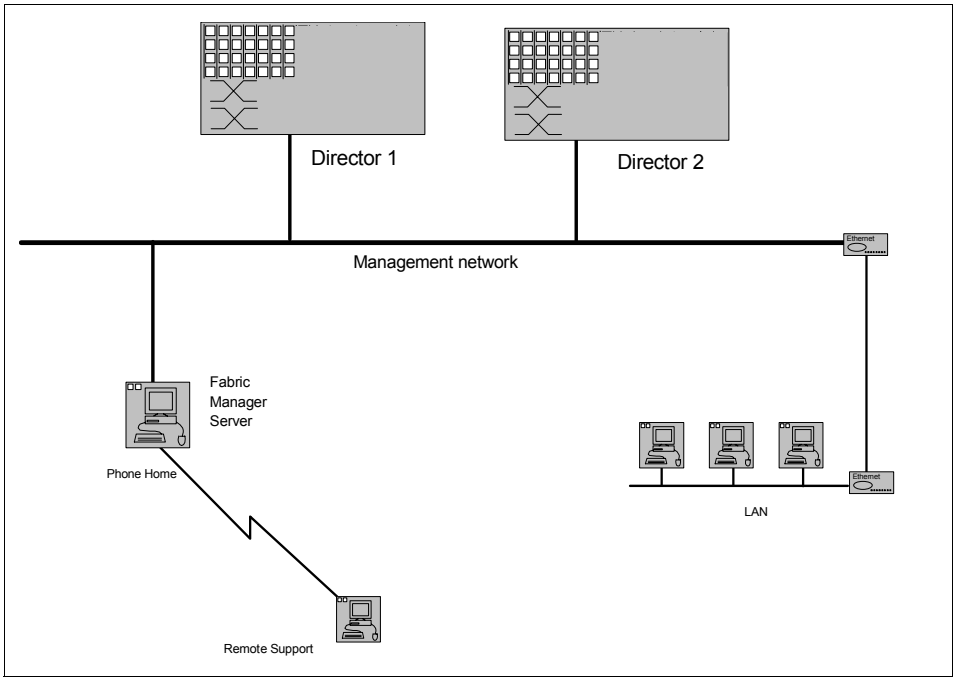


Figure 22-4 Management network

Note: This network can be part of your existing network infrastructure.

22.1.9 Design using switches

Considering the requirements as detailed in 18.1.5, “Analysis (ports and throughput)” on page 687, we have proposed in this section a switch based solution for our design.

We show the proposed design in Figure 22-5.

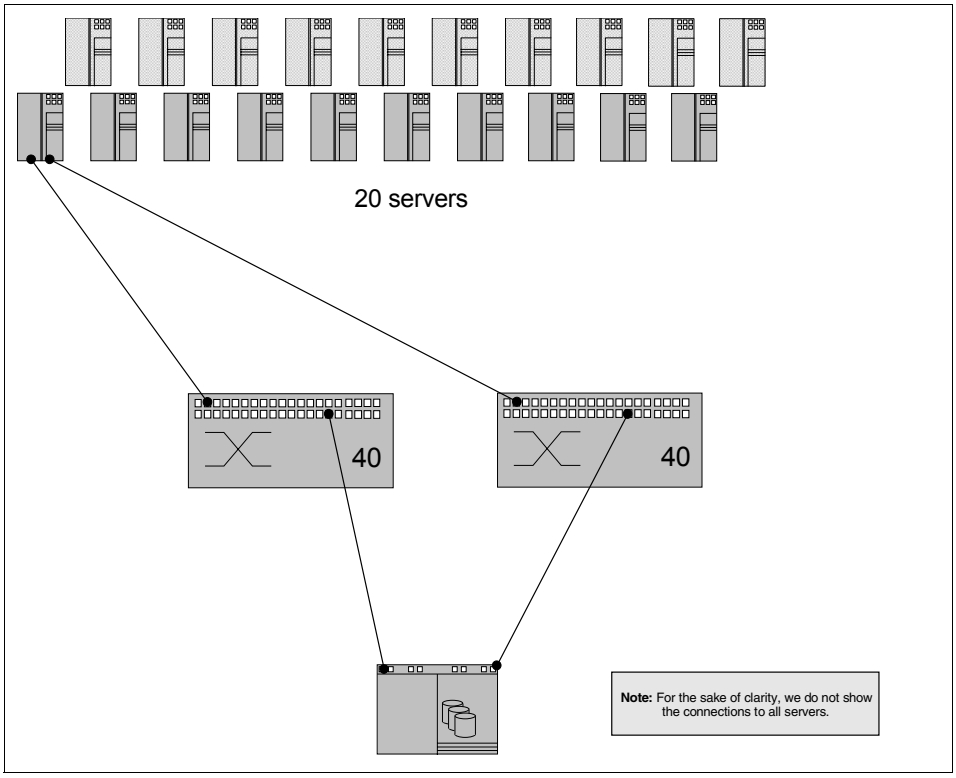


Figure 22-5 Initial design using Cisco MDS 9140 Switches

As we are using two paths from each server to the storage, we are providing redundancy and high availability. To utilize the dual connections, we will use multipathing software (such as IBM's Subsystem Device Driver) on each of the servers. The solution shown has been configured with two 40-port switches. With this design we are fulfilling these requirements:

- Redundant SAN with two switches and all servers dual connected
- All bandwidth requirements are met (40 KB/s and 4 MB/s from servers and 8.32 MB/s to storage).
- Future growth by adding more switches can be accomplished without service interruption.

With this design, we have thirty-eight SAN ports free for future expansion. This means that you could potentially connect an additional nineteen servers with redundant connections.

To expand the fabric to accommodate all 40 servers we must introduce another 40-port switch. This will give us a total of 96 ports, of which three will be used for connectivity to storage and 80 will be used for servers (dual connected).

Special consideration needs to take place when introducing the third switch to avoid bottlenecks and evenly distribute the connections from the servers to the switches. The outcome required should represent 40 servers dual connections spread across the three switches, this equates to 27:27:26 as the distribution.

As you can see, this design will only be truly balanced if the number of servers is divisible by three. In this case study, the performance requirements are quite low and we do not expect there to be an issue. However, as the number of servers increases, the relative fan-out ratio should negate any ill effect this may have on performance.

As we left the two switch implementation there were 20 connections per switch. Introducing the third switch will accommodate all 20 primary connections for the new servers. The new servers will need a total of 20 secondary connections across the original two switches which can be achieved as we are only using 21 of the 40-ports per switch (that is (40-20-1)*2=22).

However, this will leave the used port configuration as 30:30:20 which will leave an unbalanced switch-storage link utilization.

To resolve this we need to take three secondary links from each of the original switches/servers to the new switch. Although the cables will need to be physically unplugged and then plugged, no outage should occur as we are using multi-path software that can use the existing primary path. It may be wise to perform the migration one cable at a time.

The solution is highlighted in Figure 22-6.

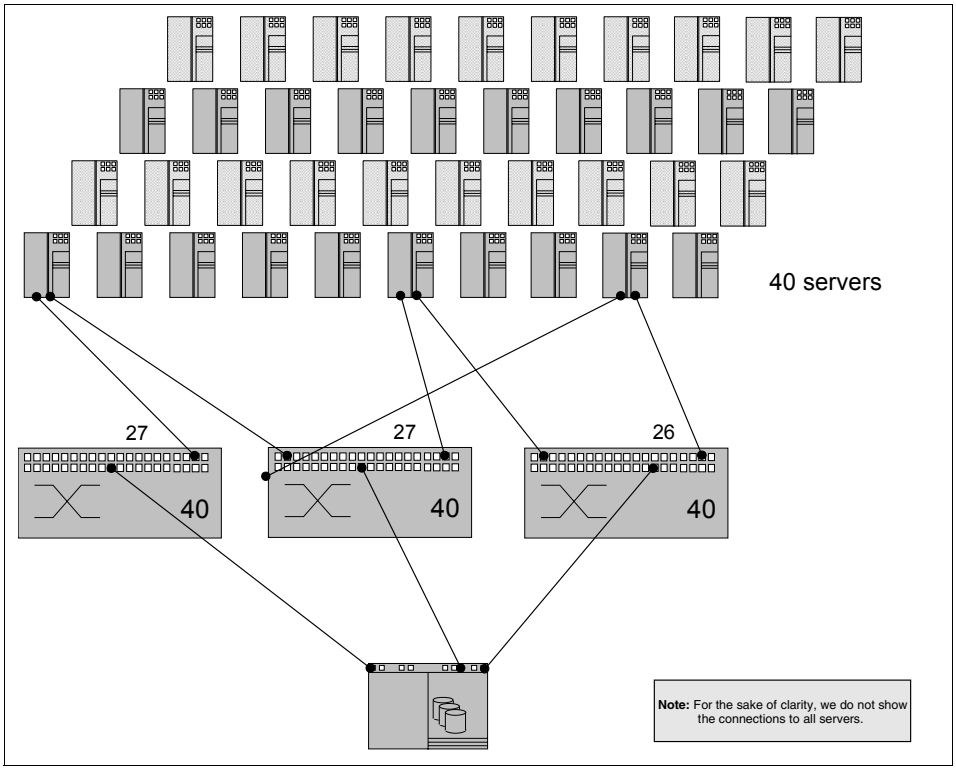


Figure 22-6 Final design to accommodate all potential servers

Important: Ensure that only secondary connections are identified and unplugged. If a primary and secondary link from the same server are removed, then that server will no longer be able to perform I/O across the SAN.

The process described above uses the least effort to achieve the desired results, other considerations and preferences for switch locations and server groups could require a more complex re-cabling effort. The use of a fiber patch panel may aid with cable relocation.

The final configuration, assuming all potential servers are added to the SAN will leave a total of 37 ports free for expansion.

In Table 22-4 we show Year 1 ports.

Table 22-4 Year 1 ports: 10 servers

Storage	Server	ISL	Spare
2	20	0	38

In Table 22-5 we show Year 2 ports.

Table 22-5 Year 2 ports: 20 servers

Storage	Server	ISL	Spare
2	40	0	58

In Table 22-6 we show Year 3 ports.

Table 22-6 Year 3 ports: 40 servers

Storage	Server	ISL	Spare
3	80	0	37

We also took these aspects into account.

22.1.10 Performance

As we can see from the design we have an initial fan-out ratio of 10:1 for each server accessing a storage device port. This means that 10 servers are accessing the same storage pool. As we only need 8.32 MB/s for the servers we have enough bandwidth on the storage port for all servers. By adding the next 10 servers we will get a fan-out ratio of 20:1. The bandwidth requirements will be 16.64 MB/s. As we introduce the third switch and servers the fan-out ratio changes somewhat. As we mentioned earlier, the end result will mean that we have fan-out ratios of 26:1 and 27:1, depending on which server is connected to which switch. Only when the number of servers is exactly divisible by three will we achieve equal fan-out ratios. In this scenario, we do not believe this to be of concern as the I/O throughput requirements are low.

22.1.11 Availability

This SAN design, with two paths from each server to the storage device, will give us the ability for each server to experience an HBA failure without impacting performance, and without downtime of the servers.

The Cisco MDS 9140 switch has a dual power supply that provides redundancy and high availabiltiy in case there is a power supply outage.

It should be noted that when introducing the third switch and performing the re-cabling functions, there will temporarily be single points of failure for six servers as the secondary link is migrated to the new switch.

22.1.12 Security

When implementing the switches we need to take into account security matters:

- **Switch security**
  - Change the default passwords to access the switches.
  - Put switches in a separate management network if one is already in place for other functions.
- **Zoning**

  In our case, we have only one platform (Windows). Because we have only one storage port for all servers there is no need to implement any zoning.

22.1.13 Distance

In all the connections, we use shortwave optics as the servers are within a radius of 300 m. It is possible to move some servers further away by using longwave optics for extending the fabric.

22.1.14 Scalability

The designs accommodate three years growth. The growth can be achieved without any interruptions in the production environment. Twelve additional servers with dual connections can be added to the existing environment. If there are additional growth requirements, we can introduce a fourth switch. At this time, we would recommend that we re-design the cable connections to ensure even distribution.

22.1.15 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: The clustering solution will failover to the passive server dynamically.
- HBA: If one of the HBA fails, multi-path software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multi-path software will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.

- Power supply: The Cisco MDS 9140 solution has dual power supplies as a standard feature and no disruption to service will occur should one fail. The replacement power supply can be installed with no outage incurred.
- Switch port: If one of the ports fails, you may replace the hot-pluggable optic. No outage will occur as the server will use its alternate path.
- Switch: Should the switch fail, the server will use its alternate path through another switch without an outage or performance degradation.

22.1.16 Manageability and management software

The switches have built-in management capabilities:

- Serial interface to the OS, typically only used to configure the network address
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use

To fully utilize the management features, the switches have to be connected to the network (Ethernet). Although inband management is supported, in our solution, we are not using ISLs, and therefore, inband management is not an option.

We show an example of setting up the management network in Figure 22-7.

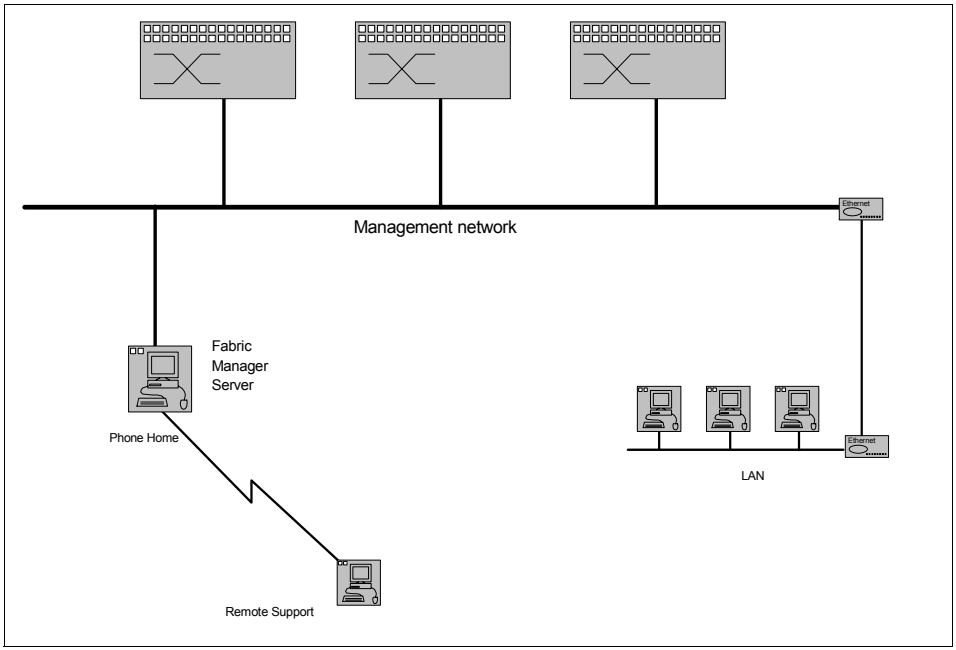


Figure 22-7 Management network

Note: This network can be part of your existing network infrastructure.

22.2 Case Study 2: Company Two

If we consider the company and its requirements, as detailed in 18.2, “Case Study 2: Company Two” on page 689, we will propose the following solution.

22.2.1 Design

Considering the requirements as detailed in 18.2.5, “Analysis (ports and throughput)” on page 692, we have proposed in this section a director based solution for our design.

The configuration for the Getwell facility can be seen in the proposed design in Figure 22-8. In this solution, we are using two Cisco MDS 9506 which have been configured with 98 ports each initially. A Cisco MDS 9506 director can have a maximum of 128-ports.

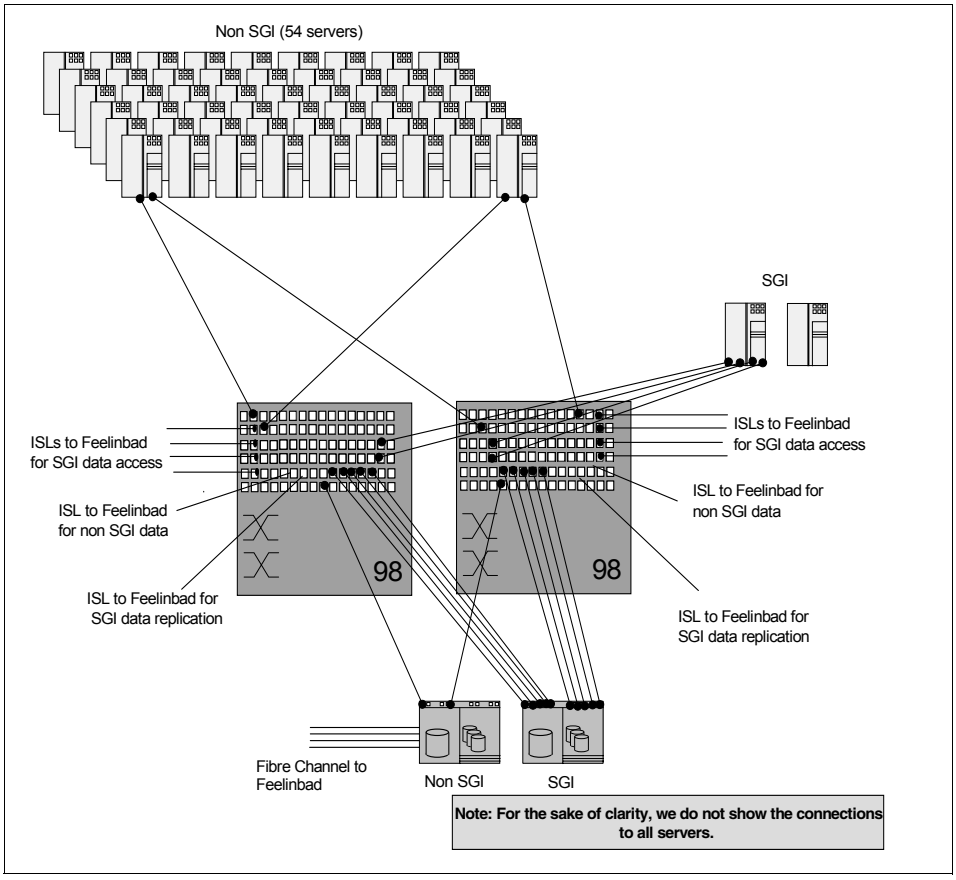


Figure 22-8 Getwell SAN design using Cisco MDS 9506 Director

To summarize the port allocation:

- One hundred and eight edge ports for non-SGI servers (dual connect)
- Two ports for non-SGI Storage
- Eight ports for SGI Servers
- Eight ports for SGI Storage
- Two ports for SGI replication (storage - switch)
- Four Fibre Channel connections from non-SGI storage device to Feelinbad
- Eight ports for ISLs for connecting to Feelinbad
- Two ports for ISLs for SGI replication
- Two ports for ISLs for non-SGI

Which makes for a total of 144 Fibre Channel connections.

Table 22-7 Initial ports

Storage	Server	ISL	Spare
16	116	12	52

The configuration for the Feelinbad facility can be seen in the proposed design in Figure 22-9

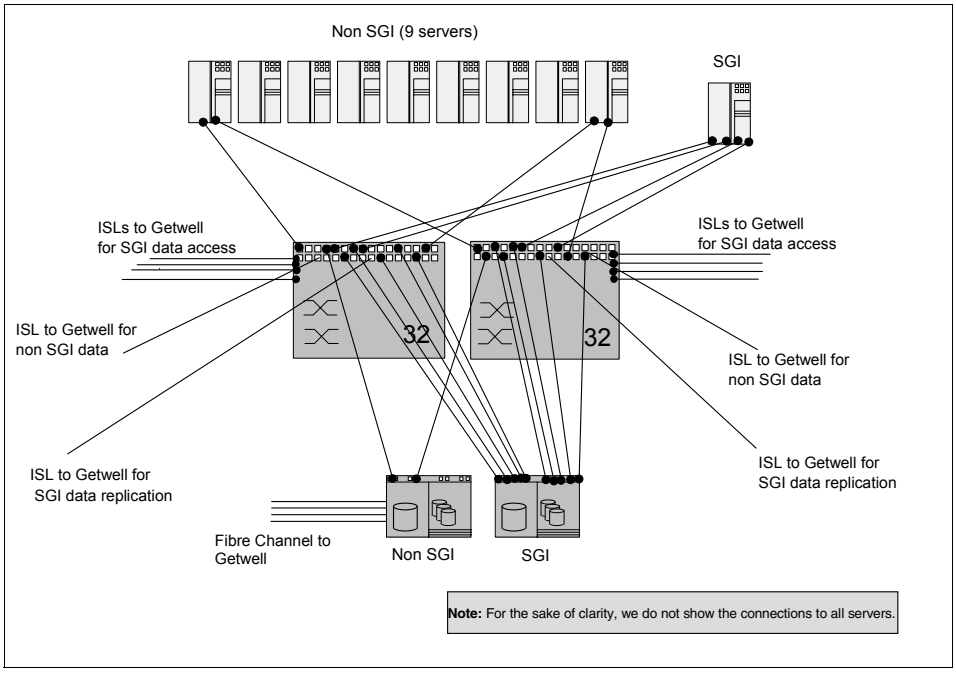


Figure 22-9 Feelinbad SAN design using Cisco MDS 9506 directors

To summarize the port allocation:

- Eighteen ports for non-SGI servers (dual connect)
- Two port for non-SGI storage
- Four ports for SGI servers
- Eight ports for SGI storage
- Four fibre channel connections from the storage device to Getwell
- Eight ports for ISLs for connecting to Getwell
- Two ports for ISLs for SGI replication
- Two ports for ISLs for non-SGI

Which makes for a total of 48 Fibre Channel connections.

Table 22-8 Initial ports

Storage	Server	ISL	Spare
12	24	12	16

With such a design we are fulfilling the requirements:

- No single point of failure, redundant SAN
- All bandwidth requirements are met.
- SAN components can be upgraded without impact on the servers. In case of upgrade or maintenance of switches, we can first upgrade one switch, while paths across the second one are still available. After the traffic is established back on the upgraded switch, we can upgrade the second switch.
- Possible growth without impact on the production servers. Because we are using a redundant SAN, we can introduce additional switches without downtime on the existing servers.

We also provided enough ports for the planned growth on both sides:

- Four SAN storage ports for SGI storage
- Four SAN server ports for SGI servers (Two dual server ports)
- Four SAN ports for ISLs between for SGI servers

In the following sections, we outline some aspects of the design.

22.2.2 Performance

As we can see from the design, we have an initial 54:1 fan-out ratio for all non-SGI servers accessing storage device port. This means that 54 servers are accessing the same storage pool. As we need 75.3 MB/s for all the servers we have enough bandwidth on the storage port for all servers. The fan-out ratio for SGI servers is 1:1. The fan-out ratio will not change in the future, because we will only increase the number of ports on SGI servers. When we increase the number of ports on the SGI servers, we also increase the SGI storage ports by the same number. As we can see from the design, the number of ISLs covers all bandwidth requirements.

Note: Refer to your storage device information to determine if a 54:1 fan-out ratio is too high.

The latency between the sites will be around 208 microseconds. We will use synchronous copying of the storage data and 208 microseconds should not cause significant time delays for applications.

Note: The rule of thumb is that the latency per 1 km of fiber optic cable is 5 microseconds.

22.2.3 Availability

The Cisco MDS 9506 Director solution provides 99.999% availability. This number will effectively be somewhat higher in a dual director configuration. This SAN design, with two paths from each server to the storage device, will give us the ability for each server to experience an HBA failure without impacting performance and without downtime of the servers.

As we have two sites, our SAN is designed to be redundant also in the area of connecting those two sites. We provide redundant paths for data replication and also for data access in the case that storage in the primary site will fail.

22.2.4 Security

When implementing the switches and directors, we need to accommodate some security related items:

- **Director security**
  - Change the default passwords to access the switches.
  - Put directors in separate management network if one is already in place for non-SGI functions.
- **Zoning**

  In our case, we have a multi-platform environment. Because we have only one storage port for all servers except SGI, there is no need to implement any zoning. If in the future we expand the number of storage ports, we can group the servers by the storage ports, and separate them into zones for performance reasons. This can be implemented non-disruptively using software zoning.

  This would result in having the following zone:

  - Zone for all SGI servers in both sites.

Note: There are no data integrity issues if you do not implement zoning. In our example, we only use zoning for performance reasons.

22.2.5 Distance

As you can see from the designs, we are using a maximum of one hop between the directors in each site. This means that there is no practical delay in the performance. All the local connections will use shortwave optics as the servers are within a radius of 500 m. It is possible to move some servers further away by using longwave optics for ISLs. Even with the fabric expansion, as shown in designs, we should have no problems with the delays in the fabric OS for name server and FSPF changes.

For connections from site to site we will use shortwave optics which will be connected over DWDM products. With the use of a DWDM product we will solve the degradation of laser signal, but we still need to adjust the buffers needed on switch and director ports for this type of connection.

22.2.6 Scalability

From the designs we accommodated three years growth. For the predicted growth we will only accommodate new ports for SGI servers and storage. In case that in the future we would need more ports the design is ready to be expanded with new or director port blades.

22.2.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: The clustering solution will failover to the passive server dynamically.
- HBA: If one of the HBA fails, multipath software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multipath software will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: Dual power supplies are supported and no disruption to service will occur if one fails. The replacement power supply can be installed concurrently.
- Director port: If one of the ports fails, you may replace it using a hot-pluggable optics.
- Director blade: If a director port blade fails, the alternate path will be used. The port on the failed blade can be dynamically plugged into the spare ports in other blades until the blade can be replaced.

- Director: The director is a 99.999% type solution, it is therefore, improbable that it will fail. Should it fail, there will be no impact if a dual director solution has been implemented.

22.2.8 Manageability and management software

The directors and switches have built-in management capabilities:

- Serial interface to the OS, typically, only used to configure the network address.
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use.

To fully utilize the management features, the directors have to be connected to the network (Ethernet). Although inband management is supported, we are using outband management.

Note: This network can be part of your existing network infrastructure.

22.3 Case Study 3: Company Three

If we consider the company and its requirements, as detailed in 18.3, “Case Study 3: Company Three” on page 697, we will propose the following solution.

22.3.1 Design

Considering the requirements as detailed in 18.3.5, “Analysis (ports and throughput)” on page 699, we have proposed in this section a switch based solution for our design.

We show our proposed design in Figure 22-10.

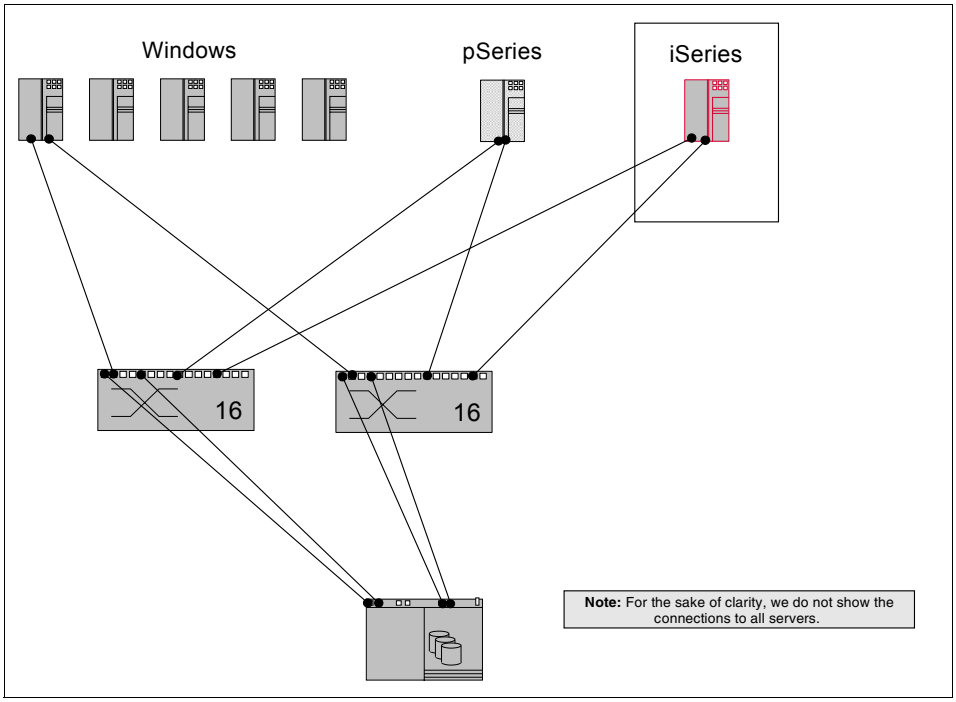


Figure 22-10 Initial design using Cisco MDS 9216 Switches

In this scenario, we are using two Cisco MDS 9216 switches with a 16 port configuration initially. Maximum port configuration of Cisco MDS 9216 is fourty-eight.

It is important to note that Cisco solutions do not support iSeries with code level lower than OS/400 V5.2.

For interoperability matrix, please refer to this website:

http://www.storage.ibm.com/index.htm

As we are using two paths from each Windows and pSeries server to the storage, and two paths from the iSeries server to the storage, we are providing redundancy and high availability. To utilize the dual connections we will use multipathing software (such as IBM's Subsystem Device Driver) on each of the servers. For iSeries server, OS/400 has its own failover support. The solution shown has been configured with two 16-port switches. With this design we are partially fulfilling the requirements:

- Redundant SAN with two switches and all servers dual connected.
- All bandwidth requirements are met.

- By incorporating some of the director class functionality, firmware upgrades and maintenance can be achieved non-disruptively. That is, there is no requirement to redirect all traffic through one switch while the other is being upgraded.
- Future growth by adding more switches can be accomplished without service interruption.

In Table 22-9 we summarize the port usage.

Table 22-9 Port usage

Storage	Server	ISL	Spare
4	14	0	14

In the following sections, we outline some of the aspects of this solution.

22.3.2 Performance

In the design we have a fan-out ratio of 6:1 for all servers except iSeries, and 1:1 for iSeries. Latency through the switches is negligible (~2 micro seconds).

22.3.3 Availability

This SAN design, with two paths from each server to the storage device, will give us the ability for each server to experience an HBA failure without impacting performance and without downtime of the servers.

Cisco MDS 9216 switch has dual power supplies which provides redundancy and high availability in case of power outage.

22.3.4 Security

When implementing the switches we need to take care about some security related things:

- **Switch security**
  - Change the default passwords to access the switches.
  - Put switches in a separate management network if one is already in place for other functions.
- **Zoning**

  Zoning will be implemented to separate the three platforms (Windows, pSeries and iSeries), and isolate the iSeries I/O to its own storage ports.

22.3.5 Distance

In all the connections we will use shortwave optics as the servers are within a radius of 300 m. It is possible to move some server further away by using longwave optics for ISLs.

22.3.6 Scalability

As you can see from the design, there are spare ports for growth. The growth can be achieved without any interruptions in the production environment.

22.3.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Open Systems Server: The clustering solution will failover to the passive server dynamically.
- iSeries server: Application is not available. Server has to be replaced or repaired.
- HBA (Open Systems): If one of the HBA fails, multi-path software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- HBA (iSeries): OS/400 failover support will handle HBA redundancy. There will be no impact of the performance.
- Cable: If a cable between a server and the switch fails, multi-path software will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: Cisco MDS 9216 solution has dual power supplies as a standard feature and no disruption to service will occur should one fail. The replacement power supply can be installed with no outage incurred.
- Switch port: If one of the ports fail, you may replace the hot-pluggable optic. No outage will occur as the server will use its alternate path.
- Switch: Should the switch fail, the server will use its alternate path through another switch without an outage or performance degradation.

22.3.8 Manageability and management software

The switches have built-in management capabilities:

- Serial interface to the OS, typically only used to configure the network address

- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use.

To fully utilize the management features, the switches have to be connected to the network (Ethernet). Although inband management is supported, in our solution we are not using ISLs, and therefore, inband management is not an option.

We show an example of setting up the management network in Figure 22-4.

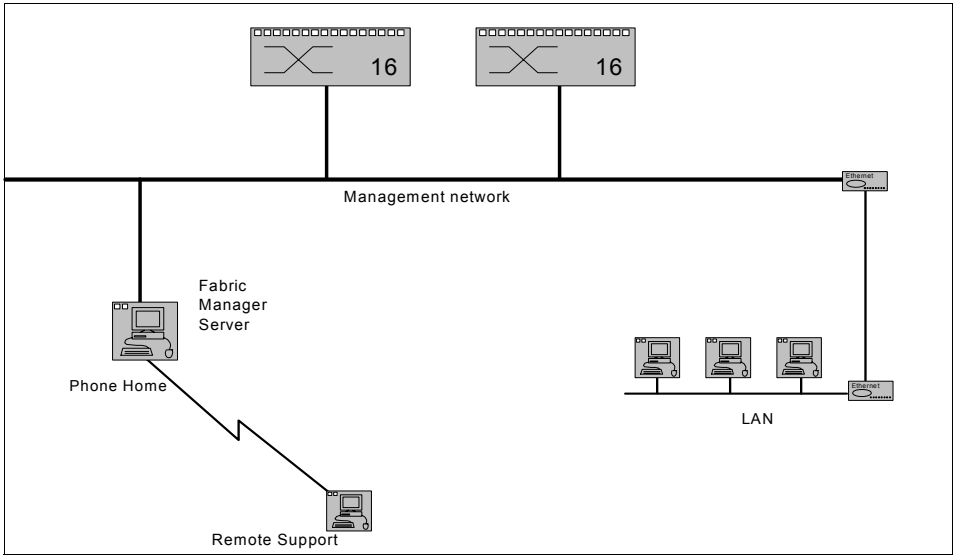


Figure 22-11 Management network

Note: This network can be part of your existing network infrastructure.

22.4 Case Study 4 - Company Four

If we consider the company and its requirements as detailed in 18.4, “Case Study 4: Company Four” on page 700, we will propose the following solution.

22.4.1 Design

Considering the requirements as detailed in 18.4.5, “Analysis (ports and throughput)” on page 703, we have proposed in this section a switch based solution for our design.

Both the open systems and zSeries platforms will share the same storage device which will provide much higher availability.

As you can see from our proposed design in Figure 22-12, we have replaced the ESCON director with Cisco MDS 9506 director for both sites. Cisco MDS 9506 is designed to support FICON and Fibre Channel intermix which is best suited for our scenario with two different platforms. Cisco MDS 9506 shown has been configured with thirty-two ports initially, a full configuration can support up to 128 ports.

.

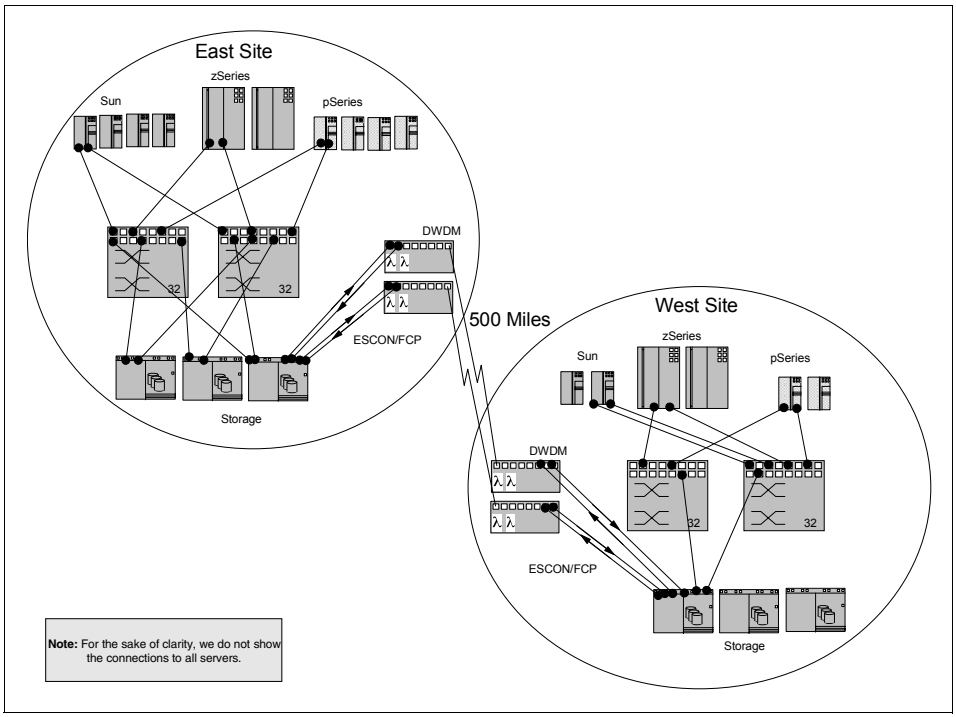


Figure 22-12 Initial design using Cisco MDS 9506 director.

As we are using redundant paths from each server and to the storage we are providing high availability. To utilize the redundant connections we will use multipathing software (such as IBM's Subsystem Device Driver) on each of the open systems servers, while for our mainframe servers we have z/OS built in failover support for path redundancy. The solution shown has been configured with two 32-ports initially. With this design, we are partially fulfilling the requirements:

- Redundant SAN with two switches and all servers dual connected.
- All bandwidth requirements are met.
- By incorporating some of the director class functionality, firmware upgrades and maintenance can be achieved non-disruptively. That is, there is no requirement to redirect all traffic through one switch while the other is being upgraded.
- Future growth by adding more switches can be accomplished without service interruption.

In the following sections, we outline some of the aspects of this solution.

For storage replication we are using ESCON links over DWDM. Depending on the implementation time frame, we may be able to substitute these with FCP using PPRC V2. The peak bandwidth used by the servers is 11.58 MB/s. As only 30% of these are writes then we need to accommodate 3.48 MB/s. Two pairs of ESCON channels will accomplish this and provide redundancy. The same DWDM devices used for open systems can be used for the zSeries requirement.

In Table 22-10 we show the ports at the East site.

Table 22-10 Ports: East site

Storage	Server	ISL	Spare
10	20	0	34

In Table 22-11 we show the ports at the West site.

Table 22-11 Ports: West Site

Storage	Server	ISL	Spare
10	12	0	42

22.4.2 Performance

In the design, we have a fan-out ratio of 20:6 for all servers in the East site. The fan-out ratio in the West site is 12:6. Latency through the switches is negligible (~2 micro seconds).

The latency between the sites will be around 4 milliseconds.

22.4.3 Availability

This SAN design, with two paths from each server to the storage device, will give us the ability for each server to experience an HBA failure without impacting performance and without downtime of the servers.

Cisco MDS 9506 director supports functionalities such as non-disruptive firmware upgrades and maintenance. That is, there is no requirement to redirect all traffic through one director while the other is being upgraded.

22.4.4 Security

When implementing the directors we need to take care about some security related things:

- **Director security**
  - Change the default passwords to access the switches.
  - Put directors in a separate management network if one is already in place for other functions.
- **Zoning**

We will be implementing zoning so we can group the servers by storage ports (FICON/FCP) and separate then into zones.

22.4.5 Distance

In all the connections, we will use shortwave SFPs as the servers are within a radius of 500 m. It is possible to move some servers further away by using longwave SFPs for ISLs.

22.4.6 Scalability

The design accommodates three years growth. The growth can be achieved without any interruptions in the production environment.

22.4.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: The clustering solution will failover to the passive server dynamically.

- HBA: If one of the HBA fails, multi-path software for open systems server and z/OS for mainframe servers, will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multi-path software or z/OS will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: Cisco MDS 9506 director solution has dual power supplies as a standard feature and no disruption to service will occur should one fail. The replacement power supply can be installed with no outage incurred.
- Director port: If one of the ports fails, you may replace the hot-pluggable optic. No outage will occur as the server will use its alternate path.
- Director: Should the director fail, the server will use its alternate path through another switch without an outage or performance degradation.
- DWDM device: Should the DWDM device fail, the alternate one will be used.

22.4.8 Manageability and management software

The switches have built-in management capabilities:

- Serial interface to the OS, typically only used to configure the network address.
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use.

To fully utilize the management features, the switches have to be connected to the network (Ethernet). Although inband management is supported, in our solution we are not using ISLs, and therefore, inband management is not an option.

We show an example of setting up the management network in Figure 22-13.

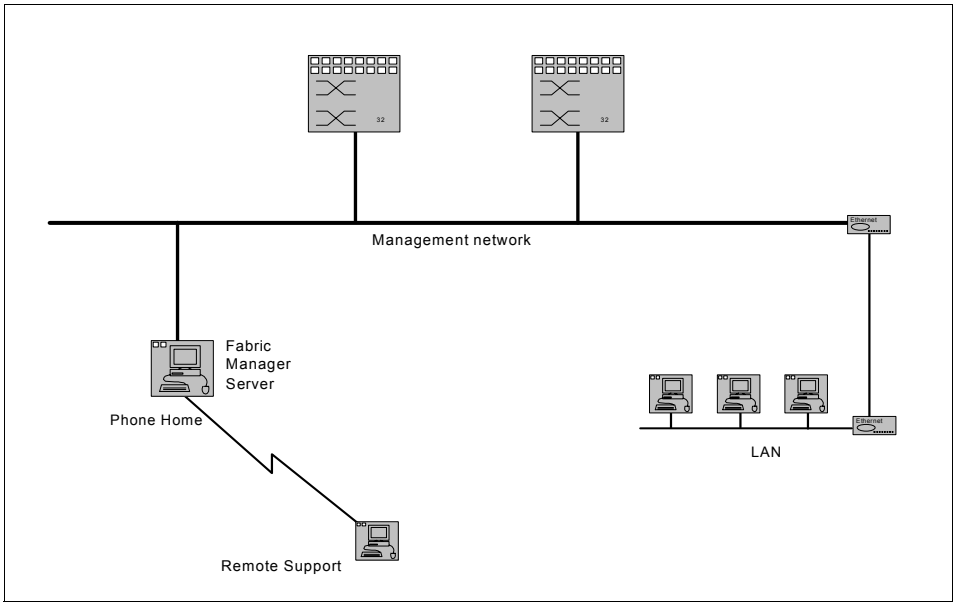


Figure 22-13 Management network

Note: This network can be part of your existing network infrastructure.

22.5 Case Study 5 - Company Five

If we consider the company and its requirements as detailed in 18.5, “Case Study 5: Company Five” on page 705, we will propose the following solution.

22.5.1 Design

Considering the requirements as detailed in 18.5.5, “Analysis (ports and throughput)” on page 708, we selected a solution based around the Cisco MDS 9216 since it supports FC-AL which is required for accessing tape drives.

The proposed design is shown in Figure 22-14.

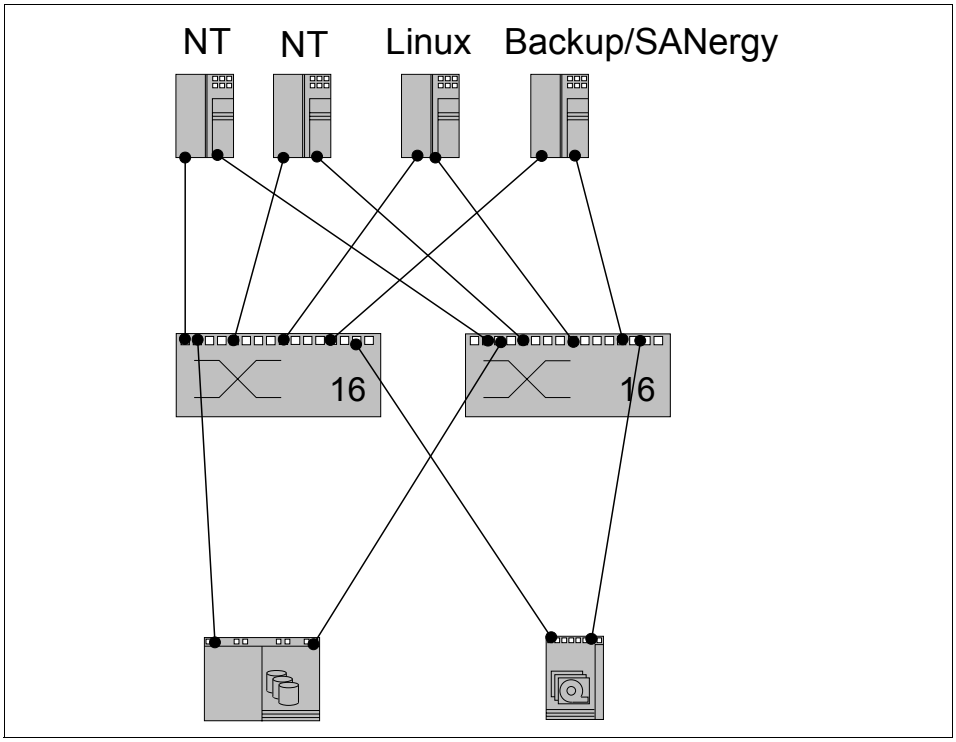


Figure 22-14 The proposed design for Company Five

As we are using two paths from each server to the storage we are providing redundancy and high availability. To utilize the dual connections, we will use multipathing software (such as IBM's Subsystem Device Driver) on each of the servers. The solution shown has been configured with two 16-port switches (additional ports can be added in groups of eight, to a maximum of 24). With this design we are partially fulfilling the requirements:

- Redundant SAN with two switches. All servers are dual connected to the switches. Tape drive is connected to the same SAN via FC-AL.
- All bandwidth requirements are met.
- By incorporating some of the director class functionality, firmware upgrades and maintenance can be achieved non-disruptively. That is, there is no requirement to redirect all traffic through one switch while the other is being upgraded.
- Future growth by adding more switches can be accomplished without service interruption.

You will notice that we have introduced a fourth server to accommodate heterogeneous file sharing (using Tivoli SANergy) and also backup/restore software (using Tivoli Storage Manager). After the migration from NT to Linux is complete, we could consider using the vacant NT server, thus reducing the number of servers to three.

In Table 22-12 we show the initial ports used.

Table 22-12 Initial ports

Storage	Server	ISL	Spare
4	8	0	20

In the following sections, we outline some of the aspects of this solution.

22.5.2 Performance

In the initial design, we have a 4:1 fan-out ratio for all servers accessing the storage. This means that four servers are accessing one storage port. All servers can also access the tape devices.

Attention: Host software (such as Tivoli Storage Manager) must be present to allow functions that enable tape sharing and/or tape library sharing.

The Cisco MDS 9216 switch supports FC-AL which enables direct access of the tape drives.

22.5.3 Availability

All servers and the storage have multiple paths giving resilience through redundancy.

The Cisco MDS 9216 switch has dual power supplies which provide redundancy and high availability in case of power outages.

22.5.4 Security

When implementing these devices, we need to address some security issues:

- **Switch security**
  - Change the default passwords.
  - Put the directors into a separate management network if one is already in place for other functions.

- **Zoning**

  We have only one disk storage port per switch for all servers, therefore, there is no requirement to implement any zoning. In the future, if we expand the number of storage ports, we can group the servers to the storage ports in separate zones for performance reasons. This can be implemented non-disruptively using software zoning.

  Note: There are no data integrity issues if you do not implement zoning. In our example we would only use zoning for performance reasons.

22.5.5 Distance

All devices (servers, switches, and storage) will use shortwave optics and cables as they are within a radius of 500 m.

22.5.6 Scalability

In the design, we accommodated three years growth. We have enough bandwidth allocated for all requirements in the next three years. If, in the future, we need more ports, the switches can accommodate another 10 servers or tape drives. The Cisco MDS 9216 is upgradeable and supports a maximum number of 48 ports.

22.5.7 “What if” failure scenarios

- Server: Application will not be available. Server has to be replaced or repaired.
- HBA: If one of the HBA fails, multi-path software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multi-path software will automatically failover workload to the alternate path. If a cable between the switch and the disk storage fails, an alternate route will be used. There is no performance loss.
- Power supply: Cisco MDS 9216 has redundant power supplies.
- Switch Port: If one of the ports fails, you may replace it using a hot-pluggable optic, without outage or performance problem.
- Switch: If a switch fails the alternate switch will accommodate the workload.

22.5.8 Manageability and management software

The switches have built-in management capabilities:

- Serial interface to the OS, typically only used to configure the network address
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use

To fully utilize the management features, switches have to be connected to the network (Ethernet). Although inband management is supported, in our solution we are not using ISLs, and therefore, inband management is not an option.

We show an example of setting up the management network in Figure 22-15.

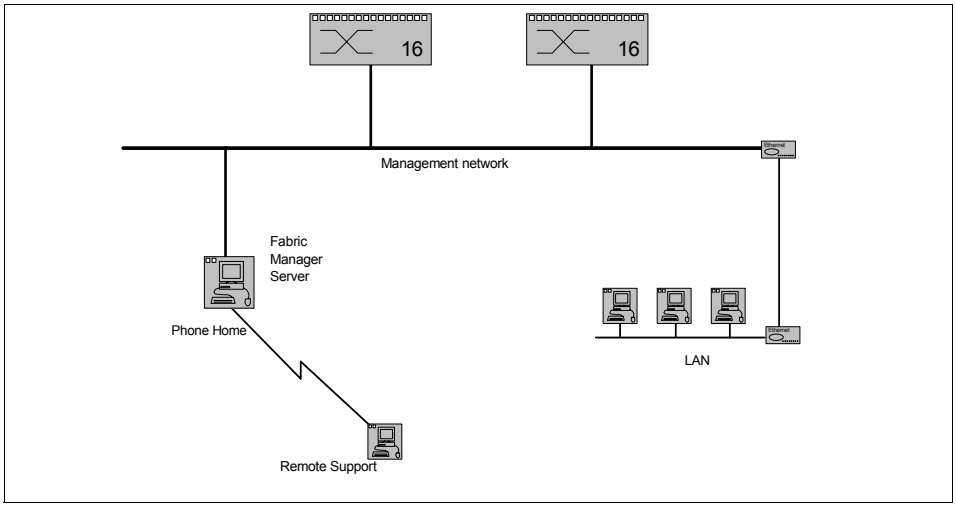


Figure 22-15 Management network

Note: This network can be part of your existing network infrastructure.

22.6 Case Study 6: Company Six

If we consider the company and its requirements, as detailed in 18.6, “Case Study 6: Company Six” on page 710, we will propose the following solution.

22.6.1 Design

Considering the requirements as detailed in 18.6.5, “Analysis (ports and throughput)” on page 712, we selected a solution based around the Cisco MDS 9216.

We show the proposed design for Primary site in Figure 22-16.

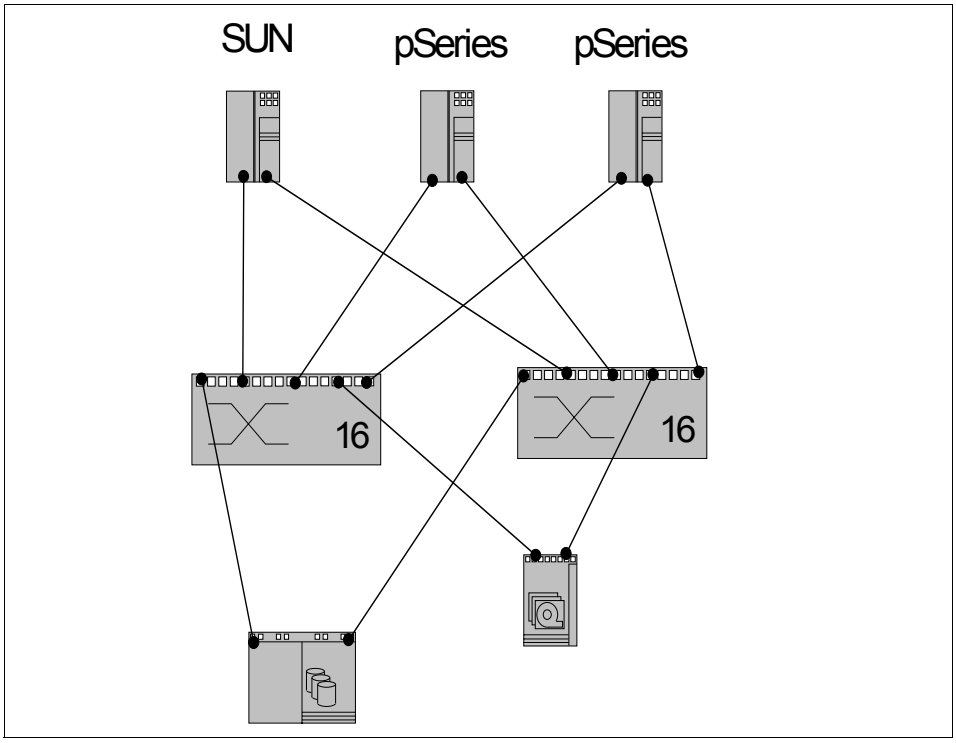


Figure 22-16 The proposed design for the Primary site

In the secondary site, we already have a SAN which has a Compaq switch. You can see the design of the Secondary site in Figure 22-17.

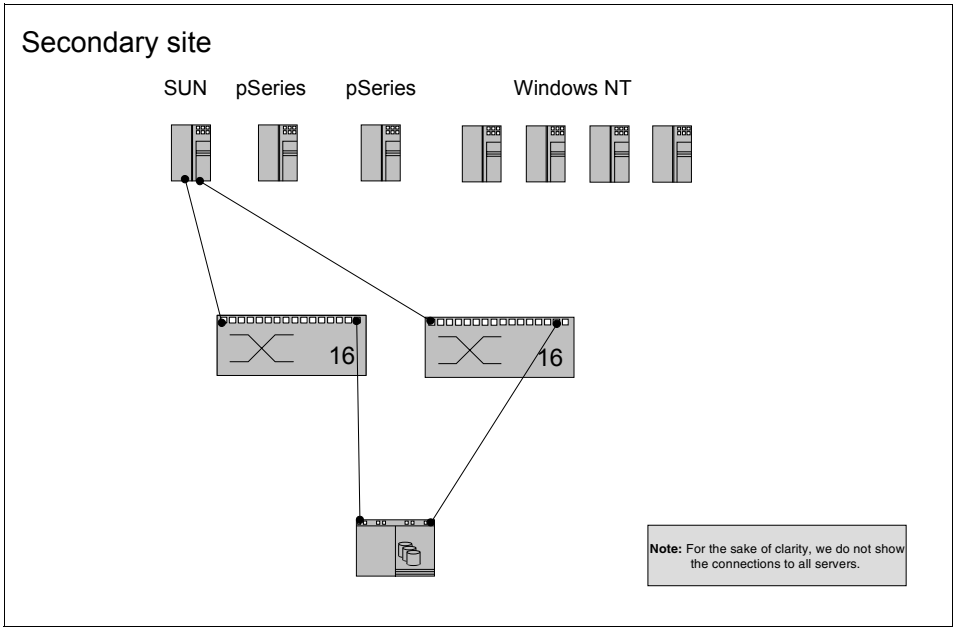


Figure 22-17 The proposed design for the Secondary site

Both sites are connected with two ISLs. Because the distance is 900 miles, we will use DWDM products to connect them. You can see the connection setup in Figure 22-18.

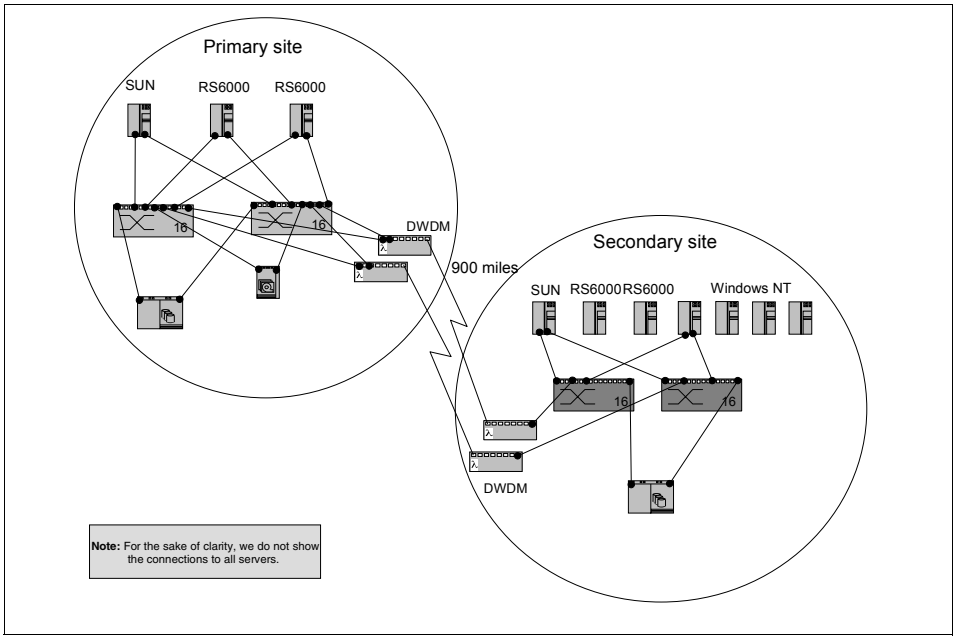


Figure 22-18 The complete solution

Note: To the best of our knowledge, this solution (with the Cisco MDS 9216 and the Compaq) has not been tested, and is therefore not certified.

As you can see we used a Cisco MDS 9216 with 16 ports configured initially, additional ports can be added. With such a design, we are fulfilling these requirements:

- Redundant SAN with two switches and all servers dual connected
- All bandwidth requirements are met.
- These directors supports functionalities such as a non-disruptive firmware upgrade. That is, there is no requirement to redirect all traffic through one switch while the other is being upgraded.
- Future growth by adding more directors can be accomplished without service interruption.
- We are re-using existing equipment (directors and tape).
- We are providing storage replication to remote site for disaster recovery.
- We are improving backup performance by providing infrastructure for LAN-free backup.

Attention: Host software (such as Tivoli Storage Manager) must be present to allow functions that enable tape sharing and/or tape library sharing.

In Table 22-13 we show the ports at the Local site.

Table 22-13 Ports at Local site

Storage	Server	ISL	Spare
4	6	4	20

There are ample ports left at the Remote site, the final port allocation for the remote site is represented in Table 22-14.

Table 22-14 Ports at Remote site

Storage	Server	ISL	Spare
2	14	2	14

In the following sections, we will outline some aspects of the design.

22.6.2 Performance

In the initial design, we have a 3:1 fan-out ratio for all servers accessing the storage on the Primary site. This means that three servers are accessing one storage port. All servers can also access the tape device with greater performance and availability, which was the requirement for LAN-free backup implementation. There are no latency issues, because the traffic is only going through the director, where the latency in the low microseconds.

At the Secondary site we have 7:1 fan-out ratio. In this case, the traffic is only going through the switches and again the latency is so low as to be of no significance.

22.6.3 Availability

All servers and the storage have multiple paths giving resilience through redundancy.

As we have two sites our SAN is designed to be redundant also in the area of connecting those two sites. We are providing redundant paths for data replication. It recommended that the connections between two DWDM boxes are also redundant. This however, could be a considerable expense.

22.6.4 Security

When implementing the directors, we need to address some security issues:

- **Director security**
  - Change the default passwords.
  - Either put the directors into an existing separate management network if one is already in place for other functions, or implement a new network.
- **Zoning**

  In our case, we have a multi-platform environment. Therefore, we will group platforms including servers accessing tape library for performance reasons. This can be implemented non-disruptively using software zoning.

22.6.5 Distance

As can be seen from the designs, we are only using one switch between the server and the storage in each site. This means that there is no practical delay in the performance. All the local connections will use shortwave optics.

With the use of a DWDM product we will solve the degradation of laser signal, but we still need to adjust the buffers needed on switch ports for this type of connection. On the switches at both sites, we need to enable the Extended Distance option for the ISL ports.

22.6.6 Scalability

From the designs we have accommodated three years growth. We have enough bandwidth allocated for all requirements in the next three years. The only expansion will be probably by adding additional storage ports, and we accommodated the space for this. In case we would need more ports, the design is ready to be expanded with new directors.

22.6.7 “What if” failure scenarios

Here are the “what if” scenarios we considered:

- Server: Application will not be available. Server has to be replaced or repaired.
- HBA: If one of the HBA fails, multi-path software will automatically failover the workload to the alternate HBA. There will be no impact on the performance.
- Cable: If a cable between a server and the switch fails, multi-path software will automatically failover workload to the alternate path. If a cable between the

switch and the disk storage fails, an alternate route will be used. There is no performance loss.

- Power supply: Cisco MDS 9216 switch has redundant power supplies.
- Switch: If one of the ports fails, you may replace it using a hot-pluggable optic, without outage or performance problem.
- Switch: If a switch fails, the server will use the alternate switch to connect to the storage, without outage or performance problem.

22.6.8 Manageability and management software

The switches have built-in management capabilities:

- Serial interface to the OS, typically, only used to configure the network address.
- HTTP interface with graphical management
- SNMP can be setup to forward messages to the management server if there is one in use.

To fully utilize the management features, the switches have to be connected to the network (Ethernet). Although inband management is supported, in our solution we are not using ISLs and therefore inband management is not an option.

We show an example of setting up the management network in Figure 22-4.

23

IBM TotalStorage SAN b-type family distance solutions

In this chapter we will discuss some solutions that can be implemented using the IBM TotalStorage SAN Switch products included in the IBM portfolio.

For details of the IBM TotalStorage SAN Switch portfolio products refer to:

- Chapter 8, “IBM TotalStorage SAN b-type family” on page 287
- *Implementing an Open IBM SAN,* SG24-6116

The solutions we will discuss are:

- Remote disk consolidation
- Two sites up to 10 km apart
- Two sites up to 100 km apart with repeaters
- Point-to-Point DWDM
- Point-to-Point DWDM with PPRC
- Ring topology DWDM
- Channel extender and WAN extension
- FCIP connection
- Remote tape vaulting
- Remote tape vaulting with redundancy

© Copyright IBM Corp. 2004. All rights reserved.

This chapter presents various implementations and uses for Storage Area Network (SAN) in high availability graphically dispersed environments. The considerations for a high availability SAN design using Fibre Channel switches and directors are covered in the following topics.

By reducing or eliminating single points of failure in the enterprise environment, SANs can help to improve overall availability of business applications. By utilizing highly available components and solutions, as well as a fault-tolerant design, enterprises can achieve the availability needed to support 24x7 uptime requirements. High distances in these solutions enable the SAN to cover a whole new host of IT availability issues, for example, we can deploy disaster tolerant solutions.

In network systems such as SANs, with their associated servers, fabric, and storage components, as well as software applications, downtime can occur even if parts of the system are highly available or fault tolerant. To improve business continuance under a variety of circumstances, SANs can incorporate redundant components, connections, software, and configurations to minimize or eliminate single points of failure.

With the emergence of the Internet and the proliferation of global e-business applications, more and more companies are implementing computing infrastructures specifically designed for continuous data and system availability. Today, even applications such as company e-mail have become mission critical for ongoing business operations. Faced with increased customer and internal user expectations, companies are currently striving to achieve at least 99.999 percent (“five nines”) availability in their computing systems — a figure equivalent to less than 5.3 minutes of downtime a year. Additional downtime can severely impact business operations and cost valuable time, money, and resources.

To ensure the highest level of system uptime, companies are implementing reliable storage networks capable of boosting the availability of data for all the users and applications that need it. These companies typically represent the industries that demand the highest levels of system and data availability — the utilities and telecommunications sector, brokerages and financial service institutions, and a wide variety of service providers.

Notes

- The IBM TotalStorage switch models Fxx and M12 have four ports per ASIC, whilst the Hxx and M14 have eight ports per ASIC. The buffer credit allocation is on an ASIC level, so ports will share a pool of buffer credits.
- This is a key consideration when introducing distance and it should be the aim of a low level design to spread the allocation of long distance ports among the ASICs for that switch.

- All long distance links within the fabric must be configured as ExtendedFabric links in order to obtain optimum I/O performance. Extended Fabric is an optional license and must be installed before long distance connectivity is enabled.
- Certain switch models will only connect to long distance links to switches of the same model.

23.1 High-level availability objectives

Developing highly available networks involves identifying specific availability requirements and predicting what potential failures might cause outages. The first step is to clearly define availability objectives, which can vary widely from company to company and even within segments of the same company. In some environments, no disruption can be tolerated, while other environments might be only minimally affected by short outages. As a result, availability is a function of the frequency of outages (caused by unplanned failures or scheduled maintenance and upgrades) and the time to recover from such outages.

Many companies are addressing their availability requirements by implementing networks fabrics of Fibre Channel devices designed to provide high-performance storage environments. These flexible SANs are based on the following principles:

- A thorough understanding of availability requirements throughout the enterprise
- A flexible design that incorporates fault tolerance through redundancy and mirroring
- Simplified fault monitoring, diagnostics, and repair capabilities to ensure fast recovery
- A minimal amount of human intervention required during failover events
- A reliable backup and recovery plan to account for a wide variety of contingencies

23.2 Storage consolidation with remote disk

This remote disk consolidation solution is an extension of the most basic SAN solutions of disk consolidation (Figure 23-1).

Disk consolidation provides a means for storage administrators to take advantage of the capabilities of Fibre Channel fabric-based Storage Area Networks to organize storage in more flexible and efficient ways, easing many aspects of storage administration.

In an attempt to extend the distance between the storage and the servers, and in order to provide business benefits like consolidation, physical separation of disk storage from the server across sites, is necessary. This solution extends the distance to up to 10 km apart, using longwave SFPs in the SAN fabric, and single-mode fiber cable interconnect.

Note: IBM TotalStorage SAN Switches support a distance of up to 80 km using extended longwave SFPs between the directors.

The example is comprised of:

- Three dual-port hosts
- Four IBM TotalStorage SAN Switch M12 directors
- One IBM TotalStorage Enterprise Storage Server Model 800 with eight FC host adapters

The host servers and ESS are connected to the SAN fabric via shortwave SFPs using multi-mode fibre cable.

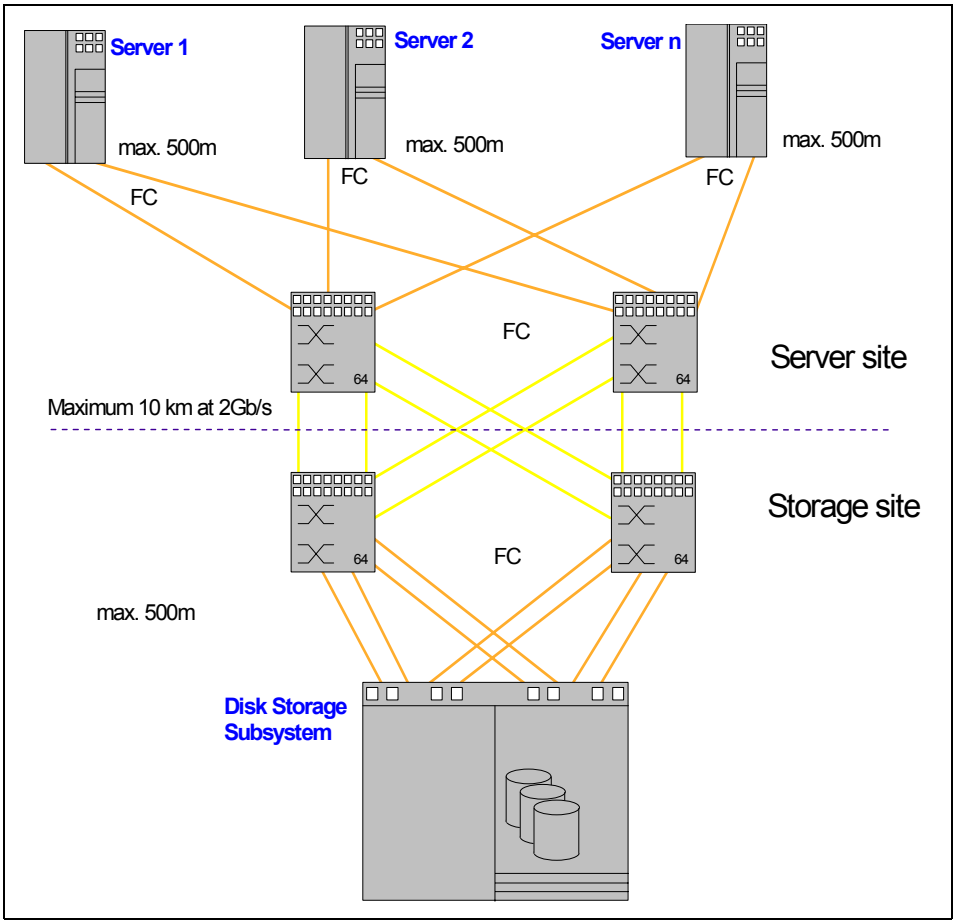


Figure 23-1 Remote disk consolidation

Checklist

We checked the following items:

- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and firmware level for proposed configuration
- IBM TotalStorage SAN Switch director longwave SFP count
- All cables clearly labelled and recorded
- The ISL links — single-mode fiber cable
- Single-mode FC cable laying and termination

- IBM TotalStorage SAN Switch high availability features
- Nickname assignments so we can quickly cross reference WWNs to devices
- ESS LUN definitions done
- Fabric Manager userids and passwords defined

Performance

In this simple implementation, performance will basically depend on the number of HBAs available on each server and the number of storage connections and most important the distance between storage and servers. It is important to consider application behavior over the distance, and this varies from application to application and we have assumed that the performance is adequate for applications running in the servers.

The IBM TotalStorage SAN Switch supports any to any connectivity so it will not affect performance by itself. Latency in the M14 director is less than 2.1 microseconds; this is not relevant compared to storage device response times typically in the millisecond range.

Knowing the requirements of our servers we can decide whether it is enough or if we need to add more connections or even more storage devices.

Without having the exact requirements we may consider that for a high profile server, a server to storage ratio of 6:1 is acceptable. This is only a starting point, we will then implement some measurement system or use statistical data to decide whether we need to add more connections or we have more bandwidth than required.

Scalability

The IBM TotalStorage SAN Switch supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation. With four ISLs between the pair of directors on the host side and the pair of directors on the storage side, and four spare ports in a 64 port director, we can have 56 dual HBA servers (64 ports - 4 ISL - 4 spare = 56 server ports). With eight ESS connections, this slightly exceeds the 6:1 server to storage ratio (6 x 8 = 48).

Security

Here are some security considerations:

- ESS LUN masking by WWN will allow each server access only to configured LUNs.

- Fabric Manager userids, passwords and user rights are defined and defaults are removed, so only authorized personnel can perform management functions.
- Remote access to Fabric Manager configured to limit access to authorized workstations.
- Physical director security — locked cabinet, restricted access site.

"What if" failure scenarios

These are some theoretical assumptions:

- **Host HBA failure**

  SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. Fabric Manager access and ESS Specialist access are required.

- **Storage host adapter failure**

  The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed eight host adapters, so a single adapter failure will reduce available bandwidth by 12.5%. For an average workload and three servers as shown it will not impact performance. When the host adapter is replaced we do not need to update zoning as the ESS WWN is maintained by the ESS.

- **Director port failure**

  The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port, as the target FC ID will have changed. Physical access to director and Fabric Manager user with maintenance rights is required. OS root access may be required.

- **Fiber failure**

  Impact will depend on whether it is a server attachment, storage attachment or ISL fiber. For ISL fiber, path routes may be updated within the fabric. The only action required is fiber replacement. Physical access to director and attached device are required.

- **Fabric Manager Server failure**

  No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example. Call home support will also be unavailable.

- **Fabric Manager Server hard drive failure**

Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from the backup.

- **Director completely down, storage completely down, or site down (power, air conditioning, site damage**)

  This will cause an interruption in normal operation.

- **Physical damage to storage causing data loss (fire, flood)**

  We will need to restore data from backup tapes.

23.3 Two sites up to 10 km apart

In Figure 23-2, we show a basic dual fabric SAN distance extension solution between two sites up to 10 km apart. Each site has a fully redundant infrastructure at the SAN fabric level comprising of two IBM TotalStorage SAN Switch M12 directors. The host servers and storage are connected to the SAN fabric via shortwave SFPs using multi-mode fiber cable. The directors are interlinked across the sites via two ISLs each using longwave SFPs and single-mode fiber cables.

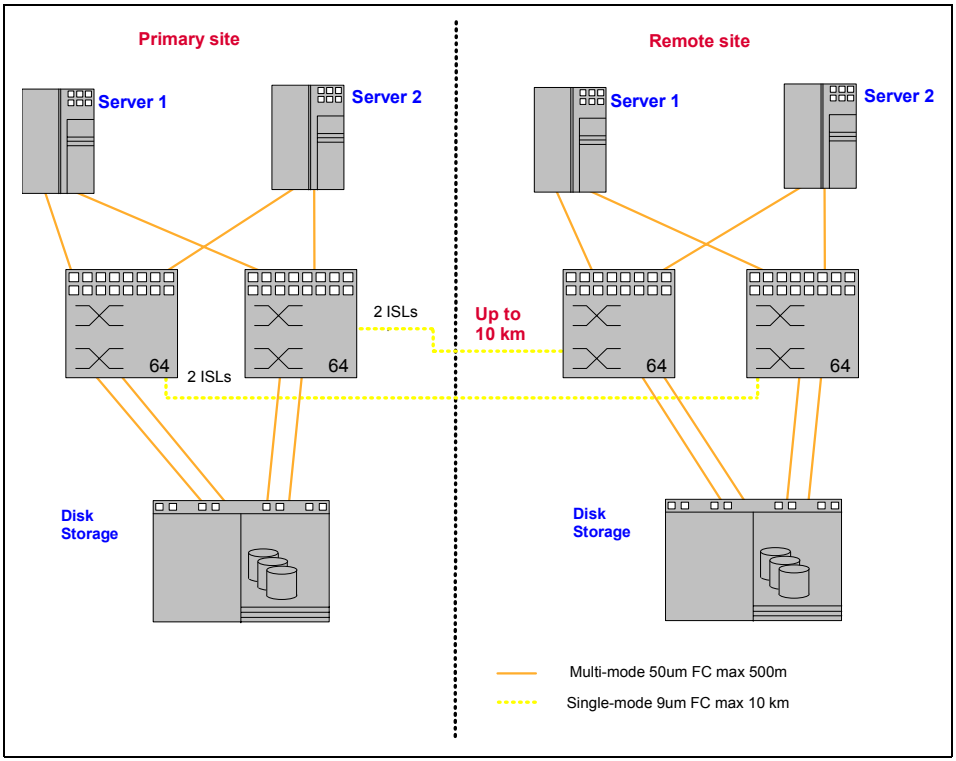


Figure 23-2 SAN distance extension up to 10 km

The solution can be used for accessing a database at the remote site as well as mirroring data at an alternative site up to 10 km.

Note: IBM TotalStorage SAN Switches support a distance of up to 80 km using extended longwave SFPs between the directors.

One can assume that a greater distance between IT sites results in greater security against wide spread disaster. However, increased distance has its price in terms of interconnection cost, business relocation effort, and network cost. Before you start to expand your storage network, you have to consider if the following solution can meet your business purpose.

The two sites are each comprised of:

- Two dual-port hosts
- Two IBM TotalStorage SAN Switch M12 directors
- One IBM TotalStorage Enterprise Storage Server Model 800 with four FC host adapters

- Software
  - IBM Subsystem Device Driver (SDD)
  - IBM StorWatch Enterprise Storage Server Specialist
  - IBM TotalStorage Expert for ESS (for historical disk performance data)

We have assumed the following for our solution:

- Both sites have SAN infrastructure.
- Server to director and director to storage use multi-mode fibre.
- Servers in production which are being used for mission critical production need data mirroring for disaster recovery.
- Some servers at the primary site need to access database in storage at the remote site, in addition to using data in the primary storage.
- Servers at the remote site are in standby-mode, and being used for backup servers at ordinary times.
- Single-mode FC cables are used as ISLs between directors at each site.
- Software mirroring is used on servers.
- No PPRC proposed.

Checklist

The following items should be considered to implement the solution:

- Fibre Channel HBA supported on servers
- Operating systems support software mirroring
- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and firmware level for proposed configuration
- Director longwave SFP count
- The ISL links — single-mode fiber cable
- Single-mode FC cable laying and termination
- E_D_TOV and R_A_TOV settings equal on all directors
- Extended distance buffering enabled for ISL ports to increase BB_Credit for 10-100 km links
- IBM TotalStorage SAN Switch high availability features
- Nickname assignments so we can quickly cross reference WWNs to devices
- ESS LUN definitions done

- Fabric Manager users and passwords defined

Performance

The major performance consideration with a long distance solution is calculating the correct number of lines between sites. This number can only be determined by performing a detailed performance profile of the servers and storage that will be remote. Over estimating the number of lines will increase costs dramatically, whilst under estimating the number of lines will dramatically impact the performance of the SAN.

It is vital that detailed performance data is available prior to sizing the number of lines required.

Typically, latency will increase over long distances – a good rule of thumb is 4.8 microseconds per kilometer. The default E_D_TOV and R_A_TOV values do not need to be modified for this distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the Directors across sites (the ISL links).

Scalability

The IBM TotalStorage SAN Switch supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation. With two ISLs between each director on the primary site and the director on the remote site, two ports for the ESS and four spare ports in a 64 port director, we can have 54 dual HBA servers (64 ports - 4 ISL - 2 ESS ports - 4 spare = 54 server ports). This may be too many for four ESS connections, depending on the type of server and workload, in which case more director ports may need to be used for ESS connections.

Security

Here are some security considerations:

- ESS LUN masking by WWN will allow each server access only to configured LUNs.
- Fabric Manager user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions.
- Remote access to Fabric Manager configured to limit access to authorized workstations.
- Physical director security — locked cabinet, restricted access site.
- The zoning information is separate for the two fabrics, so when changing the zoning in one fabric it is not automatically propagated to the other fabric.

- Our solution assumes we own private lines between sites so encryption is not required.
- For leased lines or managed services where lines are shared, encryption is normally an option available from the service provider.

"What if" failure scenarios

Here are some theoretical assumptions:

- **Host HBA failure**

  SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. Fabric Manager access and ESS Specialist access are required.

- **Storage host adapter failure**

  The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce available bandwidth by 25%. For an average workload and two servers as shown it will not impact performance. When the host adapter is replaced we do not need to update zoning as the ESS WWN is maintained by the ESS.

- **Director port failure**

  The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port, as the target FC ID will have changed. Physical access to director and Fabric Manager user with maintenance rights is required. OS root access may be required.

- **Fiber failure**

  Impact will depend on whether it is a server attachment, storage attachment or ISL fiber. For ISL fiber, path routes may be updated within the fabric. The only action required is fiber replacement. Physical access to director and attached device may be required. For the ISL fibres between sites, the fibre provider will be responsible for repair.

- **Fabric Manager Server failure**

  No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example. Call home support will also be unavailable.

- **Fabric Manager Server hard drive failure**

Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from the backup.

- **Director completely down, storage completely down, or site down (power, air conditioning, site damage)**

  This will cause an interruption in normal operation.

- **Physical damage to storage causing data loss (fire, flood)**

  We will need to restore data from backup tapes.

23.4 Two sites up to 100 km apart

In Figure 23-3 we show a basic dual fabric SAN distance extension solution between two sites up to 100 km apart. Each site has a fully redundant infrastructure at the SAN fabric level comprising of two IBM TotalStorage SAN Switch M12 directors. The host servers and storage are connected to the SAN fabric via shortwave SFPs using multi-mode fibre cable. The directors are interlinked across the sites via two ISLs each using longwave SFPs and single-mode fiber cables and repeaters. The number of repeaters used depends on the type of longwave SFP used.

One can assume that a greater distance between IT sites results in greater security against wide spread disaster. However, increased distance has its price in terms of interconnection cost, business relocation effort, and network cost. Before you start to expand your storage network, you have to consider if the following solution can meet your business purpose.

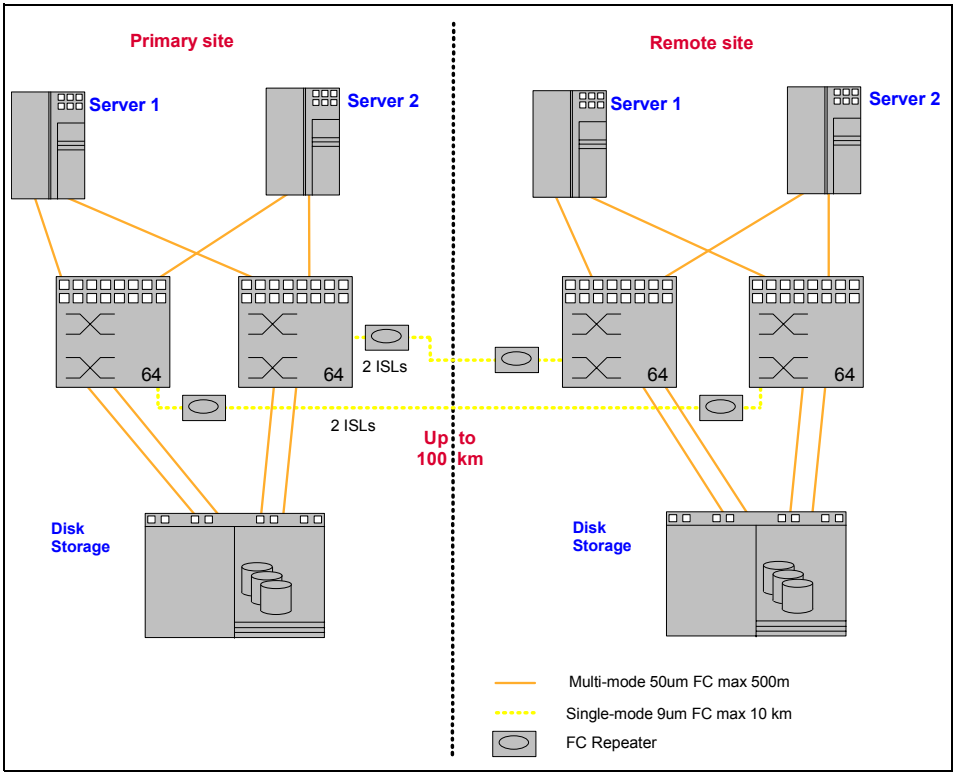


Figure 23-3 SAN distance extension up to 100 km with repeaters

The two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 886, but with the addition of:

- FC repeaters as required to reach 100 km

Checklist

The same items as for “Two sites up to 10 km apart” on page 886 should be considered, with the addition of:

- Repeater count verified for the distance in every ISL link
- Repeaters installed connected and tested

Performance

The same performance criteria apply as for “Two sites up to 10 km apart” on page 886, other than overall ISL latency increases with distance and BB_Credit utilization will increase.

Scalability

The same scalability criteria apply as for “Two sites up to 10 km apart” on page 886.

Security

The same items as for “Two sites up to 10 km apart” on page 886 should be considered, with the addition of:

- Physical repeater security — locked cabinet, restricted access site.

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 886, with the addition of:

- **Repeater failure**

  Impact will depend on whether alternate ISL links are available or not. Only action required is repeater replacement. Physical access to repeater and repeater site required.

23.5 Two sites - Point-to-point DWDM solution

In Figure 23-4 we show a basic high availability SAN design for multiple servers deployed with dual directors and a redundant fabric at each site. This design uses Dense Wavelength Division Multiplexing (DWDM) equipment for inter-site ISLs between the directors.

Note: Refer to Chapter 4, “Extending fiber over distance with DWDM” on page 183 for further information on DWDM technology and topologies.

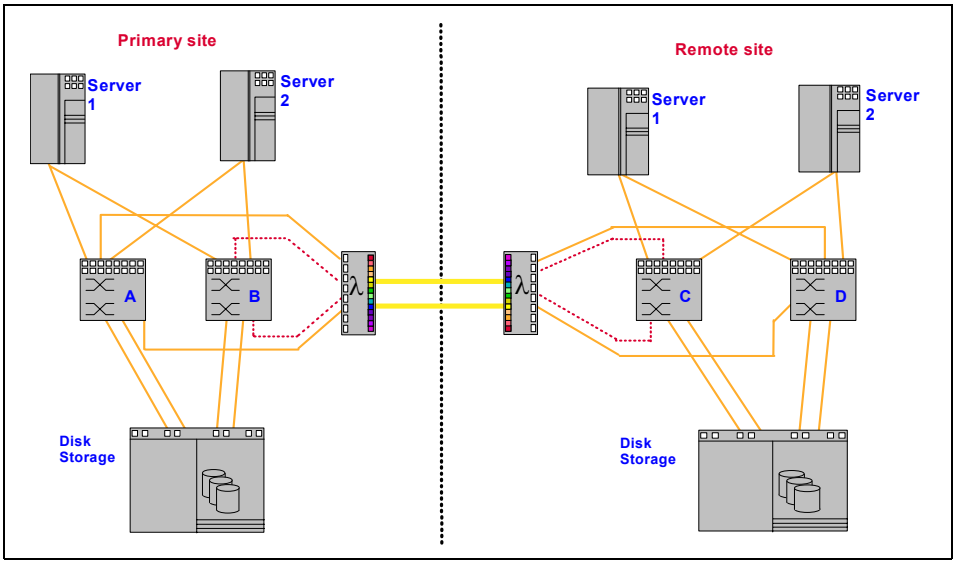


Figure 23-4 Point-to-point DWDM solution - two sites

The use of DWDM allows more channels to be sent over the same number of fiber cables. This is achieved by using different wavelengths of light to carry the different channels at the same time, and also supports transmission of multiple protocols (e.g. FC, ESCON, TCP/IP) over the same fibre. To ensure redundancy, and hence availability, the data should be sent over two completely different physical paths, but with equal length to ensure equal latency.

The method for replicating the data could be either host software based, or via storage hardware based facilities such as PPRC. In this example we are using host based software replication.

In our example, SAN AD (Directors A and D) will be a separate fabric from SAN BC (Directors B and C) fabric. Each fabric comprises of two directors, one at each site.

The two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 886, but with the addition of:

- DWDM equipment linked via dark fibre

Checklist

The same items as for “Two sites up to 10 km apart” on page 886 should be considered, with the addition of:

- The ISL link count to the DWDM equipment at each site

- Maximum distance for individual devices
- Configuration and count of the client interface adapters in the DWDM equipment
- Configuration and count of the transport network adapters in the DWDM, ensuring same frequency band and channel
- Connectivity testing between:
  - ESS to switches
  - Servers to switch
  - Switches to DWDM
- Validate connectivity
- Test server to storage on both paths
- Tests of failover/fail back operations
- Test inter-site connectivity via DWDM
- Setup LUNs on back up ESS and test from primary server, to quantify latency
- Pre-installation testing of the DWDM transport network for configuration and performance

Performance

The same performance criteria apply as for “Two sites up to 10 km apart” on page 886, with the addition of:

- The overall ISL latency increases with distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the director and the DWDM equipment. If the number of ISLs between the directors and DWDM equipment increases, it may be necessary to increase the quantity of dark fibre between the DWDMs that carry the ISL traffic.

Scalability

Most DWDMs support multiple I/O interfaces for client equipment and support up to 32 separate channels. DWDM equipment also allows mixing different interface protocols on a single physical DWDM channel.

The same scalability criteria apply as for “Two sites up to 10 km apart” on page 886.

Security

The same items as for “Two sites up to 10 km apart” on page 886 should be considered, with the addition of:

If you decide to add an ISL, ensure that all checks are done on the zone configuration changes from the management console.

- The DWDM equipment typically provides Web based management software that can be accessed by any workstation connected to the same LAN. Different user levels are provided for administrators, operators or observers. Different passwords must be set to limit access.
- The physical security of fiber connections and patch panels should be considered.

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 886, with the addition of:

- **DWDM optical channel card failure**

  As we configured the channels for high availability, there are redundant cards in the DWDM. If we lose one, the traffic will automatically be switched to the other and the channel will remain available. The failed card can be replaced concurrently.

- **DWDM optical channel manager card failure**

  The optical channel manager card performs path high availability switching. There are two cards in each shelf, if one fails the other takes control and the operation is not affected. The failed card can be replaced concurrently.

- **DWDM Optical Multiplexer failure or shelf backplane failure**

  The entire shelf will be unavailable. As we spread connections over different shelves, we will have at least half the channels of each type available. Operation will continue although performance may be affected.

- **Dark fiber failure**

  As we configured for high availability, operation will continue using the available pair with no performance impact. Customers may have additional considerations based on their business recovery policy to source the redundant dark fiber from another telco/source.

DWDM considerations

We have shown a point-to-point topology here. We would implement the DWDM solution as a protected fiber to ensure availability. In this example we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation, here we have assumed a light workload.

We show here an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one director in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits, however this is not a DWDM consideration but more to do with the general SAN solution design over a distance. The DWDM is a core architecture deployment and because of it's independence from protocol it can be used for other traffic beyond the SAN environment.

In order to maintain performance at extended distances, we need to increase the number of buffers on each interconnecting port to compensate for the number of frames that are in transit. Configuring the director ports connected to the DWDM for extended distance provides sufficient buffers for this distance.

23.6 Two sites: Point-to-point DWDM with ESS PPRC

With the IBM TotalStorage Enterprise Storage Server Model 800 running LIC 2.3.0 or later and the PPRC Version 2 licensed feature, it is no longer necessary to use dedicated ESCON ports on the ESS to connect via DWDM to a remote ESS. Remote copy traffic can now share FCP paths with normal SAN traffic, and hence the remote ESS can be accessed via the same connections as shown in Figure 23-4 on page 894.

Checklist

The same items as for "Two sites - Point-to-point DWDM solution" on page 893 should be considered, with the addition of:

- PPRC Version 2 license for the ESS
- Test PPRC - Establish paths and Test PPRC functionality - qualify timings

Performance

This is the same as for “Two sites - Point-to-point DWDM solution” on page 893.

Scalability

This is the same as for “Two sites - Point-to-point DWDM solution” on page 893.

Security

This is the same as for “Two sites - Point-to-point DWDM solution” on page 893.

“What if” failure scenarios

This is the same as for “Two sites - Point-to-point DWDM solution” on page 893.

DWDM considerations

In addition to those listed in “Two sites - Point-to-point DWDM solution” on page 893, we need to consider that the solution shown here also implements IBM’s PPRC between the two ESS subsystems. See the *IBM TotalStorage Enterprise Storage Server Implementing ESS Copy Services in Open Environments*, SG24-5757 for details on implementing PPRC with ESS.

23.7 Multiple site ring DWDM solution

Figure 23-5 illustrates four dual-fabric sites connected in a hubbed ring topology via DWDM and OADM technology. Site 1 is the hub node where all the DWDM channels originate and terminate, whilst the other three sites utilize OADM technology to add or remove channels from the ring.

Note: Refer to Chapter 4, “Extending fiber over distance with DWDM” on page 183 for further information on DWDM technology and topologies.

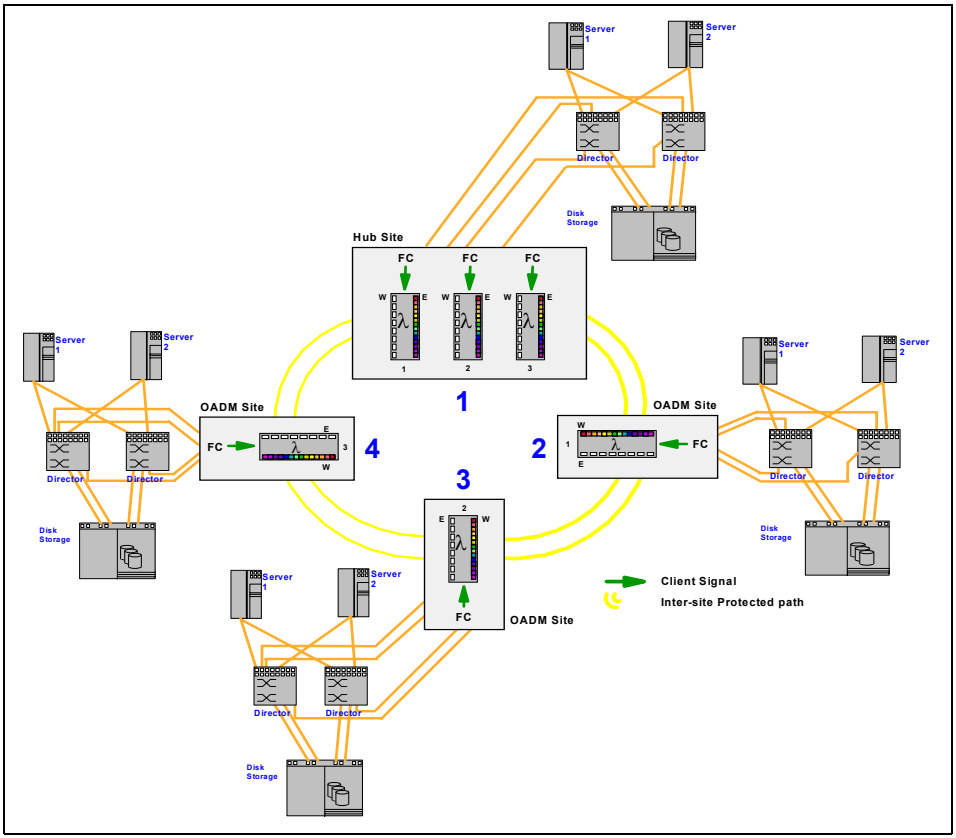


Figure 23-5 Multiple site: Ring topology DWDM solution

Each path has its own wavelength. The OADM pass through feature (express channels) enables non-contiguous sites to connect over an intermediate site as if they were directly connected. The only additional overhead of the pass through is the minimal latency (5 microseconds/km) of the second link. The pass through has no overhead since it is a passive device. The two fabrics would logically appear as fully meshed topologies.

Each of the links can operate in protected mode, which provides a redundant path in the event of a link failure. In most cases, link failures are automatically detected within 50 msec. In this case, the two wavelengths of the failed link reverse directions and reach the target port at the opposite side of the ring. If the link between 1 and 4 fails, the transmitted wavelength from 4 to 1 would reverse direction and reach 1 through 3 and 2. The transmitted wavelength from 1 to 4 would also reverse direction and reach 4 through 2 and 3.

Calculating the distance between nodes in a ring depends on the implementation of the protected path scheme. For instance, if the link between DWDM 2 and 3 fails, the path from 1 to 3 would be 1 to 2, back from 2 to 1 (due to the failed link), 1 to 4, and finally 4 to 3. This illustrates the need to utilize the entire ring circumference (and more, in a configuration with over four nodes) for failover.

Another way to calculate distance between nodes is to set up the protected path in advance (in the reverse direction) so the distance is limited to the number of hops between the two nodes. In either case, the maximum distance between nodes determines the maximum optical reach. An example of this specification is 80 to 100 km for a maximum distance between nodes and 160 to 400 km for maximum ring size. These distances should continue to increase as fiber optic technology advances.

Each site has a set of components the same as one site in the “Two sites - Point-to-point DWDM solution” on page 893, except:

- Site 1 has three sets of DWDM channels
- Sites 2, 3 and 4 each have OADM equipment
- Dark fibre between each site

Checklist

The same items as for “Two sites - Point-to-point DWDM solution” on page 893 should be considered, with the addition of:

- Installation and configuration of the OADM equipment
- Ensuring correct frequency band and channel for each site

Performance

This is the same as for “Two sites - Point-to-point DWDM solution” on page 893.

Scalability

This is the same as for “Two sites - Point-to-point DWDM solution” on page 893.

Security

This is the same as for “Two sites - Point-to-point DWDM solution” on page 893.

“What if” failure scenarios

This is the same as for “Two sites - Point-to-point DWDM solution” on page 893, except that the scenarios for DWDM failures are also applicable to OADMs.

DWDM considerations

We have shown a ring topology here, this gives a logical mesh, potentially giving any to any connectivity. We have implemented this as two rings, one ring connects to one switch in each site, the other ring connects to the remaining switch, this give two discrete SAN fabrics.

We would implement the DWDM solution as a protected ring to ensure availability. In this example we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation, here we have assumed a light workload.

We show here an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one director in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits, however this is not a DWDM consideration but more to do with the general SAN solution design over a distance. The DWDM is a core architecture deployment and because of it's independence from protocol it can be used for other traffic beyond the SAN environment.

In order to maintain performance at extended distances, we need to increase the number of buffers on each interconnecting port to compensate for the number of frames that are in transit. Configuring the director ports connected to the DWDM for extended distance provides sufficient buffers for this distance.

23.8 Two sites: Channel extender and WAN extension

Channel extenders typically use telecommunication lines for data transfer and therefore enable application and recovery sites to be located over greater distances. The use of channel extenders provides the separation for disaster recovery purposes and avoids some of the barriers imposed when customers do not have a “right of way” to lay fiber cable.

In Figure 23-6 we show a typical SAN distance extension using Optical Channel extenders at both primary and secondary sites and the ATM network.

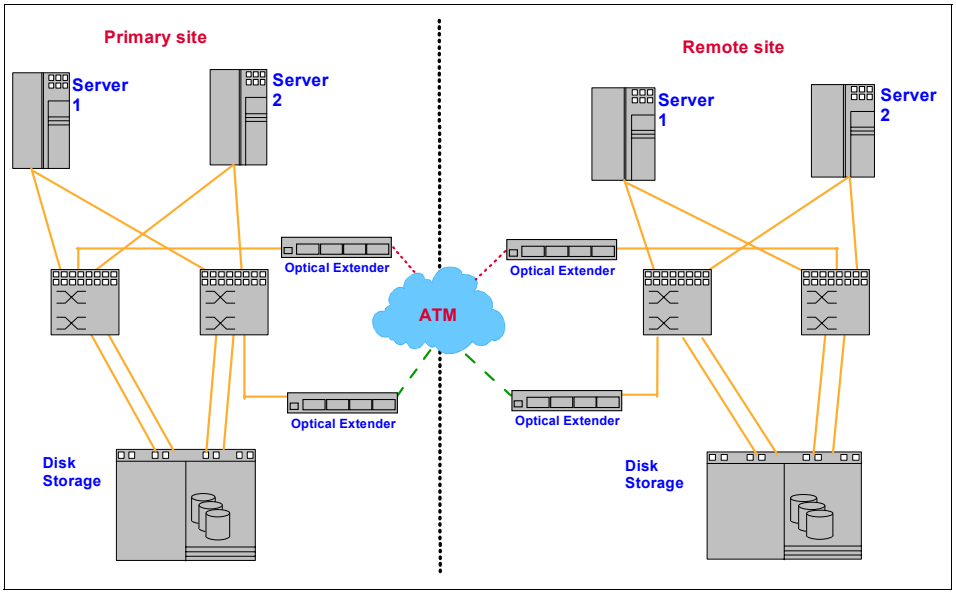


Figure 23-6 SAN extension over (ATM) WAN using channel extenders

Again, the two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 886, but with the addition of:

- Two optical channel extenders
- ATM already available

Checklist

The same items as for “Two sites up to 10 km apart” on page 886 should be considered, with the addition of:

- I/O interface (for example, FC or ESCON) selection and installation on the Optical channel extender
- Network interface (for example, ATM OC3) selection and installation on the Optical channel extender
- Pre-installation testing of the network interfaces (ATM network)

Performance

The same performance criteria apply as for “Two sites up to 10 km apart” on page 886, other than overall ISL latency increases with distance and BB_Credit utilization will increase.

Channel extenders generally compress the data before sending it over the transport network. However the compression ratio needs to be determined based on the individual application characteristics and the distance.

You must especially consider an amount of the updated data for a period of time (peak time), and reflect this in calculating the number of interconnecting links and the number of data volumes for SAN/WAN solution.

Scalability

The same scalability criteria apply as for “Two sites up to 10 km apart” on page 886. Most optical channel extenders support multiple I/O interfaces for client equipment. However in an installation, it is recommended to have a single protocol per interface on a physical channel extender.

Security

The same items as for “Two sites up to 10 km apart” on page 886 should be considered, with the addition of:

- Channel extender security — locked cabinet, restricted access site

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 886, with the addition of:

- **Channel extender failure**

  In the current solution channel extenders are not configured in a redundant mode, however they can be.

- **Transport network failure**

  The telco link failure is not planned for in the current solution, however a redundant link can be acquired from the vendor or a separate telco may be considered depending on the business recovery policy.

23.9 Two sites with FCIP connection

With the addition of FCIP it is possible to connect two remote SANs together. This enables servers or devices in the primary site to have access to SAN devices in the remote site over TCP/IP networks. The FC-FC Routing Service can use the FCIP Tunneling Service to make visible just those devices which are required from the remote site, without the need to merge the two fabrics, as would happen with pure FCIP.

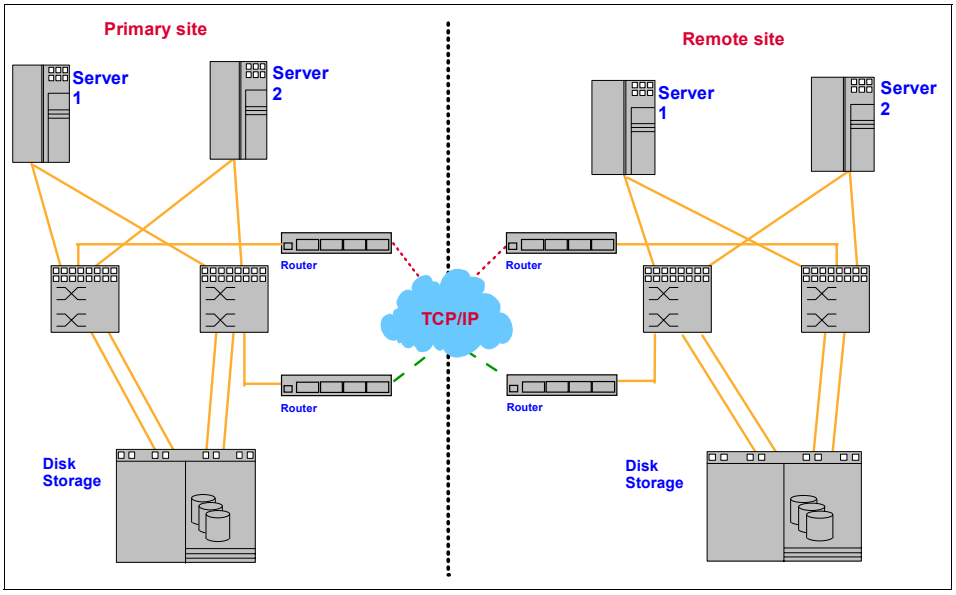


Figure 23-7 Remote access via FCIP router over TCP/IP

Again, the two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 886, but with the addition of:

- A routing interface for each director
- TCP/IP network already available

Checklist

The same items as for “Two sites up to 10 km apart” on page 886 should be considered, with the addition of:

- FCIP configuration
- FC-FC configuration
- Pre-installation testing of the network interfaces (TCP/IP network)

Performance

Extra latency will apply to those devices accessed via the TCP/IP network. For these devices, similar performance criteria apply as for “Two sites up to 10 km apart” on page 886, with the addition that overall latency increases with distance.

Scalability

As well as the normal SAN scalability issues, the WAN bandwidth and number of FCIP switch ports must be managed to support the required amount of remote traffic.

Security

The same items as for “Two sites up to 10 km apart” on page 886 should be considered, with the addition of:

- Security — locked cabinet, restricted access site

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 886, with the addition of:

- **Routing device failure**

  In the current solution, a router would cause loss of access to one of the remote fabrics. Since a dual fabric design has been used at both sites, access could still continue via the alternate fabric. It is possible to use more than one router to connect two fabrics, such that traffic can continue to flow through the second route in situations where the first route has failed.

- **TCP/IP network failure**

  WAN failure could cause complete loss of access to remote devices. Such situations are beyond the scope of this book.

23.10 Remote tape vaulting

For an existing large corporation with multiple sites, tape library resources can be consolidated to a separate site. This simplifies data movement logistics and centralizes backup software configurations. By doing this the cost of doing business is reduced as the infrastructure is efficiently utilized and fewer IT personnel required for backup data management. In addition, in the event of a disaster, the data is already located on tape in the remote location and there is no longer a need to ship the data to another site. Another application is for outsourcing services companies like Storage service providers (SSP), that are interested in providing backup and Backup management services to smaller companies.

The Figure 23-8, shows the basic layout for a remote tape vaulting solution, using IBM TotalStorage SAN Switches.

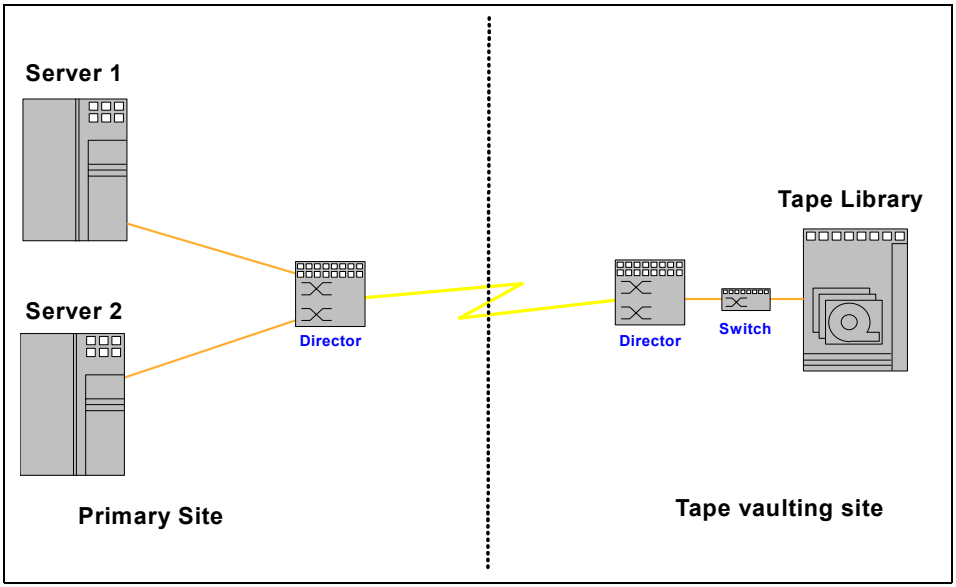


Figure 23-8 Remote tape vaulting

Any of the preceding distance solutions (dark fibre, repeaters, DWDM, iFCP, etc) can be used to connect the local and remote directors, depending on circumstances and the distance required.

Since most tape libraries fiber channel ports only support FC-AL, and IBM TotalStorage SAN Switch models M12 and M14 do not directly support FC-AL, the drives need to be connected via an Fxx or Hxx series switch. Some of the latest tape drives, such as the IBM TotalStorage Enterprise Tape Drive 3592, support direct switched fabric attachment, and so could be attached directly to the director. However, due to the relevant costs of director ports versus switch ports, it may still be advisable to use the 2026-224 switch as a tape concentrator.

If a director class product is not required at the remote site, then the connection can be made directly to the switch.

Checklist

In addition to the items considered in the disk consolidation example, we must now consider:

- Host HBA supported for tape drive / library attachment
- Use of separate HBAs for disk and tape access
- Host tape device driver levels
- Host operating system levels compatible with tape library fiber requirements

- 2026-224 switch configuration compatible with the rest of the fabric
- Zone configuration allowing tape access to required servers
- Host software tape sharing capabilities

Performance

The performance of the tape drives determines the number of drives which can be connected to the 2026-224, and hence the number of ISLs needed at the remote site to support the bandwidth. Installation of the Open Trunking feature on the directors and switch would also help to optimize ISL usage.

Certain operating system and device driver combinations support path failover for tape (e.g. AIX with Magstar 3590), and as such dual port tape drives can be connected to two 2026-224s for redundancy purposes.

Scalability

The solution can be scaled by increasing the number of tape drives in the library, and the number of associated 2026-224s. This may also require an increase in capacity of the offsite link.

Security

Here are some security considerations:

- Zoning can be used to restrict server access to specific tape devices as required.
- Proper tape management procedures will avoid servers contending for the same tape device.

“What if” failure scenarios

Here are some theoretical assumptions:

- Performance may be impacted depending on number of drives attached. Traditionally tape failover is a manual operation. Multiple path devices are configured as several logical devices, one per path. Only one of these logical devices is made active. If there is a failure the application aborts and it can then be restarted using a different logical device. For tape drives with dual paths, the latest levels of a tape device driver provide alternate pathing support and tape failover for Fibre Channel connections. With this support enabled, if an error occurs the device driver will automatically initiate error recovery and the operation will continue using the next logical path.
- Device link or device port failure — only one tape drive affected. Recovery may be manual or automatic depending on operating system and driver level as explained for ISL or switch failure.

- Switch port failures — SFPs are hot swappable. Drives can be moved to a spare port.
- Switch failure — most FRUs are hot swappable. With multi-port drives, access can continue via an alternate switch.
- User error trying to access too many drives on the same link — performance of all drives attached via the link may be degraded. Director or switch performance view may be used to find what paths are carrying traffic.
- Tape drive failure in a single tape zone — an alternate zone should be made active to get access to a working device.

23.11 Remote tape vaulting with disaster tolerance

In Figure 23-9 we show an extension or variant of the tape vaulting solution. In this solution the primary site / local site also has a tape library. Data is written to both tape libraries, however in the event of failure on the tape library at the local site, data can be backed up and restored from the tape library at the remote site.

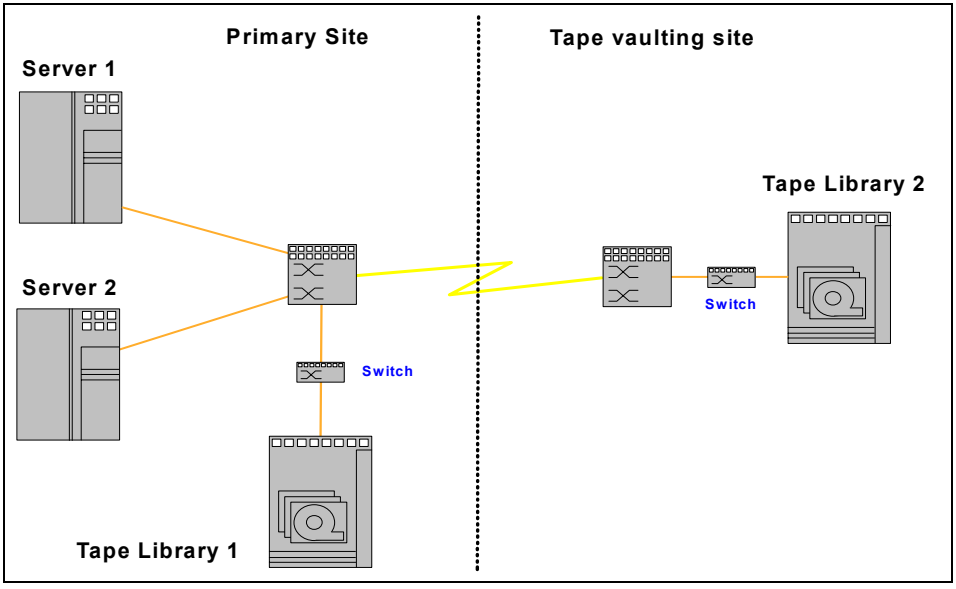


Figure 23-9 Remote tape vaulting with disaster tolerance

This solution has the same performance, scalability, security and failure attributes as the standard remote tape vaulting solution.

24

IBM TotalStorage SAN m-type family distance solutions

In this chapter we will discuss some solutions that can be implemented using the McDATA products included in the IBM portfolio.

For details of the McDATA portfolio products refer to:

- Chapter 10, “IBM TotalStorage SAN m-type family” on page 429
- *Implementing an Open IBM SAN,* SG24-6116

The solutions we will discuss are:

- Remote disk consolidation
- Two sites up to 10 km apart
- Two sites up to 100 km apart with repeaters
- Point-to-Point DWDM
- Point-to-Point DWDM with PPRC
- Ring topology DWDM
- Channel extender and WAN extension
- iFCP connection
- Remote tape vaulting
- Remote tape vaulting with redundancy

© Copyright IBM Corp. 2004. All rights reserved.

This chapter presents various implementations and uses for Storage Area Network (SAN) in high availability graphically dispersed environments. The considerations for a high availability SAN design using Fibre Channel switches and directors are covered in the following topics.

By reducing or eliminating single points of failure in the enterprise environment, SANs can help to improve overall availability of business applications. By utilizing highly available components and solutions, as well as a fault-tolerant design, enterprises can achieve the availability needed to support 24x7 uptime requirements. High distances in these solutions enable the SAN to cover a whole new host of IT availability issues, for example, we can deploy disaster tolerant solutions.

In network systems such as SANs, with their associated servers, fabric, and storage components, as well as software applications, downtime can occur even if parts of the system are highly available or fault tolerant. To improve business continuance under a variety of circumstances, SANs can incorporate redundant components, connections, software, and configurations to minimize or eliminate single points of failure.

With the emergence of the Internet and the proliferation of global e-business applications, more and more companies are implementing computing infrastructures specifically designed for continuous data and system availability. Today, even applications such as company e-mail have become mission critical for ongoing business operations. Faced with increased customer and internal user expectations, companies are currently striving to achieve at least 99.999 percent (“five nines”) availability in their computing systems — a figure equivalent to less than 5.3 minutes of downtime a year. Additional downtime can severely impact business operations and cost valuable time, money, and resources.

To ensure the highest level of system uptime, companies are implementing reliable storage networks capable of boosting the availability of data for all the users and applications that need it. These companies typically represent the industries that demand the highest levels of system and data availability — the utilities and telecommunications sector, brokerages and financial service institutions, and a wide variety of service providers.

24.1 High-level availability objectives

Developing highly available networks involves identifying specific availability requirements and predicting what potential failures might cause outages. The first step is to clearly define availability objectives, which can vary widely from company to company and even within segments of the same company. In some

environments, no disruption can be tolerated, while other environments might be only minimally affected by short outages. As a result, availability is a function of the frequency of outages (caused by unplanned failures or scheduled maintenance and upgrades) and the time to recover from such outages.

Many companies are addressing their availability requirements by implementing networks fabrics of Fibre Channel devices designed to provide high-performance storage environments. These flexible SANs are based on the following principles:

- A thorough understanding of availability requirements throughout the enterprise
- A flexible design that incorporates fault tolerance through redundancy and mirroring
- Simplified fault monitoring, diagnostics, and repair capabilities to ensure fast recovery
- A minimal amount of human intervention required during failover events
- A reliable backup and recovery plan to account for a wide variety of contingencies

24.2 Storage consolidation with remote disk

This remote disk consolidation solution is an extension of the most basic SAN solutions of disk consolidation (Figure 24-1).

Disk consolidation provides a means for storage administrators to take advantage of the capabilities of Fibre Channel fabric-based Storage Area Networks to organize storage in more flexible and efficient ways, easing many aspects of storage administration.

In an attempt to extend the distance between the storage and the servers, and in order to provide business benefits like consolidation, physical separation of disk storage from the server across sites, is necessary. This solution extends the distance to up to 10 km apart, using longwave SFPs in the SAN fabric, and single-mode fiber cable interconnect.

Note: McDATA supports a distance of up to 35 km using extended longwave SFPs between the directors.

The example is comprised of:

- Three dual-port hosts
- Four McDATA Intrepid 6064 Directors

- One IBM TotalStorage Enterprise Storage Server Model 800 with eight FC host adapters

The host servers and ESS are connected to the SAN fabric via shortwave SFPs using multi-mode fibre cable.

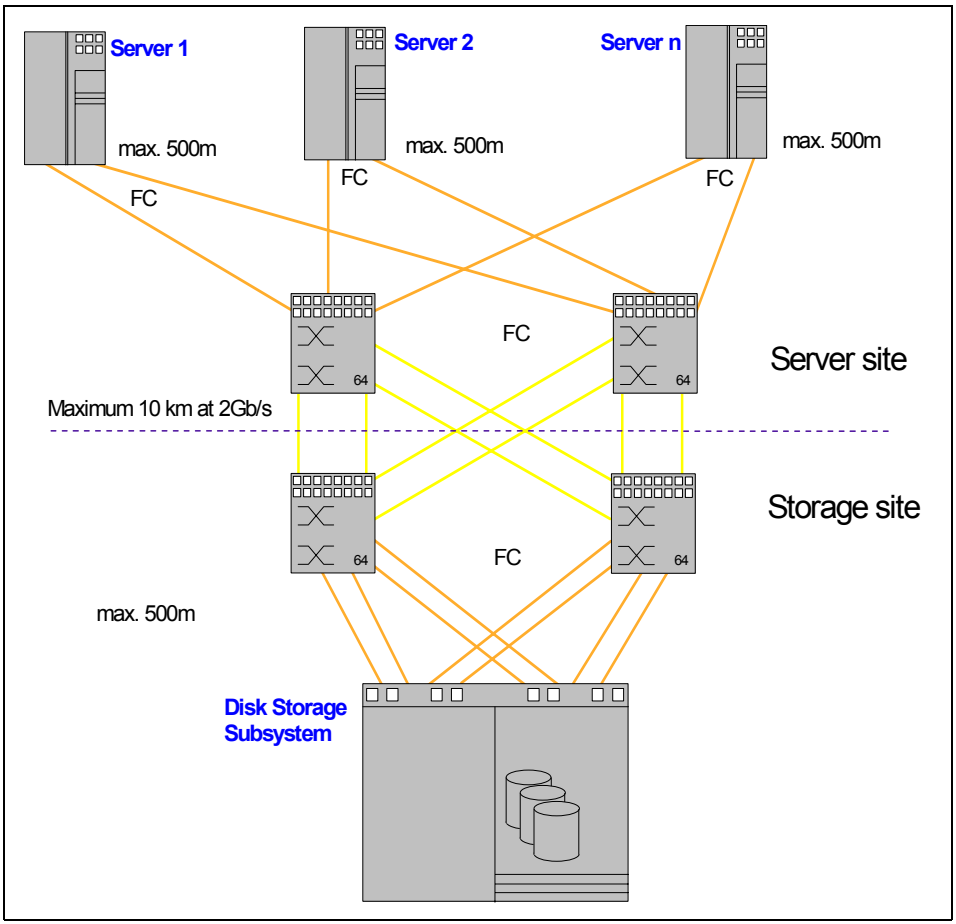


Figure 24-1 Remote disk consolidation

Checklist

We checked the following items:

- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and firmware level for proposed configuration

- McDATA switches and director longwave SFP count
- All cables clearly labelled and recorded
- The ISL links — single-mode fiber cable
- Single-mode FC cable laying and termination
- McDATA Director high availability features
- Nickname assignments so we can quickly cross reference WWNs to devices
- ESS LUN definitions done
- EFCM userids and passwords defined

Performance

In this simple implementation, performance will basically depend on the number of HBAs available on each server and the number of storage connections and most important the distance between storage and servers. It is important to consider application behavior over the distance, and this varies from application to application and we have assumed that the performance is adequate for applications running in the servers.

The McDATA director supports any to any connectivity so it will not affect performance by itself. Latency in the director is less than 2.0 microseconds; this is not relevant compared to storage device response times typically in the millisecond range.

Knowing the requirements of our servers we can decide whether it is enough or if we need to add more connections or even more storage devices.

Without having the exact requirements we may consider that for a high profile server, a server to storage ratio of 6:1 is acceptable. This is only a starting point, we will then implement some measurement system or use statistical data to decide whether we need to add more connections or we have more bandwidth than required.

Scalability

The McDATA director supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation. With four ISLs between the pair of directors on the host side and the pair of directors on the storage side, and four spare ports in a fully populated director, we can have 56 dual HBA servers (64 ports - 4 ISL - 4 spare = 56 server ports). With eight ESS connections, this slightly exceeds the 6:1 server to storage ratio (6 x 8 = 48).

Security

Here are some security considerations:

- ESS LUN masking by WWN will allow each server access only to configured LUNs.
- EFCM userids, passwords and user rights are defined and defaults are removed, so only authorized personnel can perform management functions.
- Remote access to EFCM configured to limit access to authorized workstations.
- Physical director security — locked cabinet, restricted access site.

"What if" failure scenarios

These are some theoretical assumptions:

- **Host HBA failure**

  SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. EFCM access and ESS Specialist access are required.

- **Storage host adapter failure**

  The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed eight host adapters, so a single adapter failure will reduce available bandwidth by 12.5%. For an average workload and three servers as shown it will not impact performance. When the host adapter is replaced we do not need to update zoning as the ESS WWN is maintained by the ESS.

- **Director port failure**

  The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port, as the target FC ID will have changed. Physical access to director and EFCM user with maintenance rights is required. OS root access may be required.

- **Fiber failure**

  Impact will depend on whether it is a server attachment, storage attachment or ISL fiber. For ISL fiber, path routes may be updated within the fabric. The only action required is fiber replacement. Physical access to director and attached device are required.

- **EFC Server failure**

No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example. Call home support will also be unavailable.

- **EFC Server hard drive failure**

  Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from CD-RW drive backup.

- **Director completely down, storage completely down, or site down (power, air conditioning, site damage**)

  This will cause an interruption in normal operation.

- **Physical damage to storage causing data loss (fire, flood)**

  We will need to restore data from backup tapes.

24.3 Two sites up to 10 km apart

In Figure 24-2, we show a basic dual fabric SAN distance extension solution between two sites up to 10 km apart. Each site has a fully redundant infrastructure at the SAN fabric level comprising of two McDATA Intrepid 6064 Directors. The host servers and storage are connected to the SAN fabric via shortwave SFPs using multi-mode fibre cable. The directors are interlinked across the sites via two ISLs each using longwave SFPs and single-mode fiber cables.

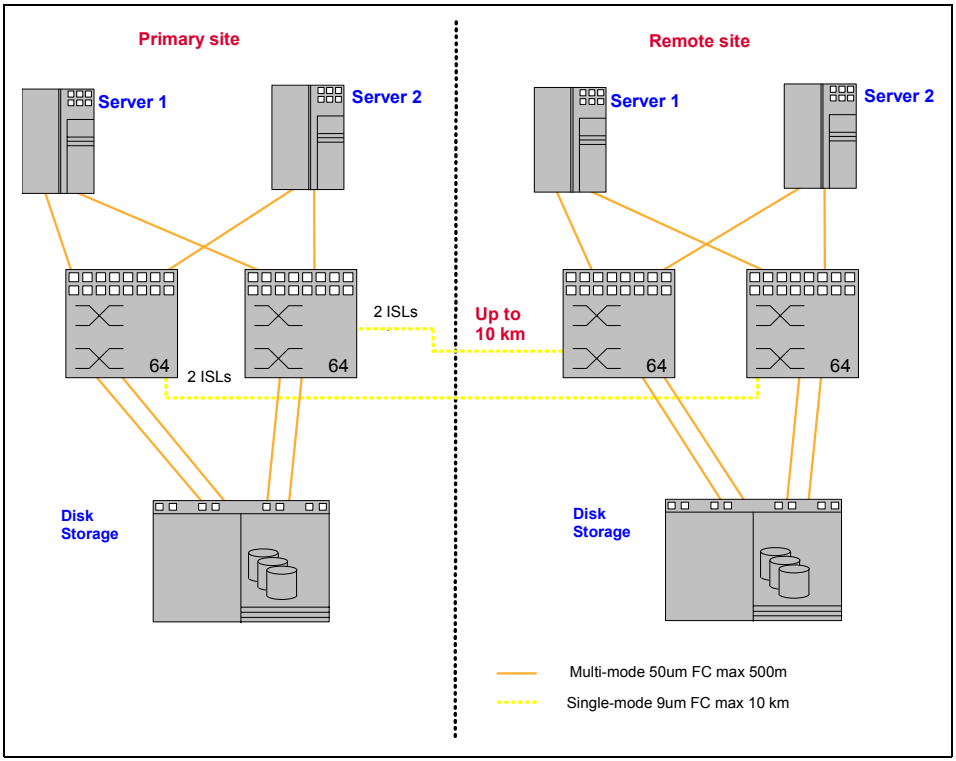


Figure 24-2 SAN distance extension up to 10 km

The solution can be used for accessing a database at the remote site as well as mirroring data at an alternative site up to 10 km.

Note: McDATA supports a distance of up to 35 km using extended longwave SFPs between the directors.

One can assume that a greater distance between IT sites results in greater security against wide spread disaster. However, increased distance has its price in terms of interconnection cost, business relocation effort, and network cost. Before you start to expand your storage network, you have to consider if the following solution can meet your business purpose.

The two sites are each comprised of:

- Two dual-port hosts
- Two McDATA Intrepid 6064 Directors
- One IBM TotalStorage Enterprise Storage Server Model 800 with four FC host adapters

- Software
  - IBM Subsystem Device Driver (SDD)
  - IBM StorWatch Enterprise Storage Server Specialist
  - IBM TotalStorage Expert for ESS (for historical disk performance data)

We have assumed the following for our solution:

- Both sites have SAN infrastructure.
- Server to director and director to storage use multi-mode fibre.
- Servers in production which are being used for mission critical production need data mirroring for disaster recovery.
- Some servers at the primary site need to access database in storage at the remote site, in addition to using data in the primary storage.
- Servers at the remote site are in standby-mode, and being used for backup servers at ordinary times.
- Single-mode FC cables are used as ISLs between directors at each site.
- Software mirroring is used on servers.
- No PPRC proposed.

Checklist

The following items should be considered to implement the solution:

- Fibre Channel HBA supported on servers
- Operating systems support software mirroring
- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and firmware level for proposed configuration
- Director longwave SFP count
- The ISL links — single-mode fiber cable
- Single-mode FC cable laying and termination
- E_D_TOV and R_A_TOV settings equal on all directors
- Extended distance buffering enabled for ISL ports to increase BB_Credit for 10-100 km links
- McDATA Director high availability features
- Nickname assignments so we can quickly cross reference WWNs to devices
- ESS LUN definitions done

- EFCM users and passwords defined

Performance

The major performance consideration with a long distance solution is calculating the correct number of lines between sites. This number can only be determined by performing a detailed performance profile of the servers and storage that will be remote. Over estimating the number of lines will increase costs dramatically, whilst under estimating the number of lines will dramatically impact the performance of the SAN.

It is vital that detailed performance data is available prior to sizing the number of lines required.

Typically, latency will increase over long distances – a good rule of thumb is 4.8 microseconds per kilometer. The default E_D_TOV and R_A_TOV values do not need to be modified for this distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the Directors across sites (the ISL links).

Scalability

The McDATA director supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation. With two ISLs between each director on the primary site and the director on the remote site, two ports for the ESS and four spare ports in a fully populated director, we can have 54 dual HBA servers (64 ports - 4 ISL - 2 ESS ports - 4 spare = 54 server ports). This may be too many for four ESS connections, depending on the type of server and workload, in which case more director ports may need to be used for ESS connections.

Security

Here are some security considerations:

- ESS LUN masking by WWN will allow each server access only to configured LUNs.
- EFCM user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions.
- Remote access to EFCM configured to limit access to authorized workstations.
- Physical director security — locked cabinet, restricted access site.
- The zoning information is separate for the two fabrics, so when changing the zoning in one fabric it is not automatically propagated to the other fabric.

- Our solution assumes we own private lines between sites so encryption is not required.
- For leased lines or managed services where lines are shared, encryption is normally an option available from the service provider.

"What if" failure scenarios

Here are some theoretical assumptions:

- **Host HBA failure**

  SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. EFCM access and ESS Specialist access are required.

- **Storage host adapter failure**

  The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce available bandwidth by 25%. For an average workload and two servers as shown it will not impact performance. When the host adapter is replaced we do not need to update zoning as the ESS WWN is maintained by the ESS.

- **Director port failure**

  The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port, as the target FC ID will have changed. Physical access to director and EFCM user with maintenance rights is required. OS root access may be required.

- **Fiber failure**

  Impact will depend on whether it is a server attachment, storage attachment or ISL fiber. For ISL fiber, path routes may be updated within the fabric. The only action required is fiber replacement. Physical access to director and attached device may be required. For the ISL fibres between sites, the fibre provider will be responsible for repair.

- **EFC Server failure**

  No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example. Call home support will also be unavailable.

- **EFC Server hard drive failure**

Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from CD-RW drive backup.

- **Director completely down, storage completely down, or site down (power, air conditioning, site damage)**

  This will cause an interruption in normal operation.

- **Physical damage to storage causing data loss (fire, flood)**

  We will need to restore data from backup tapes.

24.4 Two sites up to 100 km apart

In Figure 24-3 we show a basic dual fabric SAN distance extension solution between two sites up to 100 km apart. Each site has a fully redundant infrastructure at the SAN fabric level comprising of two McDATA Intrepid 6064 Directors. The host servers and storage are connected to the SAN fabric via shortwave SFPs using multi-mode fibre cable. The directors are interlinked across the sites via two ISLs each using longwave SFPs and single-mode fiber cables and repeaters. The number of repeaters used depends on the type of longwave SFP used.

One can assume that a greater distance between IT sites results in greater security against wide spread disaster. However, increased distance has its price in terms of interconnection cost, business relocation effort, and network cost. Before you start to expand your storage network, you have to consider if the following solution can meet your business purpose.

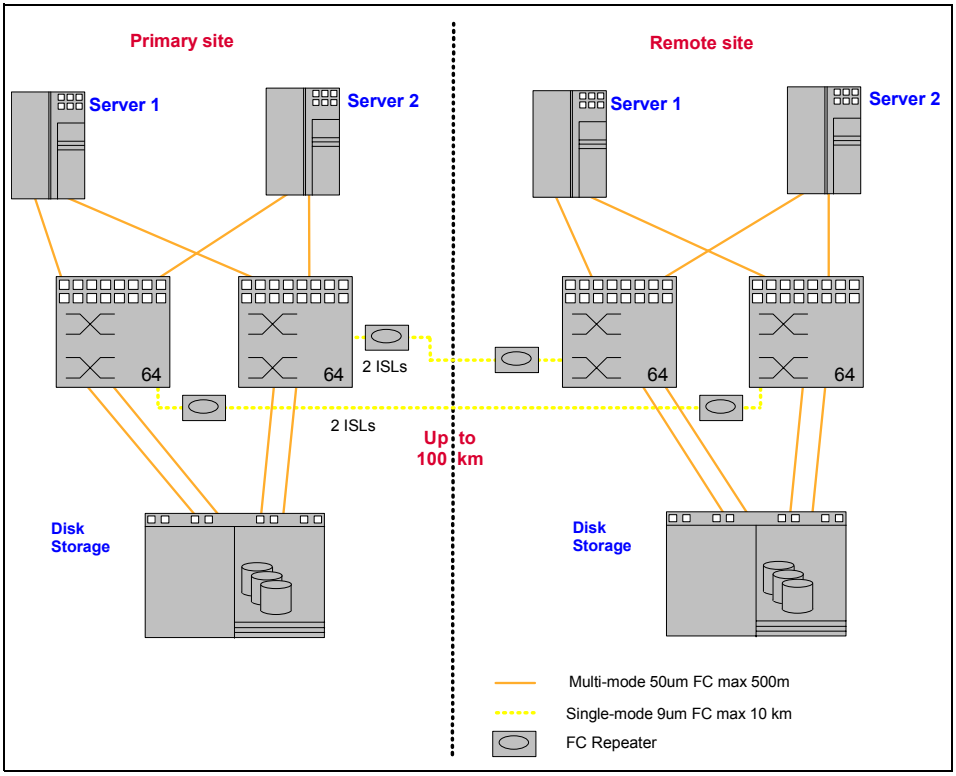


Figure 24-3 SAN distance extension up to 100 km with repeaters

The two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 915, but with the addition of:

- FC repeaters as required to reach 100 km

Checklist

The same items as for “Two sites up to 10 km apart” on page 915 should be considered, with the addition of:

- Repeater count verified for the distance in every ISL link
- Repeaters installed connected and tested

Performance

The same performance criteria apply as for “Two sites up to 10 km apart” on page 915, other than overall ISL latency increases with distance and BB_Credit utilization will increase.

Scalability

The same scalability criteria apply as for “Two sites up to 10 km apart” on page 915.

Security

The same items as for “Two sites up to 10 km apart” on page 915 should be considered, with the addition of:

- Physical repeater security — locked cabinet, restricted access site.

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 915, with the addition of:

- **Repeater failure**

  Impact will depend on whether alternate ISL links are available or not. Only action required is repeater replacement. Physical access to repeater and repeater site required.

24.5 Two sites - Point-to-point DWDM solution

In Figure 24-4 we show a basic high availability SAN design for multiple servers deployed with dual directors and a redundant fabric at each site. This design uses Dense Wavelength Division Multiplexing (DWDM) equipment for inter-site ISLs between the directors.

Note: Refer to Chapter 4, “Extending fiber over distance with DWDM” on page 183 for further information on DWDM technology and topologies.

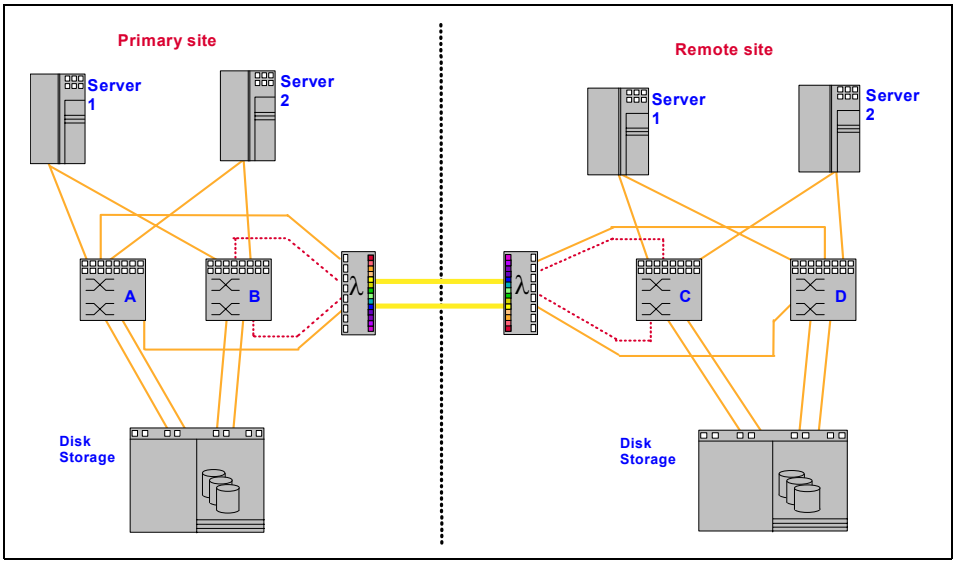


Figure 24-4 Point-to-point DWDM solution - two sites

The use of DWDM allows more channels to be sent over the same number of fiber cables. This is achieved by using different wavelengths of light to carry the different channels at the same time, and also supports transmission of multiple protocols (e.g. FC, ESCON, TCP/IP) over the same fibre. To ensure redundancy, and hence availability, the data should be sent over two completely different physical paths, but with equal length to ensure equal latency.

The method for replicating the data could be either host software based, or via storage hardware based facilities such as PPRC. In this example we are using host based software replication.

In our example, SAN AD (Directors A and D) will be a separate fabric from SAN BC (Directors B and C) fabric. Each fabric comprises of two directors, one at each site.

The two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 915, but with the addition of:

- DWDM equipment linked via dark fibre

Checklist

The same items as for “Two sites up to 10 km apart” on page 915 should be considered, with the addition of:

- The ISL link count to the DWDM equipment at each site

- Maximum distance for individual devices
- Configuration and count of the client interface adapters in the DWDM equipment
- Configuration and count of the transport network adapters in the DWDM, ensuring same frequency band and channel
- Connectivity testing between:
  - ESS to switches
  - Servers to switch
  - Switches to DWDM
- Validate connectivity
- Test server to storage on both paths
- Tests of failover/fail back operations
- Test inter-site connectivity via DWDM
- Setup LUNs on back up ESS and test from primary server, to quantify latency
- Pre-installation testing of the DWDM transport network for configuration and performance

Performance

The same performance criteria apply as for “Two sites up to 10 km apart” on page 915, with the addition of:

- The overall ISL latency increases with distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the director and the DWDM equipment. If the number of ISLs between the directors and DWDM equipment increases, it may be necessary to increase the quantity of dark fibre between the DWDMs that carry the ISL traffic.

Scalability

Most DWDMs support multiple I/O interfaces for client equipment and support up to 32 separate channels. DWDM equipment also allows mixing different interface protocols on a single physical DWDM channel.

The same scalability criteria apply as for “Two sites up to 10 km apart” on page 915.

Security

The same items as for “Two sites up to 10 km apart” on page 915 should be considered, with the addition of:

If you decide to add an ISL, ensure that all checks are done on the zone configuration changes from the management console.

- The DWDM equipment typically provides Web based management software that can be accessed by any workstation connected to the same LAN. Different user levels are provided for administrators, operators or observers. Different passwords must be set to limit access.
- The physical security of fiber connections and patch panels should be considered.

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 915, with the addition of:

- **DWDM optical channel card failure**

  As we configured the channels for high availability, there are redundant cards in the DWDM. If we lose one, the traffic will automatically be switched to the other and the channel will remain available. The failed card can be replaced concurrently.

- **DWDM optical channel manager card failure**

  The optical channel manager card performs path high availability switching. There are two cards in each shelf, if one fails the other takes control and the operation is not affected. The failed card can be replaced concurrently.

- **DWDM Optical Multiplexer failure or shelf backplane failure**

  The entire shelf will be unavailable. As we spread connections over different shelves, we will have at least half the channels of each type available. Operation will continue although performance may be affected.

- **Dark fiber failure**

  As we configured for high availability, operation will continue using the available pair with no performance impact. Customers may have additional considerations based on their business recovery policy to source the redundant dark fiber from another telco/source.

DWDM considerations

We have shown a point-to-point topology here. We would implement the DWDM solution as a protected fiber to ensure availability. In this example we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation, here we have assumed a light workload.

We show here an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one director in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits, however this is not a DWDM consideration but more to do with the general SAN solution design over a distance. The DWDM is a core architecture deployment and because of it's independence from protocol it can be used for other traffic beyond the SAN environment.

In order to maintain performance at extended distances, we need to increase the number of buffers on each interconnecting port to compensate for the number of frames that are in transit. Configuring the director ports connected to the DWDM for 10 to 100 km provides 60 buffers, that is enough for this distance.

24.6 Two sites: Point-to-point DWDM with ESS PPRC

With the IBM TotalStorage Enterprise Storage Server Model 800 running LIC 2.3.0 or later and the PPRC Version 2 licensed feature, it is no longer necessary to use dedicated ESCON ports on the ESS to connect via DWDM to a remote ESS. Remote copy traffic can now share FCP paths with normal SAN traffic, and hence the remote ESS can be accessed via the same connections as shown in Figure 24-4 on page 923.

Checklist

The same items as for "Two sites - Point-to-point DWDM solution" on page 922 should be considered, with the addition of:

- PPRC Version 2 license for the ESS
- Test PPRC - Establish paths and Test PPRC functionality - qualify timings

Performance
This is the same as for “Two sites - Point-to-point DWDM solution” on page 922.

Scalability
This is the same as for “Two sites - Point-to-point DWDM solution” on page 922.

Security
This is the same as for “Two sites - Point-to-point DWDM solution” on page 922.

“What if” failure scenarios
This is the same as for “Two sites - Point-to-point DWDM solution” on page 922.

DWDM considerations
In addition to those listed in “Two sites - Point-to-point DWDM solution” on page 922, we need to consider that the solution shown here also implements IBM’s PPRC between the two ESS subsystems. See the *IBM TotalStorage Enterprise Storage Server Implementing ESS Copy Services in Open Environments*, SG24-5757 for details on implementing PPRC with ESS.

24.7 Multiple site ring DWDM solution

Figure 24-5 illustrates four dual-fabric sites connected in a hubbed ring topology via DWDM and OADM technology. Site 1 is the hub node where all the DWDM channels originate and terminate, whilst the other three sites utilize OADM technology to add or remove channels from the ring.

Note: Refer to Chapter 4, “Extending fiber over distance with DWDM” on page 183 for further information on DWDM technology and topologies.

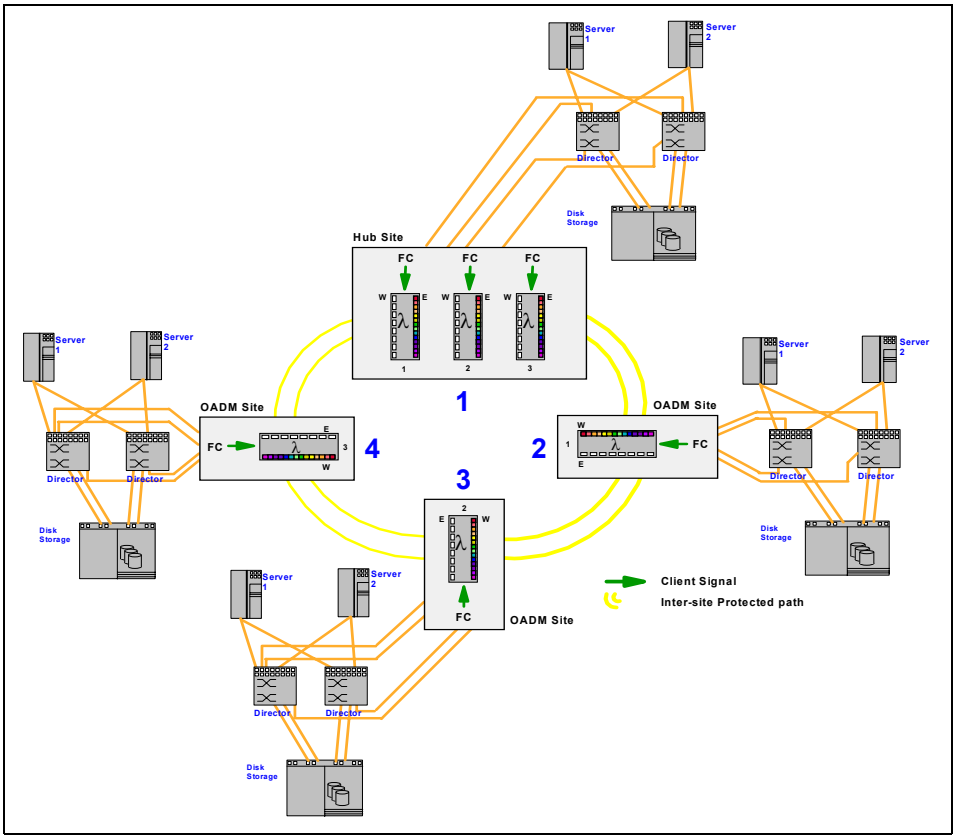


Figure 24-5 Multiple site: Ring topology DWDM solution

Each path has its own wavelength. The OADM pass through feature (express channels) enables non-contiguous sites to connect over an intermediate site as if they were directly connected. The only additional overhead of the pass through is the minimal latency (5 microseconds/km) of the second link. The pass through has no overhead since it is a passive device. The two fabrics would logically appear as fully meshed topologies.

Each of the links can operate in protected mode, which provides a redundant path in the event of a link failure. In most cases, link failures are automatically detected within 50 msec. In this case, the two wavelengths of the failed link reverse directions and reach the target port at the opposite side of the ring. If the link between 1 and 4 fails, the transmitted wavelength from 4 to 1 would reverse direction and reach 1 through 3 and 2. The transmitted wavelength from 1 to 4 would also reverse direction and reach 4 through 2 and 3.

Calculating the distance between nodes in a ring depends on the implementation of the protected path scheme. For instance, if the link between DWDM 2 and 3 fails, the path from 1 to 3 would be 1 to 2, back from 2 to 1 (due to the failed link), 1 to 4, and finally 4 to 3. This illustrates the need to utilize the entire ring circumference (and more, in a configuration with over four nodes) for failover.

Another way to calculate distance between nodes is to set up the protected path in advance (in the reverse direction) so the distance is limited to the number of hops between the two nodes. In either case, the maximum distance between nodes determines the maximum optical reach. An example of this specification is 80 to 100 km for a maximum distance between nodes and 160 to 400 km for maximum ring size. These distances should continue to increase as fiber optic technology advances.

Each site has a set of components the same as one site in the “Two sites - Point-to-point DWDM solution” on page 922, except:

- Site 1 has three sets of DWDM channels
- Sites 2, 3 and 4 each have OADM equipment
- Dark fibre between each site

Checklist

The same items as for “Two sites - Point-to-point DWDM solution” on page 922 should be considered, with the addition of:

- Installation and configuration of the OADM equipment
- Ensuring correct frequency band and channel for each site

Performance

This is the same as for “Two sites - Point-to-point DWDM solution” on page 922.

Scalability

This is the same as for “Two sites - Point-to-point DWDM solution” on page 922.

Security

This is the same as for “Two sites - Point-to-point DWDM solution” on page 922.

“What if” failure scenarios

This is the same as for “Two sites - Point-to-point DWDM solution” on page 922, except that the scenarios for DWDM failures are also applicable to OADMs.

DWDM considerations

We have shown a ring topology here, this gives a logical mesh, potentially giving any to any connectivity. We have implemented this as two rings, one ring connects to one switch in each site, the other ring connects to the remaining switch, this give two discrete SAN fabrics.

We would implement the DWDM solution as a protected ring to ensure availability. In this example we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation, here we have assumed a light workload.

We show here an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one director in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits, however this is not a DWDM consideration but more to do with the general SAN solution design over a distance. The DWDM is a core architecture deployment and because of it's independence from protocol it can be used for other traffic beyond the SAN environment.

In order to maintain performance at extended distances, we need to increase the number of buffers on each interconnecting port to compensate for the number of frames that are in transit. Configuring the director ports connected to the DWDM for 10 to 100 km provides 60 buffers, that is enough for this distance.

24.8 Two sites: Channel extender and WAN extension

Channel extenders typically use telecommunication lines for data transfer and therefore enable application and recovery sites to be located over greater distances. The use of channel extenders provides the separation for disaster recovery purposes and avoids some of the barriers imposed when customers do not have a "right of way" to lay fiber cable.

In Figure 24-6 we show a typical SAN distance extension using Optical Channel extenders at both primary and secondary sites and the ATM network.

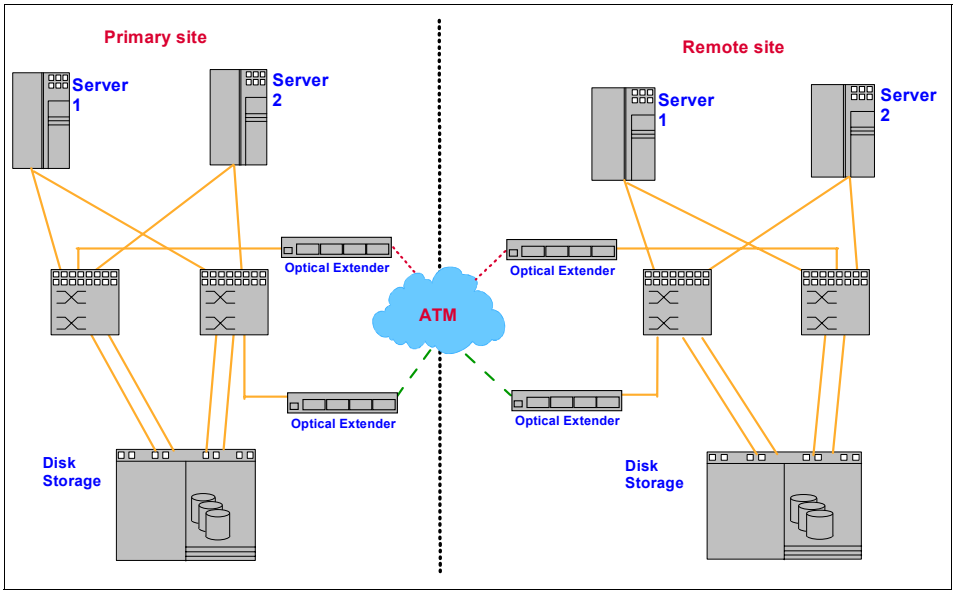


Figure 24-6 SAN extension over (ATM) WAN using channel extenders

Again, the two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 915, but with the addition of:

- Two optical channel extenders
- ATM already available

Checklist

The same items as for “Two sites up to 10 km apart” on page 915 should be considered, with the addition of:

- I/O interface (for example, FC or ESCON) selection and installation on the Optical channel extender
- Network interface (for example, ATM OC3) selection and installation on the Optical channel extender
- Pre-installation testing of the network interfaces (ATM network)

Performance

The same performance criteria apply as for “Two sites up to 10 km apart” on page 915, other than overall ISL latency increases with distance and BB_Credit utilization will increase.

Channel extenders generally compress the data before sending it over the transport network. However the compression ratio needs to be determined based on the individual application characteristics and the distance.

You must especially consider an amount of the updated data for a period of time (peak time), and reflect this in calculating the number of interconnecting links and the number of data volumes for SAN/WAN solution.

Scalability

The same scalability criteria apply as for “Two sites up to 10 km apart” on page 915. Most optical channel extenders support multiple I/O interfaces for client equipment. However in an installation, it is recommended to have a single protocol per interface on a physical channel extender.

Security

The same items as for “Two sites up to 10 km apart” on page 915 should be considered, with the addition of:

- Channel extender security — locked cabinet, restricted access site

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 915, with the addition of:

- **Channel extender failure**

  In the current solution channel extenders are not configured in a redundant mode, however they can be.

- **Transport network failure**

  The telco link failure is not planned for in the current solution, however a redundant link can be acquired from the vendor or a separate telco may be considered depending on the business recovery policy.

24.9 Two sites with iFCP connection

iFCP enables servers or devices in the primary site to have selective access to SAN devices in the remote site over TCP/IP networks. Unlike FCIP and the other distance solutions presented in this chapter, an iFCP connection does not cause the local and remote fabrics to merge into one fabric. Hence the two dual fabrics shown at each site in Figure 24-7 remain separate.

The iFCP switch connects using normal fibre channel to an E_Port on the director, but the ISL effectively terminates in the iFCP switch. This avoids the

performance overheads introduced by extended distance ISL latency when fabric-wide events occur such as principal switch selection, FSPF route selection and zone updates. It also provides a way to connect SAN switches from different vendors which would not normally interoperate.

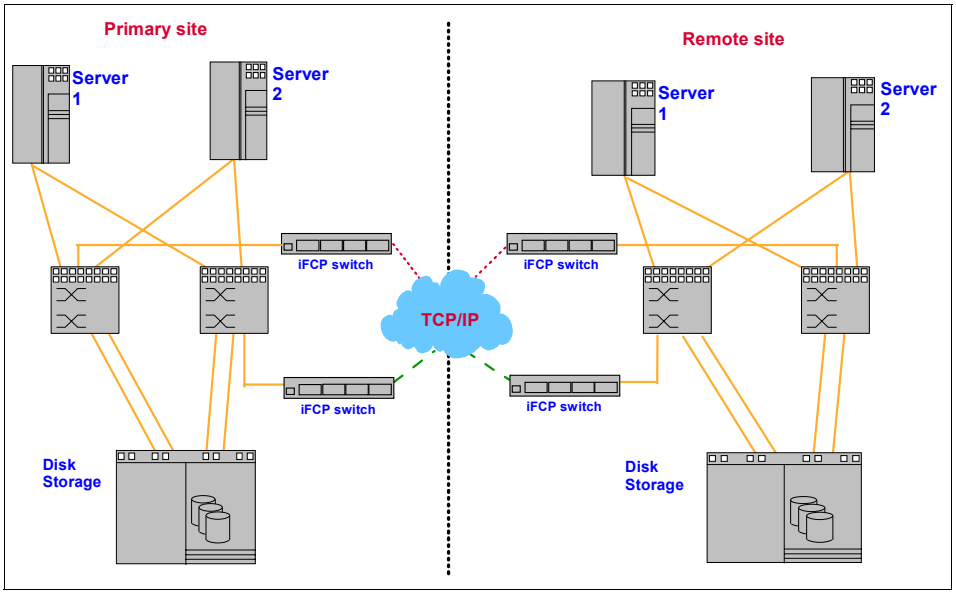


Figure 24-7 Remote access via iFCP switch over TCP/IP

Again, the two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 915, but with the addition of:

- iFCP switches for each fabric
- TCP/IP network already available

Checklist

The same items as for “Two sites up to 10 km apart” on page 915 should be considered, with the addition of:

- iFCP switch installation and configuration
- Pre-installation testing of the network interfaces (TCP/IP network)

Performance

Since the local and remote site fabrics are not merged, extra latency only applies to those devices accessed via the TCP/IP network. For these devices, similar performance criteria apply as for “Two sites up to 10 km apart” on page 915, with the addition that overall latency increases with distance.

The iFCP switch includes compression to reduce WAN bandwidth requirements, and Fast Write technology to improve performance over long distances.

Scalability

As well as the normal SAN scalability issues, the WAN bandwidth and number of iFCP switch ports must be managed to support the required amount of remote traffic.

Security

The same items as for "Two sites up to 10 km apart" on page 915 should be considered, with the addition of:

- iFCP switch security — locked cabinet, restricted access site

"What if" failure scenarios

The same failure scenarios apply as for "Two sites up to 10 km apart" on page 915, with the addition of:

- **iFCP switch failure**

  In the current solution, an iFCP switch failure would cause loss of access to one of the remote fabrics. Since a dual fabric design has been used at both sites, access could still continue via the alternate fabric. It is possible to use more than one iFCP switch to connect two fabrics, such that traffic can continue to flow through the second switch in situations where the first switch has failed.

- **TCP/IP network failure**

  WAN failure could cause complete loss of access to remote devices. Such situations are beyond the scope of this book.

24.10 Remote tape vaulting

For an existing large corporation with multiple sites, tape library resources can be consolidated to a separate site. This simplifies data movement logistics and centralizes backup software configurations. By doing this the cost of doing business is reduced as the infrastructure is efficiently utilized and fewer IT personnel required for backup data management. In addition, in the event of a disaster, the data is already located on tape in the remote location and there is no longer a need to ship the data to another site. Another application is for outsourcing services companies like Storage service providers (SSP), that are interested in providing backup and Backup management services to smaller companies.

The Figure 24-8, shows the basic layout for a remote tape vaulting solution, using McDATA directors and switches.

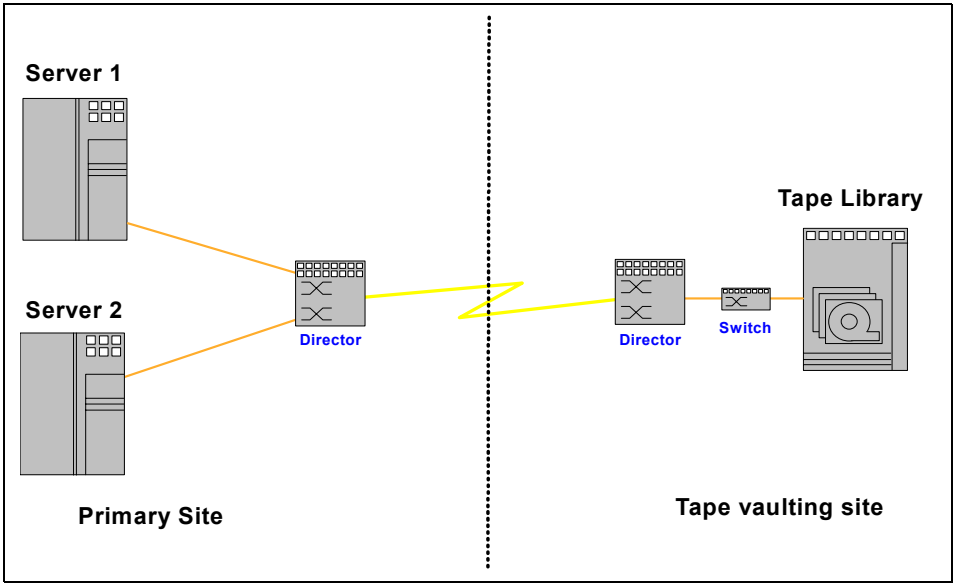


Figure 24-8 Remote tape vaulting

Any of the preceding distance solutions (dark fibre, repeaters, DWDM, iFCP, etc) can be used to connect the local and remote directors, depending on circumstances and the distance required.

Since most tape libraries fiber channel ports only support FC-AL, and McDATA directors do not directly support FC-AL, the drives need to be connected via an 2026-224 switch. Some of the latest tape drives, such as the IBM TotalStorage Enterprise Tape Drive 3592, support direct switched fabric attachment, and so could be attached directly to the director. However, due to the relevant costs of director ports versus switch ports, it may still be advisable to use the 2026-224 switch as a tape concentrator.

If a director class product is not required at the remote site, then the connection can be made directly to the switch.

Checklist

In addition to the items considered in the disk consolidation example, we must now consider:

- Host HBA supported for tape drive / library attachment
- Use of separate HBAs for disk and tape access

- Host tape device driver levels
- Host operating system levels compatible with tape library fiber requirements
- 2026-224 switch configuration compatible with the rest of the fabric
- Zone configuration allowing tape access to required servers
- Host software tape sharing capabilities

Performance

The performance of the tape drives determines the number of drives which can be connected to the 2026-224, and hence the number of ISLs needed at the remote site to support the bandwidth. Installation of the Open Trunking feature on the directors and switch would also help to optimize ISL usage.

Certain operating system and device driver combinations support path failover for tape (e.g. AIX with Magstar 3590), and as such dual port tape drives can be connected to two 2026-224s for redundancy purposes.

Scalability

The solution can be scaled by increasing the number of tape drives in the library, and the number of associated 2026-224s. This may also require an increase in capacity of the offsite link.

Security

Here are some security considerations:

- Zoning can be used to restrict server access to specific tape devices as required.
- Proper tape management procedures will avoid servers contending for the same tape device.

“What if” failure scenarios

Here are some theoretical assumptions:

- Performance may be impacted depending on number of drives attached. Traditionally tape failover is a manual operation. Multiple path devices are configured as several logical devices, one per path. Only one of these logical devices is made active. If there is a failure the application aborts and it can then be restarted using a different logical device. For tape drives with dual paths, the latest levels of a tape device driver provide alternate pathing support and tape failover for Fibre Channel connections. With this support enabled, if an error occurs the device driver will automatically initiate error recovery and the operation will continue using the next logical path.

- Device link or device port failure — only one tape drive affected. Recovery may be manual or automatic depending on operating system and driver level as explained for ISL or switch failure.
- Switch port failures — SFPs are hot swappable. Drives can be moved to a spare port.
- Switch failure — most FRUs are hot swappable. With multi-port drives, access can continue via an alternate switch.
- User error trying to access too many drives on the same link — performance of all drives attached via the link may be degraded. Director or switch performance view may be used to find what paths are carrying traffic.
- Tape drive failure in a single tape zone — an alternate zone should be made active to get access to a working device.

24.11 Remote tape vaulting with disaster tolerance

In Figure 24-9 we show an extension or variant of the tape vaulting solution. In this solution the primary site / local site also has a tape library. Data is written to both tape libraries, however in the event of failure on the tape library at the local site, data can be backed up and restored from the tape library at the remote site.

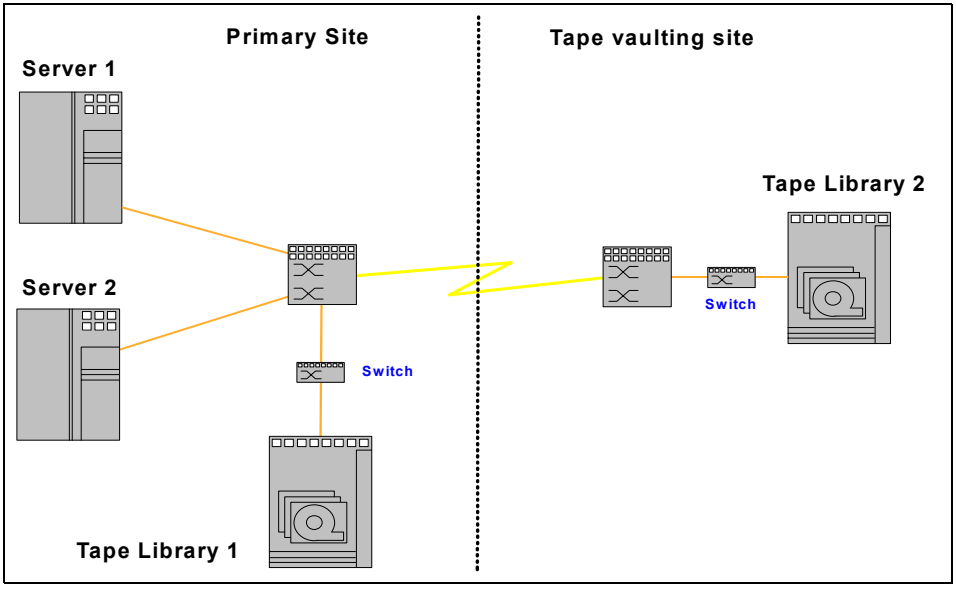


Figure 24-9 Remote tape vaulting with disaster tolerance

This solution has the same performance, scalability, security and failure attributes as the standard remote tape vaulting solution.

25

IBM TotalStorage SAN n-type family distance solutions

In this chapter we will discuss some solutions that can be implemented using the CNT products included in the IBM portfolio.

We will combine these with some of IBM's storage and other vendors' DWDM solutions to describe high distance SAN solutions.

For product details of the INRANGE FC/9000 refer to:

- Chapter 9, "IBM TotalStorage SAN n-type family and replication appliance" on page 345
- *Implementing an Open IBM SAN,* SG24-6116

The solutions we will discuss are:

- Remote disk consolidation
- Two sites up to 10 km apart
- Two sites up to 100 km apart with repeaters
- Point-to-Point DWDM
- Point-to-Point DWDM with PPRC
- Ring topology DWDM
- Channel extender and WAN extension

© Copyright IBM Corp. 2004. All rights reserved.

- FCIP connection
- Remote tape vaulting
- Remote tape vaulting with redundancy

This chapter presents various implementations and uses for Storage Area Network (SAN) in high availability graphically dispersed environments. The considerations for a high availability SAN design using Fibre Channel switches and directors are covered in the following topics.

By reducing or eliminating single points of failure in the enterprise environment, SANs can help to improve overall availability of business applications. By utilizing highly available components and solutions, as well as a fault-tolerant design, enterprises can achieve the availability needed to support 24x7 uptime requirements. High distances in these solutions enable the SAN to cover a whole new host of IT availability issues, for example, we can deploy disaster tolerant solutions.

In network systems such as SANs, with their associated servers, fabric, and storage components, as well as software applications, downtime can occur even if parts of the system are highly available or fault tolerant. To improve business continuance under a variety of circumstances, SANs can incorporate redundant components, connections, software, and configurations to minimize or eliminate single points of failure.

With the emergence of the Internet and the proliferation of global e-business applications, more and more companies are implementing computing infrastructures specifically designed for continuous data and system availability. Today, even applications such as company e-mail have become mission critical for ongoing business operations. Faced with increased customer and internal user expectations, companies are currently striving to achieve at least 99.999 percent (“five nines”) availability in their computing systems — a figure equivalent to less than 5.3 minutes of downtime a year. Additional downtime can severely impact business operations and cost valuable time, money, and resources.

To ensure the highest level of system uptime, companies are implementing reliable storage networks capable of boosting the availability of data for all the users and applications that need it. These companies typically represent the industries that demand the highest levels of system and data availability — the utilities and telecommunications sector, brokerages and financial service institutions, and a wide variety of service providers.

25.1 High-level availability objectives

Developing highly available networks involves identifying specific availability requirements and predicting what potential failures might cause outages. The first step is to clearly define availability objectives, which can vary widely from company to company and even within segments of the same company. In some environments, no disruption can be tolerated, while other environments might be only minimally affected by short outages. As a result, availability is a function of the frequency of outages (caused by unplanned failures or scheduled maintenance and upgrades) and the time to recover from such outages.

Many companies are addressing their availability requirements by implementing networks fabrics of Fibre Channel devices designed to provide high-performance storage environments. These flexible SANs are based on the following principles:

- A thorough understanding of availability requirements throughout the enterprise
- A flexible design that incorporates fault tolerance through redundancy and mirroring
- Simplified fault monitoring, diagnostics, and repair capabilities to ensure fast recovery
- A minimal amount of human intervention required during failover events
- A reliable backup and recovery plan to account for a wide variety of contingencies

25.2 Storage consolidation with remote disk

This remote disk consolidation solution is an extension of the most basic SAN solutions of disk consolidation (Figure 25-1).

Disk consolidation provides a means for storage administrators to take advantage of the capabilities of Fibre Channel fabric-based Storage Area Networks to organize storage in more flexible and efficient ways, easing many aspects of storage administration.

In an attempt to extend the distance between the storage and the servers, and in order to provide business benefits like consolidation, physical separation of disk storage from the server across sites, is necessary. This solution extends the distance to up to 10 km apart, using longwave SFPs in the SAN fabric, and single-mode fiber cable interconnect.

Note: CNT supports a distance of up to 80 km using extended longwave SFPs between the directors.

The example is comprised of:

- Three dual-port hosts
- Four CNT UltraNet Multi-service Directors
- One IBM TotalStorage Enterprise Storage Server Model 800 with eight FC host adapters

The host servers and ESS are connected to the SAN fabric via shortwave SFPs using multi-mode fibre cable.

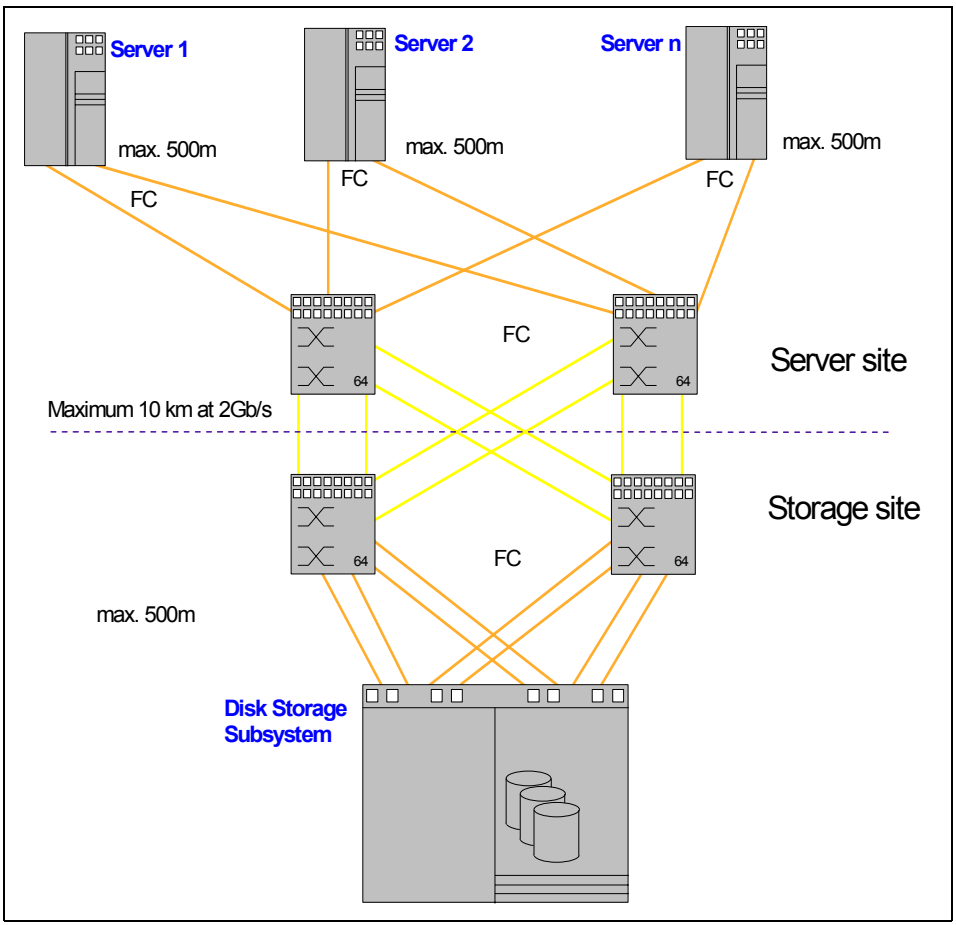


Figure 25-1 Remote disk consolidation

Checklist

We checked the following items:

- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and firmware level for proposed configuration
- CNT director longwave SFP count
- All cables clearly labelled and recorded
- The ISL links — single-mode fiber cable
- Single-mode FC cable laying and termination
- CNT Director high availability features
- Nickname assignments so we can quickly cross reference WWNs to devices
- ESS LUN definitions done
- CNT Enterprise Management System userids and passwords defined

Performance

In this simple implementation, performance will basically depend on the number of HBAs available on each server and the number of storage connections and most important the distance between storage and servers. It is important to consider application behavior over the distance, and this varies from application to application and we have assumed that the performance is adequate for applications running in the servers.

The CNT director supports any to any connectivity so it will not affect performance by itself. Latency in the director is less than 2.0 microseconds; this is not relevant compared to storage device response times typically in the millisecond range.

Knowing the requirements of our servers we can decide whether it is enough or if we need to add more connections or even more storage devices.

Without having the exact requirements we may consider that for a high profile server, a server to storage ratio of 6:1 is acceptable. This is only a starting point, we will then implement some measurement system or use statistical data to decide whether we need to add more connections or we have more bandwidth than required.

Scalability

The CNT director supports the concurrent addition of I/O modules and ports, so we can scale this solution by adding more servers or storage devices without disrupting operation. With four ISLs between the pair of directors on the host side and the pair of directors on the storage side, and four spare ports in a 64 port director, we can have 56 dual HBA servers (64 ports - 4 ISL - 4 spare = 56 server ports). With eight ESS connections, this slightly exceeds the 6:1 server to storage ratio (6 x 8 = 48). With additional I/O modules, the director can be scaled up to 256 ports.

Security

Here are some security considerations:

- ESS LUN masking by WWN will allow each server access only to configured LUNs.
- CNT Enterprise Management System userids, passwords and user rights are defined and defaults are removed, so only authorized personnel can perform management functions.
- Remote access to CNT Enterprise Management System configured to limit access to authorized workstations.
- Physical director security — locked cabinet, restricted access site.

“What if” failure scenarios

These are some theoretical assumptions:

- **Host HBA failure**

  SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. CNT Enterprise Management System access and ESS Specialist access are required.

- **Storage host adapter failure**

  The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed eight host adapters, so a single adapter failure will reduce available bandwidth by 12.5%. For an average workload and three servers as shown it will not impact performance. When the host adapter is replaced we do not need to update zoning as the ESS WWN is maintained by the ESS.

- **Director port failure**

The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port, as the target FC ID will have changed. Physical access to director and CNT Enterprise Management System user with maintenance rights is required. OS root access may be required.

- **Fiber failure**

  Impact will depend on whether it is a server attachment, storage attachment or ISL fiber. For ISL fiber, path routes may be updated within the fabric. The only action required is fiber replacement. Physical access to director and attached device are required.

- **CNT Enterprise Management System failure**

  No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example. Call home support will also be unavailable.

- **CNT Enterprise Management System hard drive failure**

  Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from the backup.

- **Director completely down, storage completely down, or site down (power, air conditioning, site damage**)

  This will cause an interruption in normal operation.

- **Physical damage to storage causing data loss (fire, flood)**

  We will need to restore data from backup tapes.

25.3 Two sites up to 10 km apart

In Figure 25-2, we show a basic dual fabric SAN distance extension solution between two sites up to 10 km apart. Each site has a fully redundant infrastructure at the SAN fabric level comprising of two CNT UltraNet Multi-service Directors. The host servers and storage are connected to the SAN fabric via shortwave SFPs using multi-mode fibre cable. The directors are interlinked across the sites via two ISLs each using longwave SFPs and single-mode fiber cables.

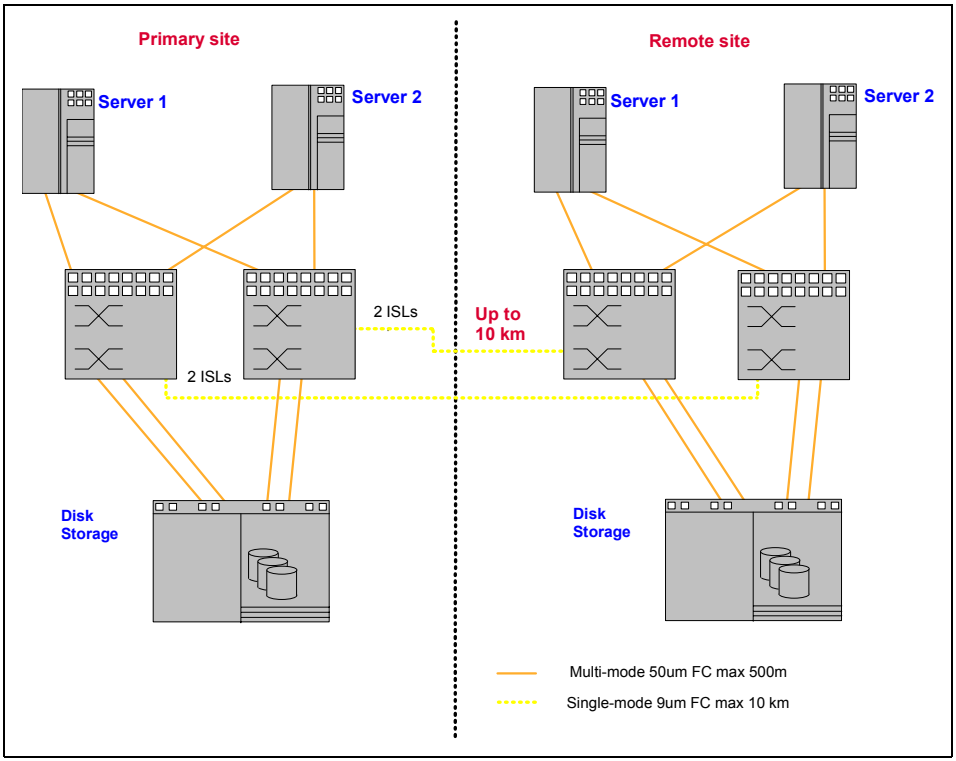


Figure 25-2 SAN distance extension up to 10 km

The solution can be used for accessing a database at the remote site as well as mirroring data at an alternative site up to 10 km.

Note: CNT supports a distance of up to 80 km using extended longwave SFPs between the directors.

One can assume that a greater distance between IT sites results in greater security against wide spread disaster. However, increased distance has its price in terms of interconnection cost, business relocation effort, and network cost. Before you start to expand your storage network, you have to consider if the following solution can meet your business purpose.

The two sites are each comprised of:

- Two dual-port hosts
- Two CNT UltraNet Multi-service Directors
- One IBM TotalStorage Enterprise Storage Server Model 800 with four FC host adapters

- Software
  - IBM Subsystem Device Driver (SDD)
  - IBM StorWatch Enterprise Storage Server Specialist
  - IBM TotalStorage Expert for ESS (for historical disk performance data)

We have assumed the following for our solution:

- Both sites have SAN infrastructure.
- Server to director and director to storage use multi-mode fibre.
- Servers in production which are being used for mission critical production need data mirroring for disaster recovery.
- Some servers at the primary site need to access database in storage at the remote site, in addition to using data in the primary storage.
- Servers at the remote site are in standby-mode, and being used for backup servers at ordinary times.
- Single-mode FC cables are used as ISLs between directors at each site.
- Software mirroring is used on servers.
- No PPRC proposed.

Checklist

The following items should be considered to implement the solution:

- Fibre Channel HBA supported on servers
- Operating systems support software mirroring
- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and firmware level for proposed configuration
- Director longwave SFP count
- The ISL links — single-mode fiber cable
- Single-mode FC cable laying and termination
- E_D_TOV and R_A_TOV settings equal on all directors
- ISL ports are configured with sufficient BB_Credits for 10-100 km links
- CNT Director high availability features
- Nickname assignments so we can quickly cross reference WWNs to devices
- ESS LUN definitions done
- CNT Enterprise Management System users and passwords defined

Performance

The major performance consideration with a long distance solution is calculating the correct number of lines between sites. This number can only be determined by performing a detailed performance profile of the servers and storage that will be remote. Over estimating the number of lines will increase costs dramatically, whilst under estimating the number of lines will dramatically impact the performance of the SAN.

It is vital that detailed performance data is available prior to sizing the number of lines required.

Typically, latency will increase over long distances – a good rule of thumb is 4.8 microseconds per kilometer. The default E_D_TOV and R_A_TOV values do not need to be modified for this distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the Directors across sites (the ISL links).

Scalability

The CNT director supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation. With two ISLs between each director on the primary site and the director on the remote site, two ports for the ESS and four spare ports in a 64 port director, we can have 54 dual HBA servers (64 ports - 4 ISL - 2 ESS ports - 4 spare = 54 server ports). This may be too many for four ESS connections, depending on the type of server and workload, in which case more director ports may need to be used for ESS connections. With additional I/O modules, the director can be scaled up to 256 ports.

Security

Here are some security considerations:

- ESS LUN masking by WWN will allow each server access only to configured LUNs.
- CNT Enterprise Management System user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions.
- Remote access to CNT Enterprise Management System configured to limit access to authorized workstations.
- Physical director security — locked cabinet, restricted access site.
- The zoning information is separate for the two fabrics, so when changing the zoning in one fabric it is not automatically propagated to the other fabric.

- Our solution assumes we own private lines between sites so encryption is not required.
- For leased lines or managed services where lines are shared, encryption is normally an option available from the service provider.

"What if" failure scenarios

Here are some theoretical assumptions:

- **Host HBA failure**

  SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. CNT Enterprise Management System access and ESS Specialist access are required.

- **Storage host adapter failure**

  The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce available bandwidth by 25%. For an average workload and two servers as shown it will not impact performance. When the host adapter is replaced we do not need to update zoning as the ESS WWN is maintained by the ESS.

- **Director port failure**

  The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port, as the target FC ID will have changed. Physical access to director and CNT Enterprise Management System user with maintenance rights is required. OS root access may be required.

- **Fiber failure**

  Impact will depend on whether it is a server attachment, storage attachment or ISL fiber. For ISL fiber, path routes may be updated within the fabric. The only action required is fiber replacement. Physical access to director and attached device may be required. For the ISL fibres between sites, the fibre provider will be responsible for repair.

- **CNT Enterprise Management System failure**

  No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example. Call home support will also be unavailable.

- **CNT Enterprise Management System hard drive failure**

  Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from the backup.

- **Director completely down, storage completely down, or site down (power, air conditioning, site damage)**

  This will cause an interruption in normal operation.

- **Physical damage to storage causing data loss (fire, flood)**

  We will need to restore data from backup tapes.

25.4 Two sites up to 100 km apart

In Figure 25-3 we show a basic dual fabric SAN distance extension solution between two sites up to 100 km apart. Each site has a fully redundant infrastructure at the SAN fabric level comprising of two CNT UltraNet Multi-service Directors. The host servers and storage are connected to the SAN fabric via shortwave SFPs using multi-mode fibre cable. The directors are interlinked across the sites via two ISLs each using longwave SFPs and single-mode fiber cables and repeaters. The number of repeaters used depends on the type of longwave SFP used.

One can assume that a greater distance between IT sites results in greater security against wide spread disaster. However, increased distance has its price in terms of interconnection cost, business relocation effort, and network cost. Before you start to expand your storage network, you have to consider if the following solution can meet your business purpose.

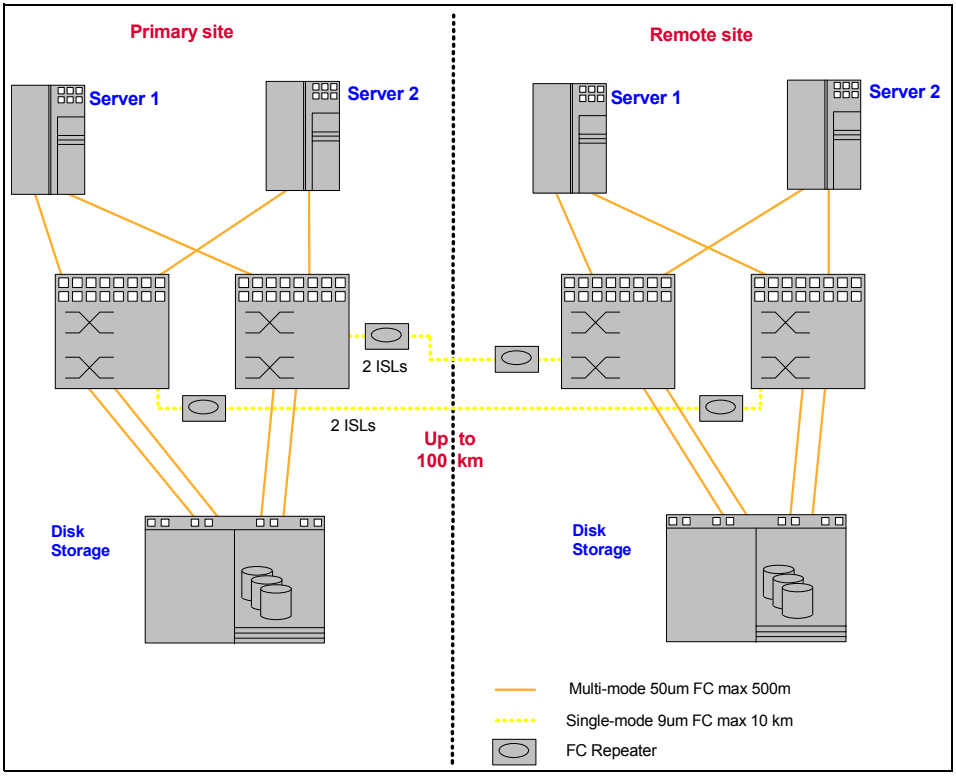


Figure 25-3 SAN distance extension up to 100 km with repeaters

The two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 945, but with the addition of:

- FC repeaters as required to reach 100 km

Note: CNT supports a distance of up to 80 km using extended longwave SFPs between the directors, hence repeaters are only necessary when exceeding 80 km.

Checklist

The same items as for “Two sites up to 10 km apart” on page 945 should be considered, with the addition of:

- Repeater count verified for the distance in every ISL link
- Repeaters installed connected and tested

Performance

The same performance criteria apply as for “Two sites up to 10 km apart” on page 945, other than overall ISL latency increases with distance and BB_Credit utilization will increase.

Scalability

The same scalability criteria apply as for “Two sites up to 10 km apart” on page 945.

Security

The same items as for “Two sites up to 10 km apart” on page 945 should be considered, with the addition of:

- Physical repeater security — locked cabinet, restricted access site.

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 945, with the addition of:

- **Repeater failure**

  Impact will depend on whether alternate ISL links are available or not. Only action required is repeater replacement. Physical access to repeater and repeater site required.

25.5 Two sites - Point-to-point DWDM solution

In Figure 25-4 we show a basic high availability SAN design for multiple servers deployed with dual directors and a redundant fabric at each site. This design uses Dense Wavelength Division Multiplexing (DWDM) equipment for inter-site ISLs between the directors.

Note: Refer to Chapter 4, “Extending fiber over distance with DWDM” on page 183 for further information on DWDM technology and topologies.

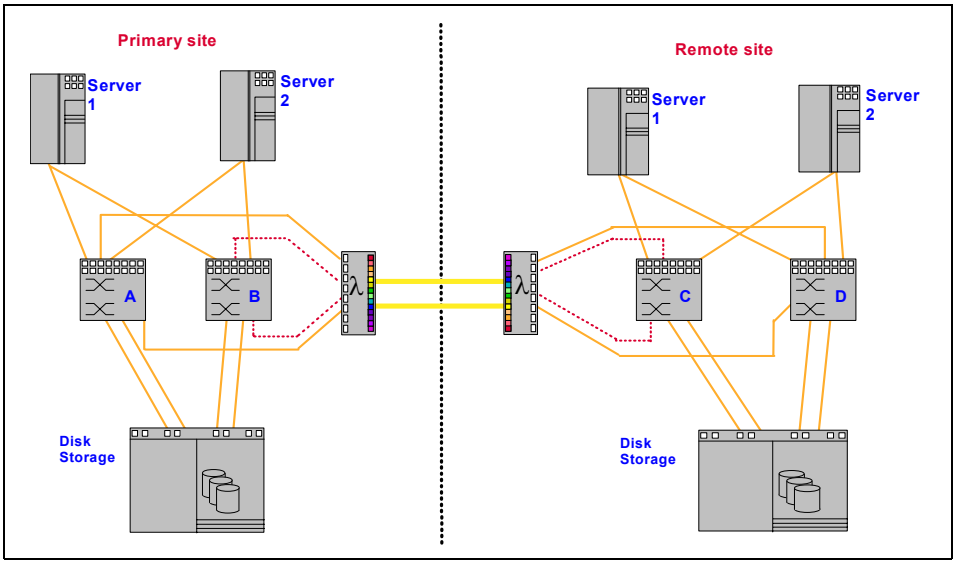


Figure 25-4 Point-to-point DWDM solution - two sites

The use of DWDM allows more channels to be sent over the same number of fiber cables. This is achieved by using different wavelengths of light to carry the different channels at the same time, and also supports transmission of multiple protocols (e.g. FC, ESCON, TCP/IP) over the same fibre. To ensure redundancy, and hence availability, the data should be sent over two completely different physical paths, but with equal length to ensure equal latency.

The method for replicating the data could be either host software based, or via storage hardware based facilities such as PPRC. In this example we are using host based software replication.

In our example, SAN AD (Directors A and D) will be a separate fabric from SAN BC (Directors B and C) fabric. Each fabric comprises of two directors, one at each site.

The two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 945, but with the addition of:

- DWDM equipment linked via dark fibre

Checklist

The same items as for “Two sites up to 10 km apart” on page 945 should be considered, with the addition of:

- The ISL link count to the DWDM equipment at each site

- Maximum distance for individual devices
- Configuration and count of the client interface adapters in the DWDM equipment
- Configuration and count of the transport network adapters in the DWDM, ensuring same frequency band and channel
- Connectivity testing between:
  - ESS to switches
  - Servers to switch
  - Switches to DWDM
- Validate connectivity
- Test server to storage on both paths
- Tests of failover/fail back operations
- Test inter-site connectivity via DWDM
- Setup LUNs on back up ESS and test from primary server, to quantify latency
- Pre-installation testing of the DWDM transport network for configuration and performance

Performance

The same performance criteria apply as for “Two sites up to 10 km apart” on page 945, with the addition of:

- The overall ISL latency increases with distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the director and the DWDM equipment. If the number of ISLs between the directors and DWDM equipment increases, it may be necessary to increase the quantity of dark fibre between the DWDMs that carry the ISL traffic.

Scalability

Most DWDMs support multiple I/O interfaces for client equipment and support up to 32 separate channels. DWDM equipment also allows mixing different interface protocols on a single physical DWDM channel.

The same scalability criteria apply as for “Two sites up to 10 km apart” on page 945.

Security

The same items as for “Two sites up to 10 km apart” on page 945 should be considered, with the addition of:

If you decide to add an ISL, ensure that all checks are done on the zone configuration changes from the management console.

- The DWDM equipment typically provides Web based management software that can be accessed by any workstation connected to the same LAN. Different user levels are provided for administrators, operators or observers. Different passwords must be set to limit access.
- The physical security of fiber connections and patch panels should be considered.

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 945, with the addition of:

- **DWDM optical channel card failure**

  As we configured the channels for high availability, there are redundant cards in the DWDM. If we lose one, the traffic will automatically be switched to the other and the channel will remain available. The failed card can be replaced concurrently.

- **DWDM optical channel manager card failure**

  The optical channel manager card performs path high availability switching. There are two cards in each shelf, if one fails the other takes control and the operation is not affected. The failed card can be replaced concurrently.

- **DWDM Optical Multiplexer failure or shelf backplane failure**

  The entire shelf will be unavailable. As we spread connections over different shelves, we will have at least half the channels of each type available. Operation will continue although performance may be affected.

- **Dark fiber failure**

  As we configured for high availability, operation will continue using the available pair with no performance impact. Customers may have additional considerations based on their business recovery policy to source the redundant dark fiber from another telco/source.

DWDM considerations

We have shown a point-to-point topology here. We would implement the DWDM solution as a protected fiber to ensure availability. In this example we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation, here we have assumed a light workload.

We show here an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one director in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits, however this is not a DWDM consideration but more to do with the general SAN solution design over a distance. The DWDM is a core architecture deployment and because of it's independence from protocol it can be used for other traffic beyond the SAN environment.

In order to maintain performance at extended distances, we need to ensure a sufficient number of buffers on each interconnecting port to compensate for the number of frames that are in transit.

25.6 Two sites: Point-to-point DWDM with ESS PPRC

With the IBM TotalStorage Enterprise Storage Server Model 800 running LIC 2.3.0 or later and the PPRC Version 2 licensed feature, it is no longer necessary to use dedicated ESCON ports on the ESS to connect via DWDM to a remote ESS. Remote copy traffic can now share FCP paths with normal SAN traffic, and hence the remote ESS can be accessed via the same connections as shown in Figure 25-4 on page 953.

Checklist

The same items as for "Two sites - Point-to-point DWDM solution" on page 952 should be considered, with the addition of:

- PPRC Version 2 license for the ESS
- Test PPRC - Establish paths and Test PPRC functionality - qualify timings

Performance

This is the same as for "Two sites - Point-to-point DWDM solution" on page 952.

Scalability

This is the same as for “Two sites - Point-to-point DWDM solution” on page 952.

Security

This is the same as for “Two sites - Point-to-point DWDM solution” on page 952.

“What if” failure scenarios

This is the same as for “Two sites - Point-to-point DWDM solution” on page 952.

DWDM considerations

In addition to those listed in “Two sites - Point-to-point DWDM solution” on page 952, we need to consider that the solution shown here also implements IBM’s PPRC between the two ESS subsystems. See the *IBM TotalStorage Enterprise Storage Server Implementing ESS Copy Services in Open Environments*, SG24-5757 for details on implementing PPRC with ESS.

25.7 Multiple site ring DWDM solution

Figure 25-5 illustrates four dual-fabric sites connected in a hubbed ring topology via DWDM and OADM technology. Site 1 is the hub node where all the DWDM channels originate and terminate, whilst the other three sites utilize OADM technology to add or remove channels from the ring.

Note: Refer to Chapter 4, “Extending fiber over distance with DWDM” on page 183 for further information on DWDM technology and topologies.

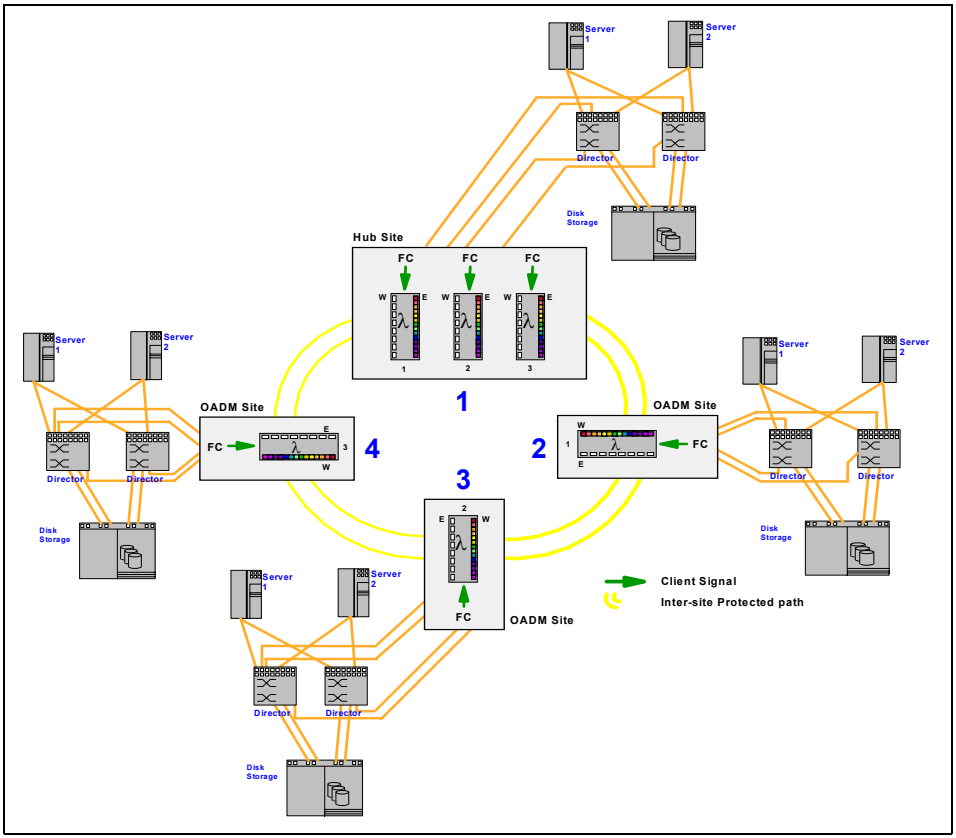


Figure 25-5 Multiple site: Ring topology DWDM solution

Each path has its own wavelength. The OADM pass through feature (express channels) enables non-contiguous sites to connect over an intermediate site as if they were directly connected. The only additional overhead of the pass through is the minimal latency (5 microseconds/km) of the second link. The pass through has no overhead since it is a passive device. The two fabrics would logically appear as fully meshed topologies.

Each of the links can operate in protected mode, which provides a redundant path in the event of a link failure. In most cases, link failures are automatically detected within 50 msec. In this case, the two wavelengths of the failed link reverse directions and reach the target port at the opposite side of the ring. If the link between 1 and 4 fails, the transmitted wavelength from 4 to 1 would reverse direction and reach 1 through 3 and 2. The transmitted wavelength from 1 to 4 would also reverse direction and reach 4 through 2 and 3.

Calculating the distance between nodes in a ring depends on the implementation of the protected path scheme. For instance, if the link between DWDM 2 and 3 fails, the path from 1 to 3 would be 1 to 2, back from 2 to 1 (due to the failed link), 1 to 4, and finally 4 to 3. This illustrates the need to utilize the entire ring circumference (and more, in a configuration with over four nodes) for failover.

Another way to calculate distance between nodes is to set up the protected path in advance (in the reverse direction) so the distance is limited to the number of hops between the two nodes. In either case, the maximum distance between nodes determines the maximum optical reach. An example of this specification is 80 to 100 km for a maximum distance between nodes and 160 to 400 km for maximum ring size. These distances should continue to increase as fiber optic technology advances.

Each site has a set of components the same as one site in the “Two sites - Point-to-point DWDM solution” on page 952, except:

- Site 1 has three sets of DWDM channels
- Sites 2, 3 and 4 each have OADM equipment
- Dark fibre between each site

Checklist

The same items as for “Two sites - Point-to-point DWDM solution” on page 952 should be considered, with the addition of:

- Installation and configuration of the OADM equipment
- Ensuring correct frequency band and channel for each site

Performance

This is the same as for “Two sites - Point-to-point DWDM solution” on page 952.

Scalability

This is the same as for “Two sites - Point-to-point DWDM solution” on page 952.

Security

This is the same as for “Two sites - Point-to-point DWDM solution” on page 952.

“What if” failure scenarios

This is the same as for “Two sites - Point-to-point DWDM solution” on page 952, except that the scenarios for DWDM failures are also applicable to OADMs.

DWDM considerations

We have shown a ring topology here, this gives a logical mesh, potentially giving any to any connectivity. We have implemented this as two rings, one ring connects to one switch in each site, the other ring connects to the remaining switch, this give two discrete SAN fabrics.

We would implement the DWDM solution as a protected ring to ensure availability. In this example we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation, here we have assumed a light workload.

We show here an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one director in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits, however this is not a DWDM consideration but more to do with the general SAN solution design over a distance. The DWDM is a core architecture deployment and because of it's independence from protocol it can be used for other traffic beyond the SAN environment.

In order to maintain performance at extended distances, we need to ensure a sufficient number of buffers on each interconnecting port to compensate for the number of frames that are in transit.

25.8 Two sites: Channel extender and WAN extension

Channel extenders typically use telecommunication lines for data transfer and therefore enable application and recovery sites to be located over greater distances. The use of channel extenders provides the separation for disaster recovery purposes and avoids some of the barriers imposed when customers do not have a "right of way" to lay fiber cable.

In Figure 25-6 we show a typical SAN distance extension using Optical Channel extenders at both primary and secondary sites and the ATM network.

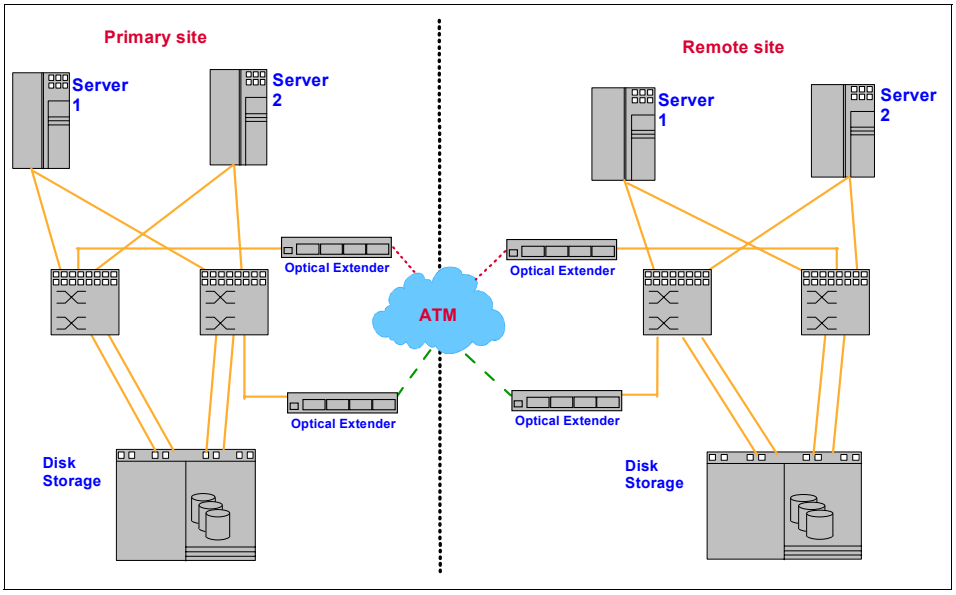


Figure 25-6 SAN extension over (ATM) WAN using channel extenders

Again, the two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 945, but with the addition of:

- Two optical channel extenders
- ATM already available

Checklist

The same items as for “Two sites up to 10 km apart” on page 945 should be considered, with the addition of:

- I/O interface (for example, FC or ESCON) selection and installation on the Optical channel extender
- Network interface (for example, ATM OC3) selection and installation on the Optical channel extender
- Pre-installation testing of the network interfaces (ATM network)

Performance

The same performance criteria apply as for “Two sites up to 10 km apart” on page 945, other than overall ISL latency increases with distance and BB_Credit utilization will increase.

Channel extenders generally compress the data before sending it over the transport network. However the compression ratio needs to be determined based on the individual application characteristics and the distance.

You must especially consider an amount of the updated data for a period of time (peak time), and reflect this in calculating the number of interconnecting links and the number of data volumes for SAN/WAN solution.

Scalability

The same scalability criteria apply as for “Two sites up to 10 km apart” on page 945. Most optical channel extenders support multiple I/O interfaces for client equipment. However in an installation, it is recommended to have a single protocol per interface on a physical channel extender.

Security

The same items as for “Two sites up to 10 km apart” on page 945 should be considered, with the addition of:

- Channel extender security — locked cabinet, restricted access site

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 945, with the addition of:

- **Channel extender failure**

  In the current solution channel extenders are not configured in a redundant mode, however they can be.

- **Transport network failure**

  The telco link failure is not planned for in the current solution, however a redundant link can be acquired from the vendor or a separate telco may be considered depending on the business recovery policy.

25.9 Two sites with FCIP connection

Note to Reviewer: Insufficient information available on this product to guarantee accuracy.

With the addition of the UltraNet Edge Storage Router (see “CNT Ultranet Edge Storage Router” on page 347) it is possible to connect two remote SANs together. This enables servers or devices in the primary site to have access to SAN devices in the remote site over TCP/IP networks. You can then make visible just those devices which are required from the remote site.

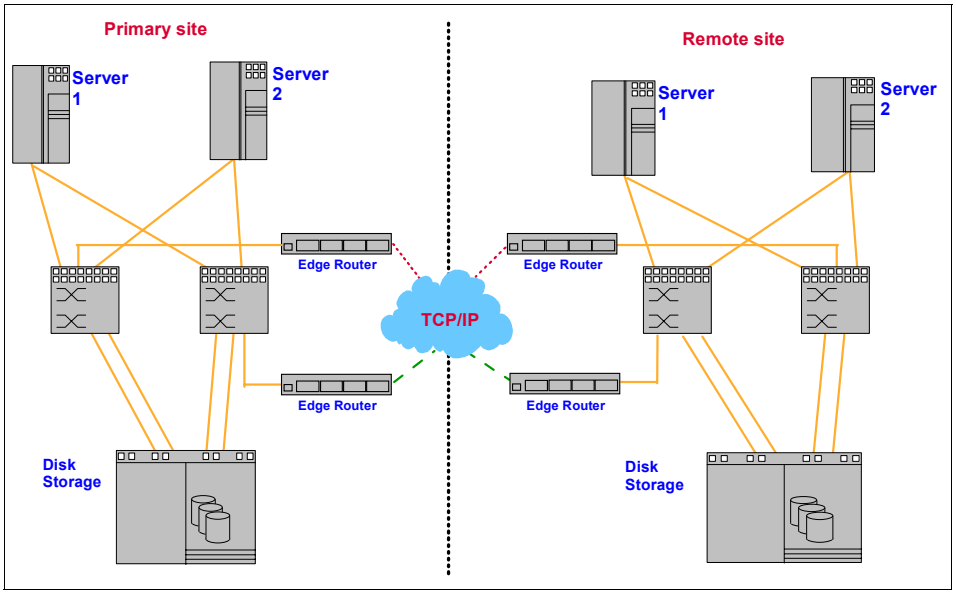


Figure 25-7 Remote access via FCIP router over TCP/IP

Again, the two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 945, but with the addition of:

- UltraNet Edge Storage Router for each director
- TCP/IP network already available

Checklist

The same items as for “Two sites up to 10 km apart” on page 945 should be considered, with the addition of:

- FCIP configuration
- Pre-installation testing of the network interfaces (TCP/IP network)

Performance

Extra latency will apply to those devices accessed via the TCP/IP network. For these devices, similar performance criteria apply as for “Two sites up to 10 km apart” on page 945, with the addition that overall latency increases with distance.

The UltraNet Edge Storage Router includes compression to reduce WAN bandwidth requirements.

Scalability

As well as the normal SAN scalability issues, the WAN bandwidth and number of FCIP switch ports must be managed to support the required amount of remote traffic.

Security

The same items as for “Two sites up to 10 km apart” on page 945 should be considered, with the addition of:

- Router security — locked cabinet, restricted access site

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 945, with the addition of:

- **UltraNet Edge Storage Router failure**

  In the current solution, a router would cause loss of access to one of the remote fabrics. Since a dual fabric design has been used at both sites, access could still continue via the alternate fabric. It is possible to use more than one router to connect two fabrics, such that traffic can continue to flow through the second route in situations where the first route has failed.

- **TCP/IP network failure**

  WAN failure could cause complete loss of access to remote devices. Such situations are beyond the scope of this book.

25.10 Remote tape vaulting

For an existing large corporation with multiple sites, tape library resources can be consolidated to a separate site. This simplifies data movement logistics and centralizes backup software configurations. By doing this the cost of doing business is reduced as the infrastructure is efficiently utilized and fewer IT personnel required for backup data management. In addition, in the event of a disaster, the data is already located on tape in the remote location and there is no longer a need to ship the data to another site. Another application is for outsourcing services companies like Storage service providers (SSP), that are interested in providing backup and Backup management services to smaller companies.

The Figure 25-8, shows the basic layout for a remote tape vaulting solution, using CNT directors and IBM TotalStorage SAN switches (required for tape FC-AL support).

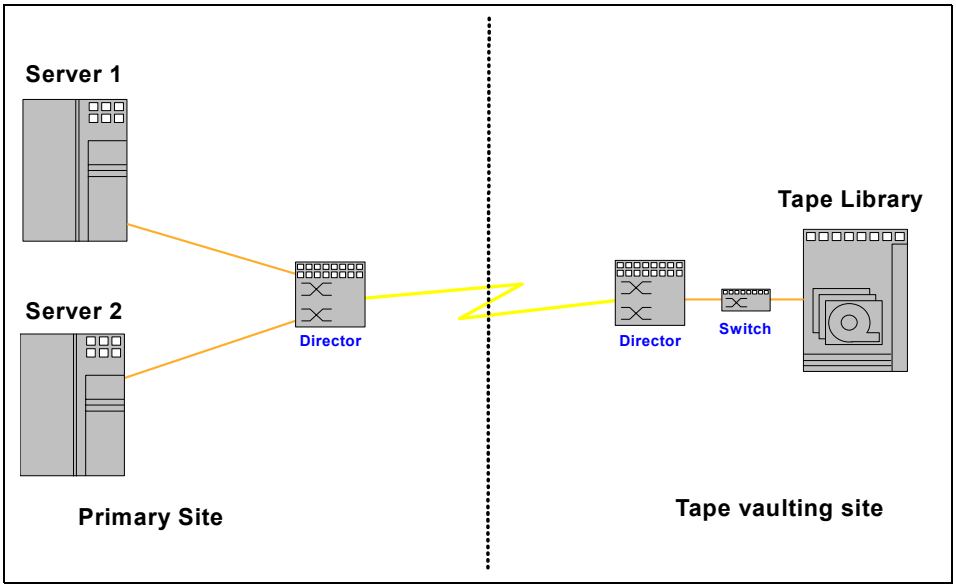


Figure 25-8 Remote tape vaulting

Any of the preceding distance solutions (dark fibre, repeaters, DWDM, iFCP, etc.) can be used to connect the local and remote directors, depending on circumstances and the distance required.

Since most tape libraries fiber channel ports only support FC-AL, and CNT directors do not directly support FC-AL, the drives need to be connected via an IBM TotalStorage switch. Some of the latest tape drives, such as the IBM TotalStorage Enterprise Tape Drive 3592, support direct switched fabric attachment, and so could be attached directly to the director. However, due to the relevant costs of director ports versus switch ports, it may still be advisable to use the IBM TotalStorage switch as a tape concentrator.

If a director class product is not required at the remote site, then the connection can be made directly to the switch.

Checklist

In addition to the items considered in the disk consolidation example, we must now consider:

- Host HBA supported for tape drive / library attachment

- Use of separate HBAs for disk and tape access
- Host tape device driver levels
- Host operating system levels compatible with tape library fiber requirements
- IBM TotalStorage switch configuration compatible with the rest of the fabric
- Zone configuration allowing tape access to required servers
- Host software tape sharing capabilities

Performance

The performance of the tape drives determines the number of drives which can be connected to the IBM TotalStorage switch, and hence the number of ISLs needed at the remote site to support the bandwidth. Installation of the Open Trunking feature on the directors and switch would also help to optimize ISL usage.

Certain operating system and device driver combinations support path failover for tape (e.g. AIX with Magstar 3590), and as such dual port tape drives can be connected to two IBM TotalStorage switches for redundancy purposes.

Scalability

The solution can be scaled by increasing the number of tape drives in the library, and the number of associated IBM TotalStorage switches. This may also require an increase in capacity of the offsite link.

Security

Here are some security considerations:

- Zoning can be used to restrict server access to specific tape devices as required.
- Proper tape management procedures will avoid servers contending for the same tape device.

“What if” failure scenarios

Here are some theoretical assumptions:

- Performance may be impacted depending on number of drives attached. Traditionally tape failover is a manual operation. Multiple path devices are configured as several logical devices, one per path. Only one of these logical devices is made active. If there is a failure the application aborts and it can then be restarted using a different logical device. For tape drives with dual paths, the latest levels of a tape device driver provide alternate pathing support and tape failover for Fibre Channel connections. With this support

enabled, if an error occurs the device driver will automatically initiate error recovery and the operation will continue using the next logical path.

- Device link or device port failure — only one tape drive affected. Recovery may be manual or automatic depending on operating system and driver level as explained for ISL or switch failure.
- Switch port failures — SFPs are hot swappable. Drives can be moved to a spare port.
- Switch failure — most FRUs are hot swappable. With multi-port drives, access can continue via an alternate switch.
- User error trying to access too many drives on the same link — performance of all drives attached via the link may be degraded. Director or switch performance view may be used to find what paths are carrying traffic.
- Tape drive failure in a single tape zone — an alternate zone should be made active to get access to a working device.

25.11 Remote tape vaulting with disaster tolerance

In Figure 25-9 we show an extension or variant of the tape vaulting solution. In this solution the primary site / local site also has a tape library. Data is written to both tape libraries, however in the event of failure on the tape library at the local site, data can be backed up and restored from the tape library at the remote site.

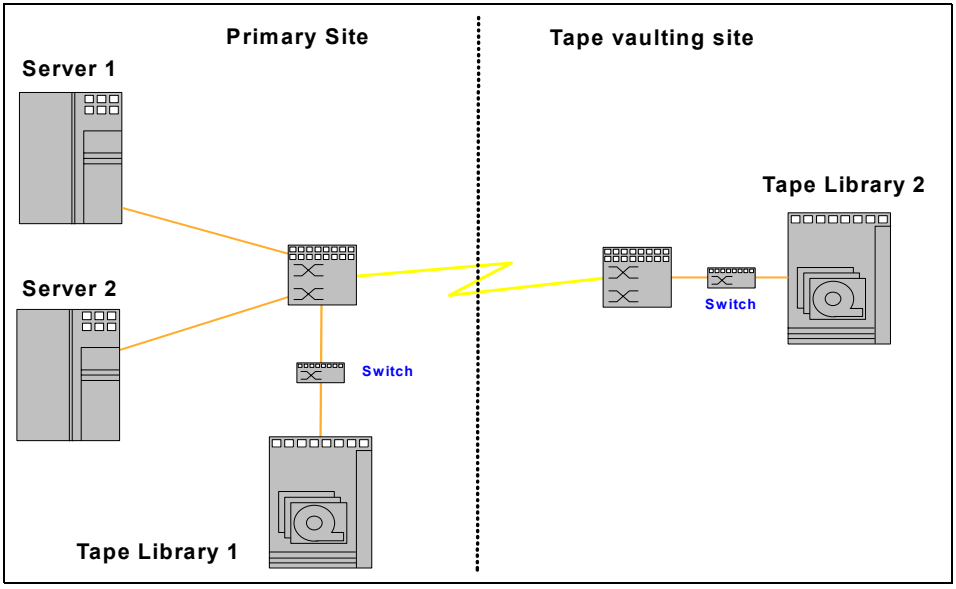


Figure 25-9 Remote tape vaulting with disaster tolerance

This solution has the same performance, scalability, security and failure attributes as the standard remote tape vaulting solution.

26

Cisco distance solutions

In this chapter we will discuss some solutions that can be implemented using the Cisco products included in the IBM portfolio.

For product details of the Cisco portfolio products refer to:

- Chapter 11, “Cisco switches and directors” on page 489
- *Implementing an Open IBM SAN,* SG24-6116

The solutions we will discuss are:

- Remote disk consolidation
- Two sites up to 10 km apart
- Two sites up to 100 km apart with repeaters
- Point-to-Point DWDM
- Point-to-Point DWDM with PPRC
- Ring topology DWDM
- Channel extender and WAN extension
- FCIP connection
- Remote tape vaulting
- Remote tape vaulting with redundancy

This chapter presents various implementations and uses for Storage Area Network (SAN) in high availability graphically dispersed environments. The considerations for a high availability SAN design using IBM SAN Fibre Channel switch are covered in the following topics.

© Copyright IBM Corp. 2004. All rights reserved.

By reducing or eliminating single points of failure in the enterprise environment, SANs can help to improve overall availability of business applications. By utilizing highly available components and solutions as well as a fault-tolerant design, enterprises can achieve the availability needed to support 24x7 uptime requirements. High distances in these solutions enable the SAN to cover a whole new host of IT availability issues, for example, we can deploy disaster tolerant solutions.

In network systems such as SANs, with their associated servers, fabric, and storage components, as well as software applications, downtime can occur even if parts of the system are highly available or fault tolerant. To improve business continuance under a variety of circumstances, SANs can incorporate redundant components, connections, software, and configurations to minimize or eliminate single points of failure.

With the emergence of the Internet and the proliferation of global e-business applications, more and more companies are implementing computing infrastructures specifically designed for continuous data and system availability. Today, even applications such as company e-mail have become mission critical for ongoing business operations. Faced with increased customer and internal user expectations, companies are currently striving to achieve at least 99.999 percent (“five nines”) availability in their computing systems — a figure equivalent to less than 5.3 minutes of downtime a year. Additional downtime can severely impact business operations and cost valuable time, money, and resources.

To ensure the highest level of system uptime, companies are implementing reliable storage networks capable of boosting the availability of data for all the users and applications that need it. These companies typically represent the industries that demand the highest levels of system and data availability — the utilities and telecommunications sector, brokerages and financial service institutions, and a wide variety of service providers.

26.1 High-level availability objectives

Developing highly available networks involves identifying specific availability requirements and predicting what potential failures might cause outages. The first step is to clearly define availability objectives, which can vary widely from company to company and even within segments of the same company. In some environments, no disruption can be tolerated, while other environments might be only minimally affected by short outages. As a result, availability is a function of the frequency of outages (caused by unplanned failures or scheduled maintenance and upgrades) and the time to recover from such outages.

Many companies are addressing their availability requirements by implementing networks fabrics of Fibre Channel devices designed to provide high-performance storage environments. These flexible SANs are based on the following principles:

- A thorough understanding of availability requirements throughout the enterprise
- A flexible design that incorporates fault tolerance through redundancy and mirroring
- Simplified fault monitoring, diagnostics, and repair capabilities to ensure fast recovery
- A minimal amount of human intervention required during failover events
- A reliable backup and recovery plan to account for a wide variety of contingencies

26.2 Storage consolidation with remote disk

This remote disk consolidation solution is an extension of the most basic SAN solutions of disk consolidation (Figure 26-1).

Disk consolidation provides a means for storage administrators to take advantage of the capabilities of Fibre Channel fabric-based Storage Area Networks to organize storage in more flexible and efficient ways, easing many aspects of storage administration.

In an attempt to extend the distance between the storage and the servers, and in order to provide business benefits like consolidation, physical separation of disk storage from the server across sites, is necessary. This solution extends the distance to up to 10 km apart, using longwave SFPs in the SAN fabric, and single-mode fiber cable interconnect.

The example is comprised of:

- Three dual-port hosts
- Four Cisco MDS 9506 Multilayer Directors
- One IBM TotalStorage Enterprise Storage Server Model 800 with eight FC host adapters

The host servers and ESS are connected to the SAN fabric via shortwave SFPs using multi-mode fibre cable.

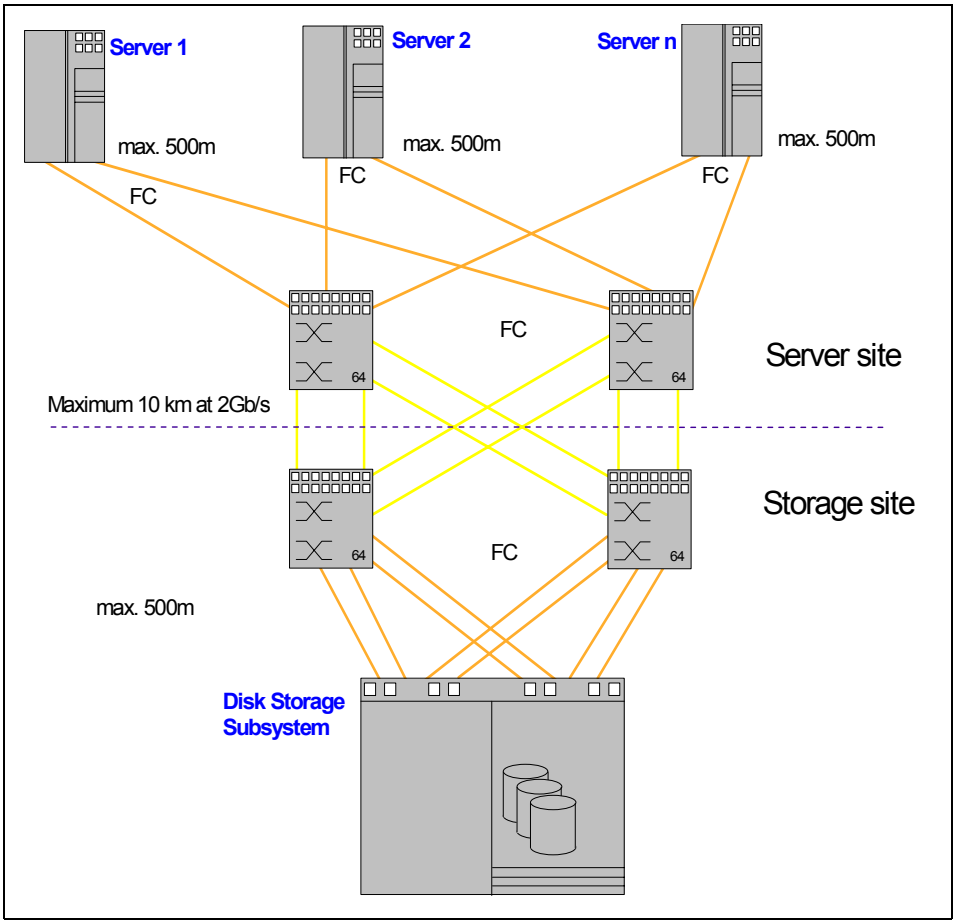


Figure 26-1 Remote disk consolidation

Checklist

We checked the following items:

- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and firmware level for proposed configuration
- Cisco switches and director longwave SFP count
- All cables clearly labelled and recorded
- The ISL links — single-mode fiber cable
- Single-mode FC cable laying and termination

- Cisco Director high availability features
- Nickname assignments so we can quickly cross reference WWNs to devices
- ESS LUN definitions done
- Fabric Manager userids and passwords defined

Performance

In this simple implementation, performance will basically depend on the number of HBAs available on each server and the number of storage connections and most important the distance between storage and servers. It is important to consider application behavior over the distance, and this varies from application to application and we have assumed that the performance is adequate for applications running in the servers.

The Cisco director supports any to any connectivity so it will not affect performance by itself. Latency in the director is less than 2.0 microseconds; this is not relevant compared to storage device response times typically in the millisecond range.

Knowing the requirements of our servers we can decide whether it is enough or if we need to add more connections or even more storage devices.

Without having the exact requirements we may consider that for a high profile server, a server to storage ratio of 6:1 is acceptable. This is only a starting point, we will then implement some measurement system or use statistical data to decide whether we need to add more connections or we have more bandwidth than required.

Scalability

The Cisco director supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation. With four ISLs between the pair of directors on the host side and the pair of directors on the storage side, and four spare ports in a fully populated director, we can have 56 dual HBA servers (64 ports - 4 ISL - 4 spare = 56 server ports). With eight ESS connections, this slightly exceeds the 6:1 server to storage ratio (6 x 8 = 48).

Security

Here are some security considerations:

- ESS LUN masking by WWN will allow each server access only to configured LUNs.

- Fabric Manager userids, passwords and user rights are defined and defaults are removed, so only authorized personnel can perform management functions.
- Remote access to Fabric Manager configured to limit access to authorized workstations.
- Physical director security — locked cabinet, restricted access site.

"What if" failure scenarios

These are some theoretical assumptions:

- **Host HBA failure**

  SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. Fabric Manager access and ESS Specialist access are required.

- **Storage host adapter failure**

  The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed eight host adapters, so a single adapter failure will reduce available bandwidth by 12.5%. For an average workload and three servers as shown it will not impact performance. When the host adapter is replaced we do not need to update zoning as the ESS WWN is maintained by the ESS.

- **Director port failure**

  The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port, as the target FC ID will have changed. Physical access to director and Fabric Manager user with maintenance rights is required. OS root access may be required.

- **Fiber failure**

  Impact will depend on whether it is a server attachment, storage attachment or ISL fiber. For ISL fiber, path routes may be updated within the fabric. The only action required is fiber replacement. Physical access to director and attached device are required.

- **Fabric Manager Server failure**

  No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example. Call home support will also be unavailable.

- **Fabric Manager Server hard drive failure**

Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from CD-RW drive backup.

- **Director completely down, storage completely down, or site down (power, air conditioning, site damage**)

  This will cause an interruption in normal operation.

- **Physical damage to storage causing data loss (fire, flood)**

  We will need to restore data from backup tapes.

26.3 Two sites up to 10 km apart

In Figure 26-2, we show a basic dual fabric SAN distance extension solution between two sites up to 10 km apart. Each site has a fully redundant infrastructure at the SAN fabric level comprising of two Cisco MDS 9506 Multilayer Directors. The host servers and storage are connected to the SAN fabric via shortwave SFPs using multi-mode fibre cable. The directors are interlinked across the sites via two ISLs each using longwave SFPs and single-mode fiber cables.

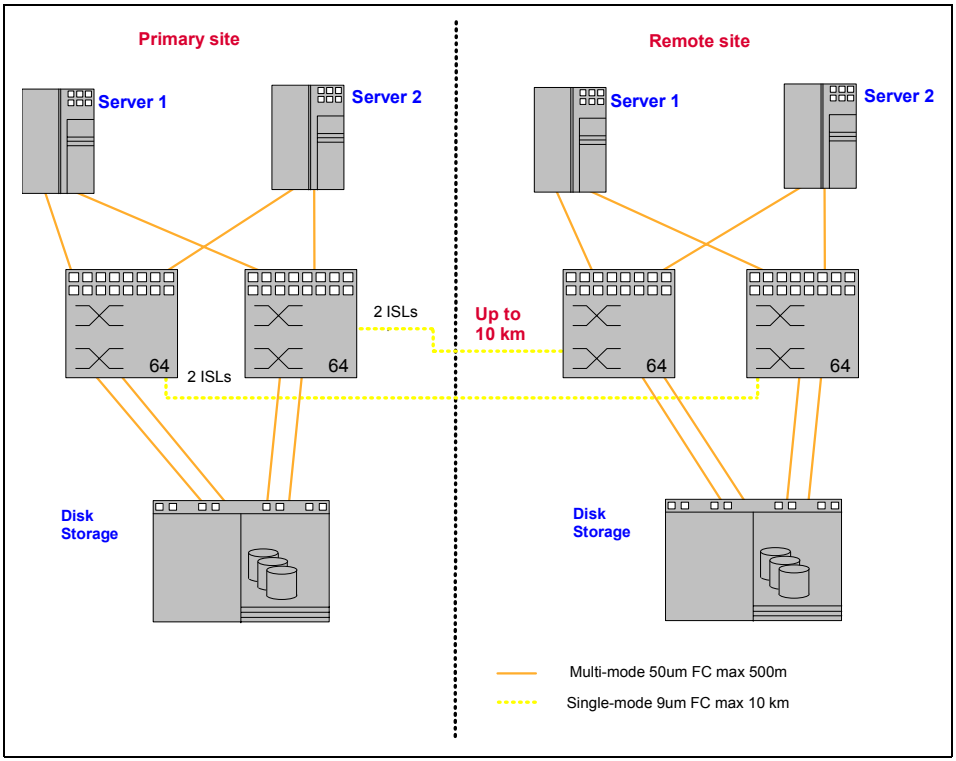


Figure 26-2 SAN distance extension up to 10 km

The solution can be used for accessing a database at the remote site as well as mirroring data at an alternative site up to 10 km.

Note: Cisco supports a distance of up to 35 km using extended longwave SFPs between the directors.

One can assume that a greater distance between IT sites results in greater security against wide spread disaster. However, increased distance has its price in terms of interconnection cost, business relocation effort, and network cost. Before you start to expand your storage network, you have to consider if the following solution can meet your business purpose.

The two sites are each comprised of:

- Two dual-port hosts
- Two Cisco MDS 9506 Multilayer Directors
- One IBM TotalStorage Enterprise Storage Server Model 800 with four FC host adapters

- Software
  - IBM Subsystem Device Driver (SDD)
  - IBM StorWatch Enterprise Storage Server Specialist
  - IBM TotalStorage Expert for ESS (for historical disk performance data)

We have assumed the following for our solution:

- Both sites have SAN infrastructure.
- Server to director and director to storage use multi-mode fibre.
- Servers in production which are being used for mission critical production need data mirroring for disaster recovery.
- Some servers at the primary site need to access database in storage at the remote site, in addition to using data in the primary storage.
- Servers at the remote site are in standby-mode, and being used for backup servers at ordinary times.
- Single-mode FC cables are used as ISLs between directors at each site.
- Software mirroring is used on servers.
- No PPRC proposed.

Checklist

The following items should be considered to implement the solution:

- Fibre Channel HBA supported on servers
- Operating systems support software mirroring
- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and firmware level for proposed configuration
- Director longwave SFP count
- The ISL links — single-mode fiber cable
- Single-mode FC cable laying and termination
- E_D_TOV and R_A_TOV settings equal on all directors
- Extended distance buffering enabled for ISL ports to increase BB_Credit for 10-100 km links
- Cisco Director high availability features
- Nickname assignments so we can quickly cross reference WWNs to devices
- ESS LUN definitions done

- Fabric Manager users and passwords defined

Performance

The major performance consideration with a long distance solution is calculating the correct number of lines between sites. This number can only be determined by performing a detailed performance profile of the servers and storage that will be remote. Over estimating the number of lines will increase costs dramatically, whilst under estimating the number of lines will dramatically impact the performance of the SAN.

It is vital that detailed performance data is available prior to sizing the number of lines required.

Typically, latency will increase over long distances – a good rule of thumb is 4.8 microseconds per kilometer. The default E_D_TOV and R_A_TOV values do not need to be modified for this distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the Directors across sites (the ISL links).

Scalability

The Cisco director supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation. With two ISLs between each director on the primary site and the director on the remote site, two ports for the ESS and four spare ports in a fully populated director, we can have 54 dual HBA servers (64 ports - 4 ISL - 2 ESS ports - 4 spare = 54 server ports). This may be too many for four ESS connections, depending on the type of server and workload, in which case more director ports may need to be used for ESS connections.

Security

Here are some security considerations:

- ESS LUN masking by WWN will allow each server access only to configured LUNs.
- Fabric Manager user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions.
- Remote access to Fabric Manager configured to limit access to authorized workstations.
- Physical director security — locked cabinet, restricted access site.
- The zoning information is separate for the two fabrics, so when changing the zoning in one fabric it is not automatically propagated to the other fabric.

- Our solution assumes we own private lines between sites so encryption is not required.
- For leased lines or managed services where lines are shared, encryption is normally an option available from the service provider.

"What if" failure scenarios

Here are some theoretical assumptions:

- **Host HBA failure**

  SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. Fabric Manager access and ESS Specialist access are required.

- **Storage host adapter failure**

  The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce available bandwidth by 25%. For an average workload and two servers as shown it will not impact performance. When the host adapter is replaced we do not need to update zoning as the ESS WWN is maintained by the ESS.

- **Director port failure**

  The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port, as the target FC ID will have changed. Physical access to director and Fabric Manager user with maintenance rights is required. OS root access may be required.

- **Fiber failure**

  Impact will depend on whether it is a server attachment, storage attachment or ISL fiber. For ISL fiber, path routes may be updated within the fabric. The only action required is fiber replacement. Physical access to director and attached device may be required. For the ISL fibres between sites, the fibre provider will be responsible for repair.

- **Fabric Manager Server failure**

  No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example. Call home support will also be unavailable.

- **Fabric Manager Server hard drive failure**

Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from CD-RW drive backup.

- **Director completely down, storage completely down, or site down (power, air conditioning, site damage)**

  This will cause an interruption in normal operation.

- **Physical damage to storage causing data loss (fire, flood)**

  We will need to restore data from backup tapes.

26.4 Two sites up to 100 km apart

In Figure 26-3 we show a basic dual fabric SAN distance extension solution between two sites up to 100 km apart. Each site has a fully redundant infrastructure at the SAN fabric level comprising of two Cisco MDS 9506 Multilayer Directors. The host servers and storage are connected to the SAN fabric via shortwave SFPs using multi-mode fibre cable. The directors are interlinked across the sites via two ISLs each using longwave SFPs and single-mode fiber cables and repeaters. The number of repeaters used depends on the type of longwave SFP used.

One can assume that a greater distance between IT sites results in greater security against wide spread disaster. However, increased distance has its price in terms of interconnection cost, business relocation effort, and network cost. Before you start to expand your storage network, you have to consider if the following solution can meet your business purpose.

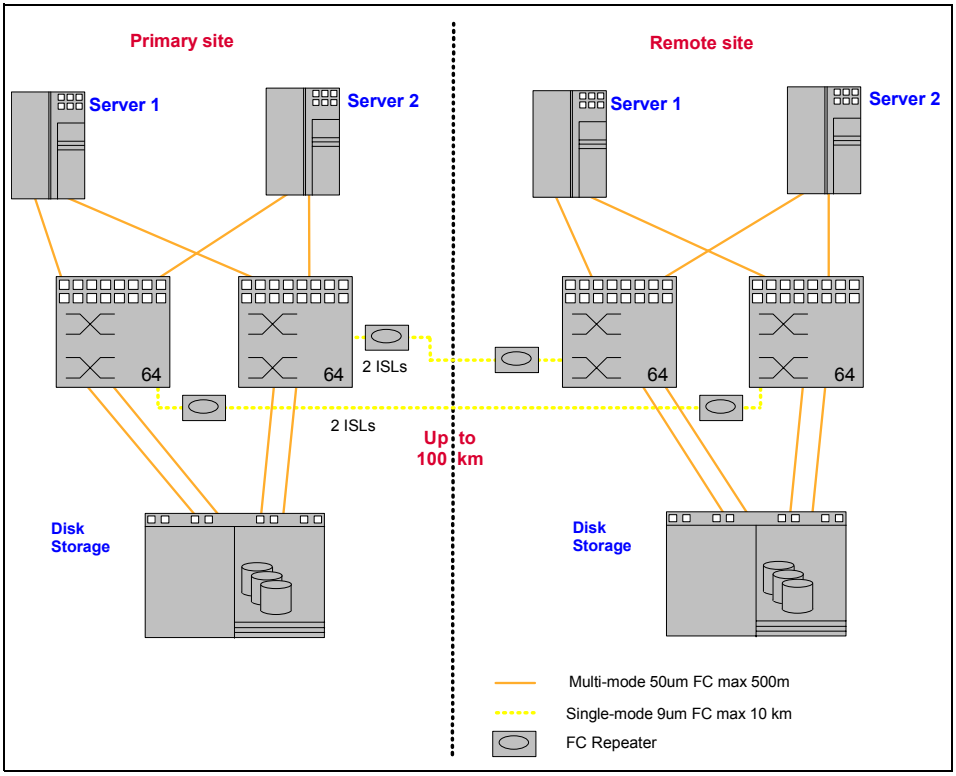


Figure 26-3 SAN distance extension up to 100 km with repeaters

The two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 975, but with the addition of:

- FC repeaters as required to reach 100 km

Checklist

The same items as for “Two sites up to 10 km apart” on page 975 should be considered, with the addition of:

- Repeater count verified for the distance in every ISL link
- Repeaters installed connected and tested

Performance

The same performance criteria apply as for “Two sites up to 10 km apart” on page 975, other than overall ISL latency increases with distance and BB_Credit utilization will increase.

Scalability

The same scalability criteria apply as for “Two sites up to 10 km apart” on page 975.

Security

The same items as for “Two sites up to 10 km apart” on page 975 should be considered, with the addition of:

- Physical repeater security — locked cabinet, restricted access site.

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 975, with the addition of:

- **Repeater failure**

  Impact will depend on whether alternate ISL links are available or not. Only action required is repeater replacement. Physical access to repeater and repeater site required.

26.5 Two sites - Point-to-point DWDM solution

In Figure 26-4 we show a basic high availability SAN design for multiple servers deployed with dual directors and a redundant fabric at each site. This design uses Dense Wavelength Division Multiplexing (DWDM) equipment for inter-site ISLs between the directors.

Note: Refer to Chapter 4, “Extending fiber over distance with DWDM” on page 183 for further information on DWDM technology and topologies.

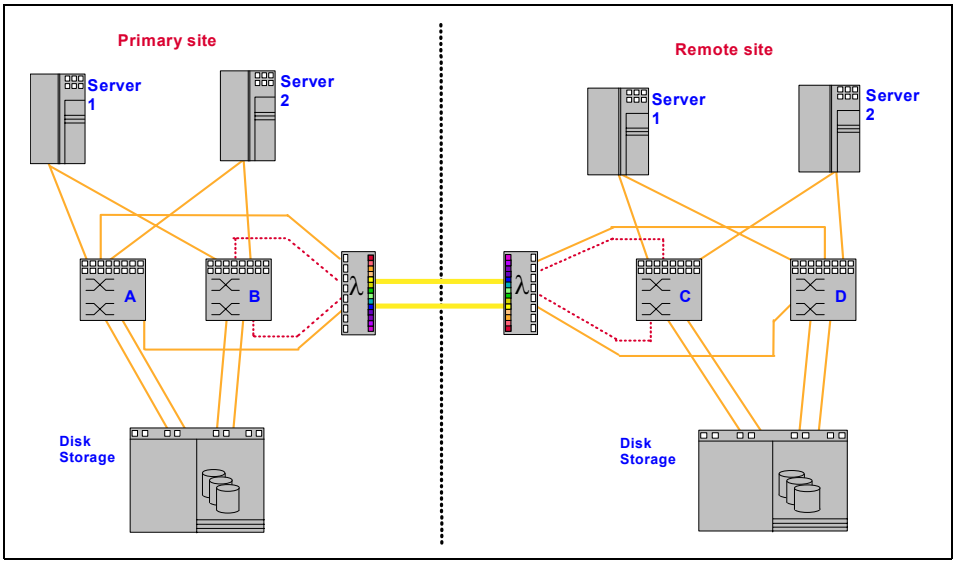


Figure 26-4 Point-to-point DWDM solution - two sites

The use of DWDM allows more channels to be sent over the same number of fiber cables. This is achieved by using different wavelengths of light to carry the different channels at the same time, and also supports transmission of multiple protocols (e.g. FC, ESCON, TCP/IP) over the same fibre. To ensure redundancy, and hence availability, the data should be sent over two completely different physical paths, but with equal length to ensure equal latency.

The method for replicating the data could be either host software based, or via storage hardware based facilities such as PPRC. In this example we are using host based software replication.

In our example, SAN AD (Directors A and D) will be a separate fabric from SAN BC (Directors B and C) fabric. Each fabric comprises of two directors, one at each site.

The two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 975, but with the addition of:

- DWDM equipment linked via dark fibre

Checklist

The same items as for “Two sites up to 10 km apart” on page 975 should be considered, with the addition of:

- The ISL link count to the DWDM equipment at each site

- Maximum distance for individual devices
- Configuration and count of the client interface adapters in the DWDM equipment
- Configuration and count of the transport network adapters in the DWDM, ensuring same frequency band and channel
- Connectivity testing between:
  - ESS to switches
  - Servers to switch
  - Switches to DWDM
- Validate connectivity
- Test server to storage on both paths
- Tests of failover/fail back operations
- Test inter-site connectivity via DWDM
- Setup LUNs on back up ESS and test from primary server, to quantify latency
- Pre-installation testing of the DWDM transport network for configuration and performance

Performance

The same performance criteria apply as for “Two sites up to 10 km apart” on page 975, with the addition of:

- The overall ISL latency increases with distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the director and the DWDM equipment. If the number of ISLs between the directors and DWDM equipment increases, it may be necessary to increase the quantity of dark fibre between the DWDMs that carry the ISL traffic.

Scalability

Most DWDMs support multiple I/O interfaces for client equipment and support up to 32 separate channels. DWDM equipment also allows mixing different interface protocols on a single physical DWDM channel.

The same scalability criteria apply as for “Two sites up to 10 km apart” on page 975.

Security

The same items as for “Two sites up to 10 km apart” on page 975 should be considered, with the addition of:

If you decide to add an ISL, ensure that all checks are done on the zone configuration changes from the management console.

- The DWDM equipment typically provides Web based management software that can be accessed by any workstation connected to the same LAN. Different user levels are provided for administrators, operators or observers. Different passwords must be set to limit access.
- The physical security of fiber connections and patch panels should be considered.

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 975, with the addition of:

- **DWDM optical channel card failure**

  As we configured the channels for high availability, there are redundant cards in the DWDM. If we lose one, the traffic will automatically be switched to the other and the channel will remain available. The failed card can be replaced concurrently.

- **DWDM optical channel manager card failure**

  The optical channel manager card performs path high availability switching. There are two cards in each shelf, if one fails the other takes control and the operation is not affected. The failed card can be replaced concurrently.

- **DWDM Optical Multiplexer failure or shelf backplane failure**

  The entire shelf will be unavailable. As we spread connections over different shelves, we will have at least half the channels of each type available. Operation will continue although performance may be affected.

- **Dark fiber failure**

  As we configured for high availability, operation will continue using the available pair with no performance impact. Customers may have additional considerations based on their business recovery policy to source the redundant dark fiber from another telco/source.

DWDM considerations

We have shown a point-to-point topology here. We would implement the DWDM solution as a protected fiber to ensure availability. In this example we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation, here we have assumed a light workload.

We show here an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one director in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits, however this is not a DWDM consideration but more to do with the general SAN solution design over a distance. The DWDM is a core architecture deployment and because of it's independence from protocol it can be used for other traffic beyond the SAN environment.

In order to maintain performance at extended distances, we need to increase the number of buffers on each interconnecting port to compensate for the number of frames that are in transit. Configuring the director ports connected to the DWDM for 10 to 100 km provides 60 buffers, that is enough for this distance.

26.6 Two sites: Point-to-point DWDM with ESS PPRC

With the IBM TotalStorage Enterprise Storage Server Model 800 running LIC 2.3.0 or later and the PPRC Version 2 licensed feature, it is no longer necessary to use dedicated ESCON ports on the ESS to connect via DWDM to a remote ESS. Remote copy traffic can now share FCP paths with normal SAN traffic, and hence the remote ESS can be accessed via the same connections as shown in Figure 26-4 on page 983.

Checklist

The same items as for "Two sites - Point-to-point DWDM solution" on page 982 should be considered, with the addition of:

- PPRC Version 2 license for the ESS
- Test PPRC - Establish paths and Test PPRC functionality - qualify timings

Performance

This is the same as for “Two sites - Point-to-point DWDM solution” on page 982.

Scalability

This is the same as for “Two sites - Point-to-point DWDM solution” on page 982.

Security

This is the same as for “Two sites - Point-to-point DWDM solution” on page 982.

“What if” failure scenarios

This is the same as for “Two sites - Point-to-point DWDM solution” on page 982.

DWDM considerations

In addition to those listed in “Two sites - Point-to-point DWDM solution” on page 982, we need to consider that the solution shown here also implements IBM’s PPRC between the two ESS subsystems. See the *IBM TotalStorage Enterprise Storage Server Implementing ESS Copy Services in Open Environments*, SG24-5757 for details on implementing PPRC with ESS.

26.7 Multiple site ring DWDM solution

Figure 26-5 illustrates four dual-fabric sites connected in a hubbed ring topology via DWDM and OADM technology. Site 1 is the hub node where all the DWDM channels originate and terminate, whilst the other three sites utilize OADM technology to add or remove channels from the ring.

Note: Refer to Chapter 4, “Extending fiber over distance with DWDM” on page 183 for further information on DWDM technology and topologies.

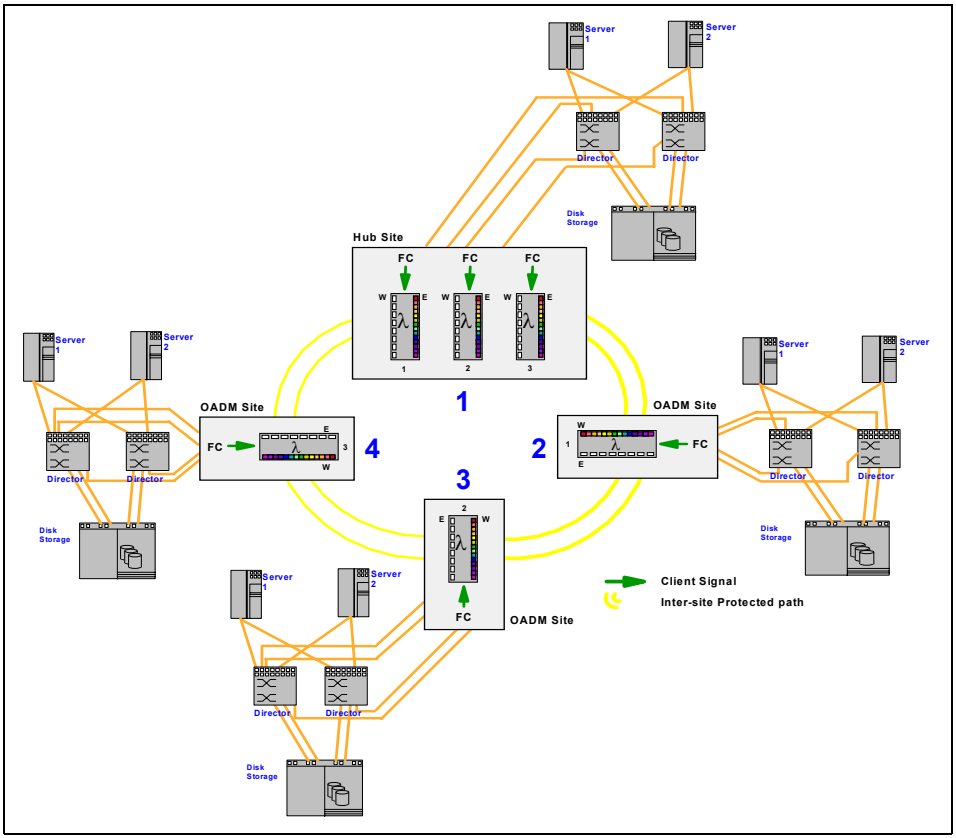


Figure 26-5 Multiple site: Ring topology DWDM solution

Each path has its own wavelength. The OADM pass through feature (express channels) enables non-contiguous sites to connect over an intermediate site as if they were directly connected. The only additional overhead of the pass through is the minimal latency (5 microseconds/km) of the second link. The pass through has no overhead since it is a passive device. The two fabrics would logically appear as fully meshed topologies.

Each of the links can operate in protected mode, which provides a redundant path in the event of a link failure. In most cases, link failures are automatically detected within 50 msec. In this case, the two wavelengths of the failed link reverse directions and reach the target port at the opposite side of the ring. If the link between 1 and 4 fails, the transmitted wavelength from 4 to 1 would reverse direction and reach 1 through 3 and 2. The transmitted wavelength from 1 to 4 would also reverse direction and reach 4 through 2 and 3.

Calculating the distance between nodes in a ring depends on the implementation of the protected path scheme. For instance, if the link between DWDM 2 and 3 fails, the path from 1 to 3 would be 1 to 2, back from 2 to 1 (due to the failed link), 1 to 4, and finally 4 to 3. This illustrates the need to utilize the entire ring circumference (and more, in a configuration with over four nodes) for failover.

Another way to calculate distance between nodes is to set up the protected path in advance (in the reverse direction) so the distance is limited to the number of hops between the two nodes. In either case, the maximum distance between nodes determines the maximum optical reach. An example of this specification is 80 to 100 km for a maximum distance between nodes and 160 to 400 km for maximum ring size. These distances should continue to increase as fiber optic technology advances.

Each site has a set of components the same as one site in the “Two sites - Point-to-point DWDM solution” on page 982, except:

- Site 1 has three sets of DWDM channels
- Sites 2, 3 and 4 each have OADM equipment
- Dark fibre between each site

Checklist

The same items as for “Two sites - Point-to-point DWDM solution” on page 982 should be considered, with the addition of:

- Installation and configuration of the OADM equipment
- Ensuring correct frequency band and channel for each site

Performance

This is the same as for “Two sites - Point-to-point DWDM solution” on page 982.

Scalability

This is the same as for “Two sites - Point-to-point DWDM solution” on page 982.

Security

This is the same as for “Two sites - Point-to-point DWDM solution” on page 982.

“What if” failure scenarios

This is the same as for “Two sites - Point-to-point DWDM solution” on page 982, except that the scenarios for DWDM failures are also applicable to OADMs.

DWDM considerations

We have shown a ring topology here, this gives a logical mesh, potentially giving any to any connectivity. We have implemented this as two rings, one ring connects to one switch in each site, the other ring connects to the remaining switch, this give two discrete SAN fabrics.

We would implement the DWDM solution as a protected ring to ensure availability. In this example we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation, here we have assumed a light workload.

We show here an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one director in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits, however this is not a DWDM consideration but more to do with the general SAN solution design over a distance. The DWDM is a core architecture deployment and because of it's independence from protocol it can be used for other traffic beyond the SAN environment.

In order to maintain performance at extended distances, we need to increase the number of buffers on each interconnecting port to compensate for the number of frames that are in transit. Configuring the director ports connected to the DWDM for 10 to 100 km provides 60 buffers, that is enough for this distance.

26.8 Two sites: Channel extender and WAN extension

Channel extenders typically use telecommunication lines for data transfer and therefore enable application and recovery sites to be located over greater distances. The use of channel extenders provides the separation for disaster recovery purposes and avoids some of the barriers imposed when customers do not have a “right of way” to lay fiber cable.

In Figure 26-6 we show a typical SAN distance extension using Optical Channel extenders at both primary and secondary sites and the ATM network.

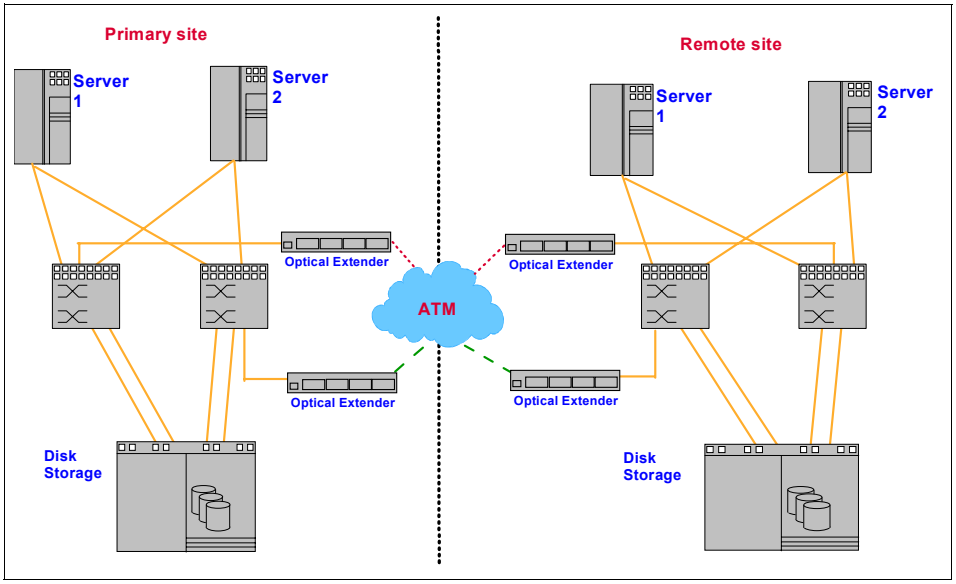


Figure 26-6 SAN extension over (ATM) WAN using channel extenders

Again, the two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 975, but with the addition of:

- Two optical channel extenders
- ATM already available

Checklist

The same items as for “Two sites up to 10 km apart” on page 975 should be considered, with the addition of:

- I/O interface (for example, FC or ESCON) selection and installation on the Optical channel extender
- Network interface (for example, ATM OC3) selection and installation on the Optical channel extender
- Pre-installation testing of the network interfaces (ATM network)

Performance

The same performance criteria apply as for “Two sites up to 10 km apart” on page 975, other than overall ISL latency increases with distance and BB_Credit utilization will increase.

Channel extenders generally compress the data before sending it over the transport network. However the compression ratio needs to be determined based on the individual application characteristics and the distance.

You must especially consider an amount of the updated data for a period of time (peak time), and reflect this in calculating the number of interconnecting links and the number of data volumes for SAN/WAN solution.

Scalability

The same scalability criteria apply as for “Two sites up to 10 km apart” on page 975. Most optical channel extenders support multiple I/O interfaces for client equipment. However in an installation, it is recommended to have a single protocol per interface on a physical channel extender.

Security

The same items as for “Two sites up to 10 km apart” on page 975 should be considered, with the addition of:

- Channel extender security — locked cabinet, restricted access site

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 975, with the addition of:

- **Channel extender failure**

  In the current solution channel extenders are not configured in a redundant mode, however they can be.

- **Transport network failure**

  The telco link failure is not planned for in the current solution, however a redundant link can be acquired from the vendor or a separate telco may be considered depending on the business recovery policy.

26.9 Two sites with FCIP connection

Using the IP Services Module (see “IP Services Module” on page 507) and the SAN Extension Package feature it is possible to connect two remote SANs together to form a single fabric. This enables servers or devices in the primary site to have access to SAN devices in the remote site over TCP/IP networks.

If required, VSANs (see 11.4.8, “Virtual SAN” on page 519) can be used to effectively separate the primary and remote SANs. Inter VSAN Routing (IVR -

see “Inter VSAN Routing” on page 525) can then be used to make visible just those devices which are required from the remote site.

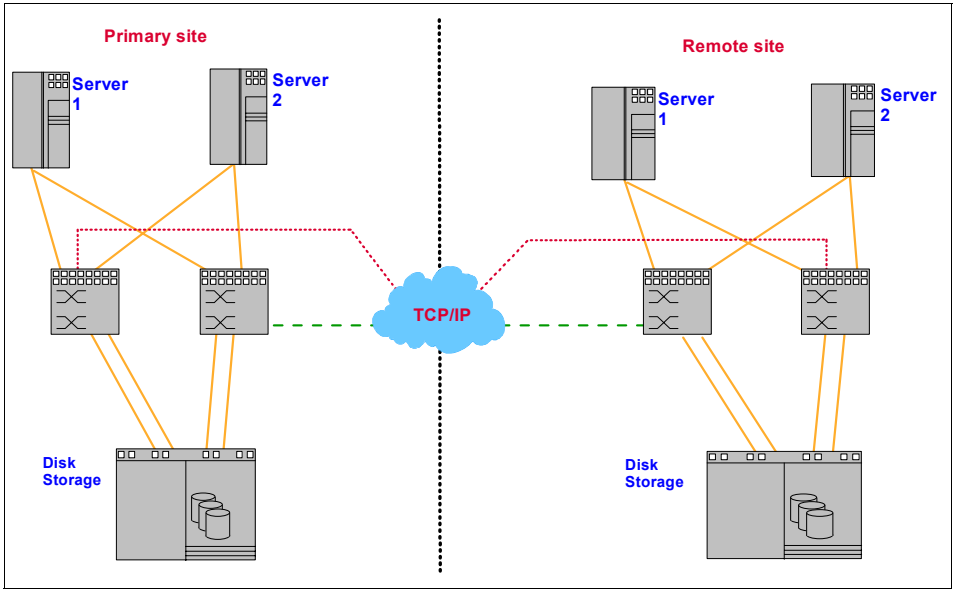


Figure 26-7 Remote access via FCIP switch over TCP/IP

Again, the two sites are each comprised of the same equipment as for “Two sites up to 10 km apart” on page 975, but with the addition of:

- IP Services Module blades for each director
- TCP/IP network already available

Checklist

The same items as for “Two sites up to 10 km apart” on page 975 should be considered, with the addition of:

- SAN Extension Package feature
- Enterprise Package feature for IVR
- FCIP configuration
- Pre-installation testing of the network interfaces (TCP/IP network)

Performance

If VSANs are exploited, the local and remote site fabrics are not merged, so extra latency only applies to those devices accessed via the TCP/IP network. For these devices, similar performance criteria apply as for “Two sites up to 10 km apart” on page 975, with the addition that overall latency increases with distance.

The SAN Extension Package, which activates the FCIP protocol on the IP Services Module blade, includes compression to reduce WAN bandwidth requirements.

Scalability

As well as the normal SAN scalability issues, the WAN bandwidth and number of FCIP switch ports must be managed to support the required amount of remote traffic.

Security

The same items as for “Two sites up to 10 km apart” on page 975 should be considered.

“What if” failure scenarios

The same failure scenarios apply as for “Two sites up to 10 km apart” on page 975, with the addition of:

- **IP Services Module failure**

  In the current solution, a blade failure would cause loss of access to one of the remote fabrics. Since a dual fabric design has been used at both sites, access could still continue via the alternate fabric. It is possible to use more than one FCIP blade to connect two fabrics, such that traffic can continue to flow through the second route in situations where the first route has failed.

- **TCP/IP network failure**

  WAN failure could cause complete loss of access to remote devices. Such situations are beyond the scope of this book.

26.10 Remote tape vaulting

For an existing large corporation with multiple sites, tape library resources can be consolidated to a separate site. This simplifies data movement logistics and centralizes backup software configurations. By doing this the cost of doing business is reduced as the infrastructure is efficiently utilized and fewer IT personnel required for backup data management. In addition, in the event of a disaster, the data is already located on tape in the remote location and there is

no longer a need to ship the data to another site. Another application is for outsourcing services companies like Storage service providers (SSP), that are interested in providing backup and Backup management services to smaller companies.

The Figure 26-8, shows the basic layout for a remote tape vaulting solution, using Cisco directors and switches.

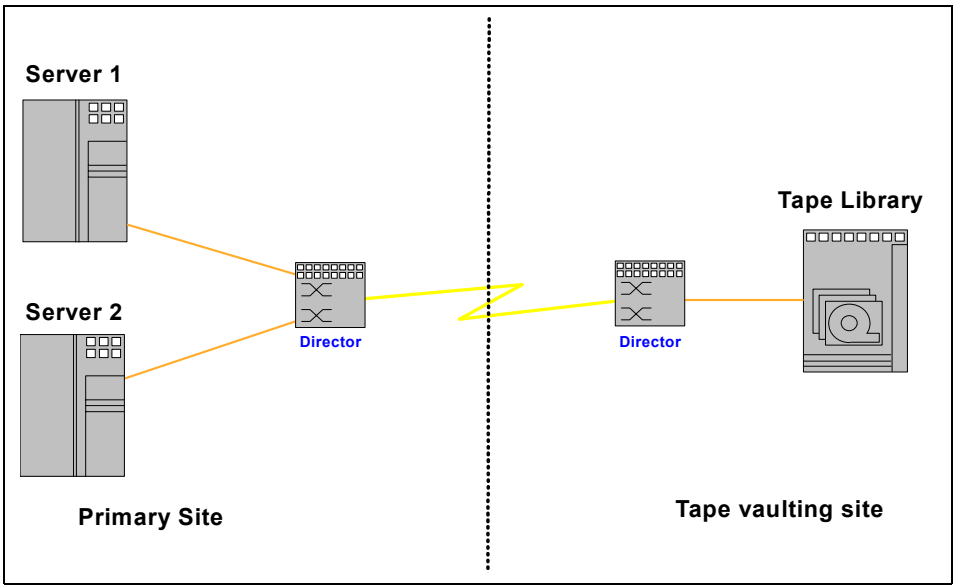


Figure 26-8 Remote tape vaulting

Any of the preceding distance solutions (dark fibre, repeaters, DWDM, FICP, etc) can be used to connect the local and remote directors, depending on circumstances and the distance required.

Cisco directors support FC-AL device connectivity, and hence tape drives can be directly connected to the director. However, due to the relevant costs of director ports versus switch ports, it may still be advisable to use a switch as a tape concentrator. A switch would also be advisable if a director class product is not required at the remote site. Alternatively, install 32 port cards (see “32 port switching module” on page 506) in the director in place of 16 port cards and use them for tape connectivity.

Checklist

In addition to the items considered in the disk consolidation example, we must now consider:

- Host HBA supported for tape drive / library attachment

- Use of separate HBAs for disk and tape access
- Host tape device driver levels
- Host operating system levels compatible with tape library fiber requirements
- Zone configuration allowing tape access to required servers
- Host software tape sharing capabilities

Performance

The performance of the tape drives determines the number of drives which can be connected to the switch, and hence the number of ISLs needed at the remote site to support the bandwidth. Utilization of Port-Channels (see “PortChannels” on page 514) between the directors and switch would also help to optimize ISL usage.

Certain operating system and device driver combinations support path failover for tape (e.g. AIX with Magstar 3590), and as such dual port tape drives can be connected to two switches for redundancy purposes.

Scalability

The solution can be scaled by increasing the number of tape drives in the library, and the number of associated switches. This may also require an increase in capacity of the offsite link.

Security

Here are some security considerations:

- Zoning can be used to restrict server access to specific tape devices as required.
- Proper tape management procedures will avoid servers contending for the same tape device.

“What if” failure scenarios

Here are some theoretical assumptions:

- Performance may be impacted depending on number of drives attached. Traditionally tape failover is a manual operation. Multiple path devices are configured as several logical devices, one per path. Only one of these logical devices is made active. If there is a failure the application aborts and it can then be restarted using a different logical device. For tape drives with dual paths, the latest levels of a tape device driver provide alternate pathing support and tape failover for Fibre Channel connections. With this support enabled, if an error occurs the device driver will automatically initiate error recovery and the operation will continue using the next logical path.

- Device link or device port failure — only one tape drive affected. Recovery may be manual or automatic depending on operating system and driver level as explained for ISL or switch failure.
- Switch port failures — SFPs are hot swappable. Drives can be moved to a spare port.
- Switch failure — most FRUs are hot swappable. With multi-port drives, access can continue via an alternate switch.
- User error trying to access too many drives on the same link — performance of all drives attached via the link may be degraded. Director or switch performance view may be used to find what paths are carrying traffic.
- Tape drive failure in a single tape zone — an alternate zone should be made active to get access to a working device.

26.11 Remote tape vaulting with disaster tolerance

In Figure 26-9 we show an extension or variant of the tape vaulting solution. In this solution the primary site / local site also has a tape library. Data is written to both tape libraries, however in the event of failure on the tape library at the local site, data can be backed up and restored from the tape library at the remote site.

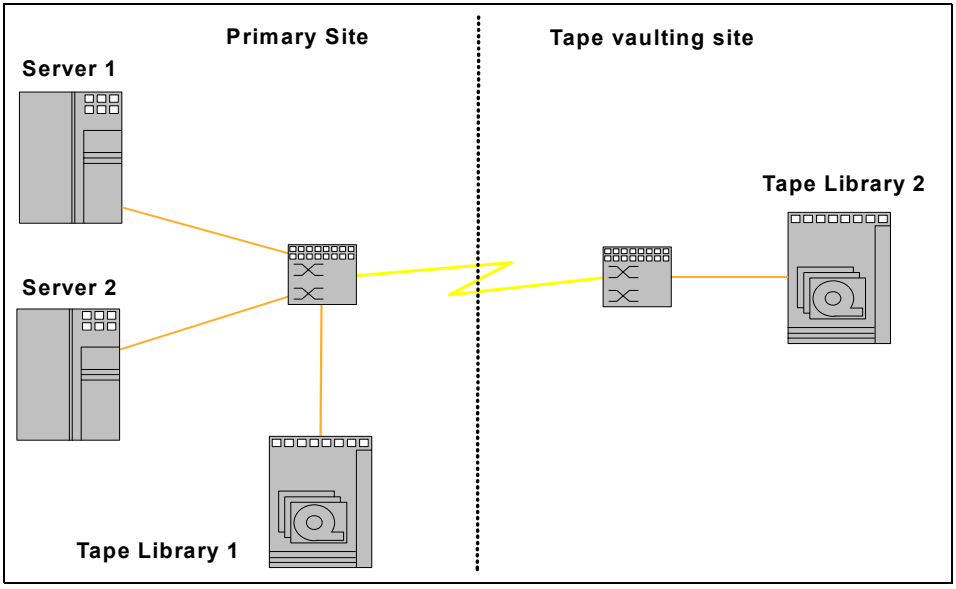


Figure 26-9 Remote tape vaulting with disaster tolerance

This solution has the same performance, scalability, security and failure attributes as the standard remote tape vaulting solution.

27

SAN best practices

In this chapter we take into account the fact that you may have already implemented a SAN, and now you need to make your life easier by managing the SAN more effectively. Here are some recommendations applicable to all four SAN switch/director manufacturers: Brocade, Cisco, CNT and McData (IBM either resells or OEM's these solutions) about the best ways to manage it.

© Copyright IBM Corp. 2004. All rights reserved.

27.1 Scaling

During our case study examples, we have planned for additional growth. When your SAN was initially implemented, there was a strong possibility that it was not planned with sufficient growth to accommodate future projects and functions.

We recommend that you perform an inventory of all HBAs, switches, storage, and ports to determine where you are today. Then apply what you know with regards to anticipated growth and new projects to determine your requirements for the next 18 months. When you have gathered all of this information, we recommend that you review whether the current SAN topology is the most efficient or whether a redesign may be in order.

Ensure that you plan for growth in the following areas:

- Switches/Directors
- Ports
- Space
- Cables
- Bandwidth

27.1.1 How to scale easily

We recommend that you use redundant fabrics wherever possible to remove single points-of-failure and allow for devices to be down when appropriate. This will give you the option to grow one part of the redundant SAN first without disturbing the other part. After the first step is completed, and traffic is restored, you can grow the second part in the same manner.

We recommend that you have a clear definition of who is responsible for performing what actions when it comes to changing the SAN fabric environment, and that you communicate with all parties when a change is going to occur. Document the change within your change management process to allow for refer back at a later date. We recommend that you establish clear and standard procedures for each person that must perform functions that will change the fabric.

27.1.2 How to avoid downtime

We recommend that you use dual connections from each server into two different switches/directors. With multi-path software, this will create an environment that will tolerate HBA and cable failures and replacement.

Similarly, we recommend at least dual connections to the storage device.

When you design a redundant SAN we recommend that you provide enough bandwidth through one part of the SAN. This will give you enough bandwidth in case of upgrades and maintenance.

Ensure that all of your switches, directors and storage devices have redundant power supplies and these power supplies have different power feeds. Some products do not have dual power supplies as standard, so you should order them when you order the switches.

We recommend that you have a supply of spare SFPs and cables. These components are the most susceptible to failure. Use dust caps on these components whenever possible.

27.1.3 Adding a switch/director

When adding a switch or director to your SAN fabric, it is imperative to plan. Know what the end result will look like.

Consider using SAN certified professionals to perform the installation and integration for you. From their experience, they can establish and manage the project plan and may be less prone to causing an outage scenario.

The process also needs considerable planning. Perform changes one port at a time whenever possible to minimize the duration of single points-of-failure. By moving one at a time, you will be sure to move secondary cables (that is, not both connections to the same host) to the new switch or director.

Backup all your configuration before you perform any changes.

Before adding a new switch or director to your SAN ensure that you have performed these actions on the *new* switch/director:

- Clear the configuration, so you do not get conflicts.
- Firmware checks (compatibility) and hardware; try to use the same firmware as on existing switches if possible.

If you are adding a different model of switch/director then check with the manufacturer to ensure that it is compatible, and that it can interoperate with your existing equipment.

27.1.4 Adding ISLs

The most important aspect to consider when adding ISLs is whether or not you are being cost effective in your use of them. We recommend that you consider the use of new SAN functions, such as trunking, to enable more efficient use of multiple ISLs between the same devices. All vendors of SAN components have or are planning to introduce trunking features to their products.

27.1.5 Performance monitoring and reporting

A large factor that needs to be considered when attempting to plan your SAN for growth are the related performance growth requirements. The tools used to accomplish this are considered below in 27.5, “Tools” on page 1007.

27.2 Know your workloads

Before designing your SAN be aware of the following:

- Collect the I/O bandwidth required for your applications.
- Design you SAN for the peaks, not for average traffic.
- Group the servers with high demand directly with storage devices and separate them from other workloads, especially when you have ISLs. In the core-edge design this would imply connecting servers with high demand to the core switches.

  We show an example of this setup in Figure 27-1.

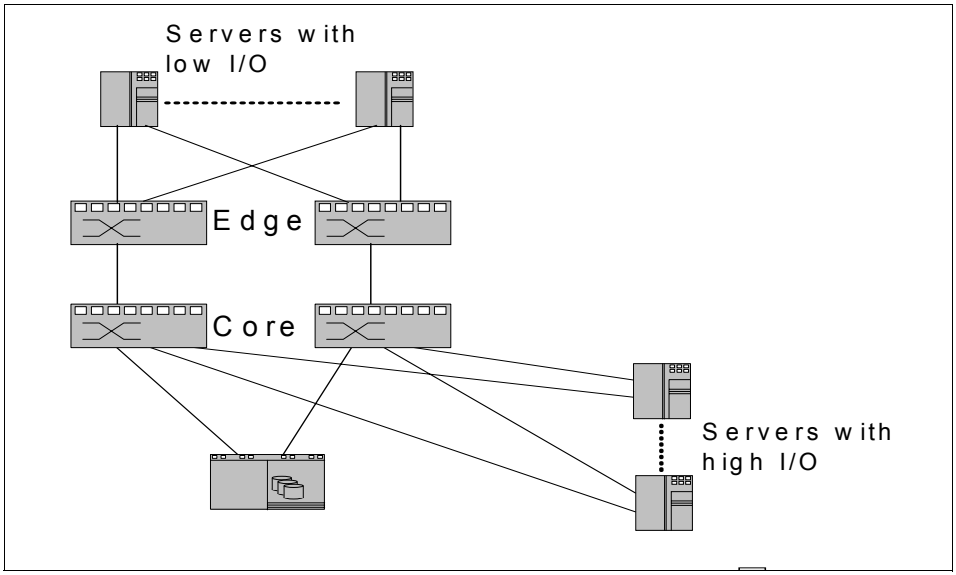


Figure 27-1 Connecting high I/O servers to the core switches

- Constantly monitor the utilization to see if the real world matches your design expectations.
- Increase or decrease the number of server ports or ISLs if this will solve your performance problems.

27.3 Port placement

Some port placement thoughts.

IBM TotalStorage SAN Switch

When connecting devices to the SAN it is important to follow some rules:

- Put the devices (servers and storage) which communicate a lot to each other on the same switch/director if possible.
- Group the devices which communicate a lot to each other on the same ASIC . The H08, H16, and M14 models have eight ports per ASIC; while models F16, F32, and M12 have four ports per ASIC.
- Separate two connections from the same device (server or storage) to two switches/directors if possible. Build redundant fabrics whenever possible.
- Try to avoid single points of failure by using multipathing software and multiple connections to storage devices.

- When connecting the same device on the same switch/director (if you really want to do this), spread the connections to different blades and different ASICs.
- If you plan to use some ports for extra long distances (for example, more than 10 km) keep in mind that some of the vendors have a shared pool of buffers per ASIC. So in that case, you might only be able to allocate one port for long distance communication.
- When you have more switches in the same SAN, it is recommended that you use the same port numbers for the same functions, for easier management. For example, when connecting storage devices to several switches always use Port 0 and 3 on the switch for them.

Cisco switch/director

When connecting devices to the SAN, it is important to follow some rules:

- Place servers and storage devices which communicate a lot with each other on the same switch/Director if possible.
- Group the devices which communicate a lot with each other on the same blade inside the Director.
- Build redundant fabrics whenever it is possible by separating two connections from the same device (servers and storage) to two switches/directors if possible.
- Prevent single points-of-failure by using multipathing software and multiple connections to storage devices.
- If you have to connect the same device on the same switch/Director, spread the connections to different blades and different ASICs.
- When you have more switches or Directors in the same SAN, it is recommended that you use the same port numbers for the same functions, for easier management.

CNT

When connecting devices to the SAN it is important to follow some rules:

- For devices (servers and storage) which have a large volume of communication with each other:
  - Put them on the same switch/Director if possible.
  - Group them on the same blade inside the switch/Director.
  - Consider putting them on ports which share an ASIC.
- Separate two connections from the same device (server or storage) to two switches/Directors if possible. Build a redundant fabric whenever it is possible.

- Try to avoid single points-of-failure by using multipathing software and multiple connections to storage devices.
- When connecting two or more ports from one device to one switch/Director (for extra bandwidth rather than resilience) the connections should be spread across different blades.
- If you plan to use some ports for extra long distances, for example, more than 10 km, keep in mind that when interconnecting equipment from different vendors, that:
  - Some vendors have a shared pool of buffers per ASIC. So, in that case you may only be able to allocate one port for long distance communication.
  - CNT have separate buffer pools for each port.
  - All ports on the CNT FC/9000 have either 64 BB_Credits for 1 Gb/s or 120 BB_Credits for 2Gb/s ports.
- When there is more than one switch or Director in the same SAN, it is recommended that the same port numbers be used for the same functions, for easier management. For example, the recommendation from CNT is that servers be connected across the Director starting at the top left and that storage be connected across the Director starting at the bottom left. This is shown in Figure 27-2.

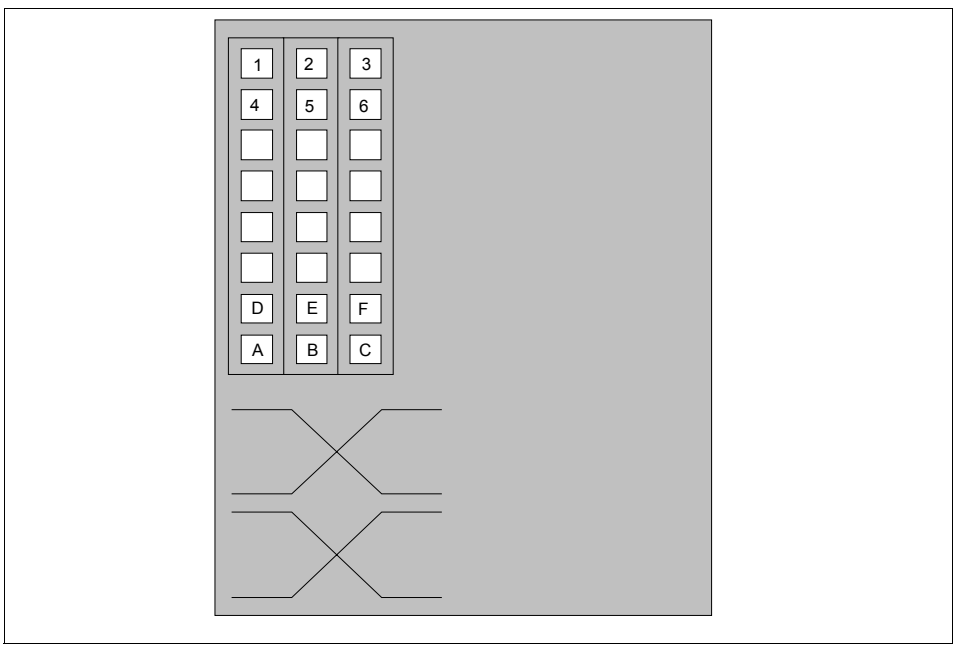


Figure 27-2 The CNT port population guidelines

In this case, the server ports are connected to the ports labelled 1, 2, 3, 4, and so on. The storage ports will be connected to the ports labelled A, B, C, D, and so on.

Note: The actual ports on an CNT FC/9000 FIO blade are numbered 1 through 8 from the top to the bottom. The top left hand port is referred to as Blade 1, port 1, the one to its right is Blade 2, port 1 and the one below that is Blade 2, port 2.

McDATA

When connecting devices to the SAN, it is important to follow some rules:

- Place servers and storage devices which communicate a lot with each other on the same switch/Director if possible.
- Group the devices which communicate a lot with each other on the same ASIC and blade inside the switch/Director.
- Build redundant fabrics whenever it is possible by separating two connections from the same device (servers and storage) to two switches/directors if possible.
- Prevent single points-of-failure by using multipathing software and multiple connections to storage devices.
- If you have to connect the same device on the same switch/Director, spread the connections to different blades and different ASICs.
- If you plan to use some ports for extra long distances, for example more than 10 km, keep in mind that with McDATA switches/Directors, additional buffer credits are available but are not enabled by default. This means that they should be enabled if you want to use them for long distance connections. This can be performed through the EFC management server.
- When you have more switches or Directors in the same SAN, it is recommended that you use the same port numbers for the same functions, for easier management.

27.4 WWNs

You should keep track of the WWNs used in the SAN. For example, keep documentation which will show which port WWN is used in which device and to which SAN port in the fabric it is connected. Also, for ease of management it is recommended that you assign aliases to the WWN used. Later on when you

introduce zoning, you can use those aliases instead of the WWNs. Some devices (e.g. IBM ESS) have their node WWN available, and we can expect that in the future this will become a common practice. This node WWN can be used for zoning purposes instead of using each port WWN in that device.

27.5 Tools

Typical features that the GUI tools provide from all manufacturers are:

- Global view of vendor specific SAN (both logical and physical)
- Ability to view current firmware levels and potentially propagate new levels
- View and change Zones and/or Configurations
- Ability to configure and activate SNMP monitoring
- Ability to configure and activate syslog monitoring
- Status indicators (usually graphical representation of LEDs)
- Performance monitors (usually capturing instantaneous data but storing high and watermarks)

The tools that you use to manage your SAN will probably support both inband and/or outband mechanisms for accessing the switches. If you use ISLs between all of your switches, then inband becomes an option. We recommend that you always use outband management, and compliment with inband management where appropriate. If an ISL does experience difficulties, and you rely solely on inband management, then you will lose all management capabilities to that device.

SNMP support is available and they tend to include standard Fibre Channel Alliance MIBs support along with their custom private MIB.

Scripting can be implemented in a number of ways. If we consider scripting programs, then telnet functions can be scripted to execute commands and capture responses. Scripting can also imply having hardcopy procedures to follow, this is covered further in 27.6, “Documentation” on page 1009.

Analyzers are typically used by field engineers to perform fault isolation by examining traced FC traffic. These devices are typically connected in-line between one switch port and a device. We recommend that the infrequent use of analyzers be left with the consultant or switch manufacturer, as they will fully understand the format of the frames being traced and how to use the hardware and software known collectively as the analyzer.

All switch devices store some forms of logs. There may be more than one log on each switch and these may become consolidated at the management server level. Logs tend to be proprietary in format and may be separated between error logs, event logs, audit logs, and hardware logs.

Beside using the built in features and management tools from the vendors, you can also use other tools which give you the option to manage components from different vendors. The tool for this from the IBM portfolio is Tivoli Storage Network Manager, which is partially based on the award wining network management software Tivoli NetView.

Tivoli Storage Network Manager is a comprehensive SAN and disk storage resource management software solution. It provides four main value points:

- Assists in the maintenance and health of the SAN infrastructure providing improved availability of the SAN and therefore continuous data access for application processing (minimizes application downtime).
- Provides a simple, secure, and efficient method to identify and allocate heterogeneous disk resources to host systems across the SAN (data integrity).
- Applies monitoring policies that allow automated execution of tasks to add and extend needed disk resources across the SAN to maintain application processing and to reduce administrative workload.
- Gathers and stores relevant SAN infrastructure performance, capacity, activity data to assist in making decisions for SAN growth and stability through the use of Tivoli Decision Support guides.

This product is a key component of the overall Tivoli Storage Management Solutions strategy. It is Tivoli's first offering for comprehensive SAN and storage resource management and will help establish our presence in the emerging SAN marketplace. In addition, it is a product that can operate stand-alone or integrate with Tivoli NetView, Tivoli Enterprise Console, Tivoli Decision Support, and ultimately with GEM (or its replacement) as an application that feeds an overall system management view of the enterprise for operational integrity of the business.

Tivoli Storage Network Manager is for enterprise customers who need to confidently deploy and maintain their SAN infrastructure and disk storage resources. Tivoli Storage Network Manager is a single comprehensive solution package that discovers, allocates, monitors, automates and manages SAN fabric components and disk storage resources that is built upon a scalable architecture to manage large complex configurations. Unlike other SAN offerings that are

limited in function and may use limiting proprietary interfaces, Tivoli Storage Network Manager utilizes industry standards, reduces the number of management interfaces, and when integrated with Tivoli NetView, provides LAN, WAN, and SAN management functions from the same console.

27.6 Documentation

You should keep the following information or documentation ready at your disposal:

- Configuration diagrams
- Logical configuration
- Physical setup of the SAN with cabling infrastructure
- Phone numbers of the devices which support a dial in function, so this can be used in case intervention is needed from the support organization
- What the firmware levels are of the SAN components
- Serial numbers in case the components have to be replaced or somehow identified by it
- Phone number of the support staff to be called in case of problems, and the procedures for reporting problems
- Access information, IPs, user IDs, passwords, keys for the people who perform any kind of operations in the SAN
- Change management should be implemented, so that all the changes (reallocation, re-cabling) are documented and accessible to someone who was not involved in the change process
- All the documentation from the supplier of the equipment has to be accessible if needed

27.7 Configurations

After the SAN is implemented, it is important to document the implementation. This also implies that all configuration and setup changes performed on the SAN components should be documented and also saved in a form which can be easily reused by the setup tools in case components are replaced.

It is also recommended that you back up the configurations after each change, and that you keep a history of changes so you can easily recover from mistakes.

27.8 Common practice faults

Here are some of the most common mistakes made in SAN setup and management:

- Bad cabling practice: Try to mark all your cables and have a clear picture how it is all connected.
- Broken cables, because of bad cabling practices. For example:
  - Pulling cable ties too tight
  - Having too small a radius of curvature
  - Leaving cables hanging from connections with no support
  - Lack of strain relief on cables
  - Not using dust caps
- When you introduce an additional switch, you should clear its configuration.
- When you zone, do not forget to propagate zones to all switches if this is not done automatically.
- When you create the first zone, this does not imply that all the other devices outside this zone are now in some kind of a general zone. Every SAN port must belong to at least one zone after zones are introduced.
- If you overlap zones, check them twice before you activate them.
- If you change zones based on policy. For example, if you have a different zone for backups, do not forget to change them back to normal operations when you do not need them anymore.
- Do not just add an ISL because you think that you have congestion on existing ones. Measure the traffic and predict what will happen with the introduction of a new ISL. Keep in mind that FSPF will react to this and the situation can become worse.

27.9 Zoning

In Fibre Channel environments, the grouping together of the multiple ports to form a virtual private network is called zoning. Ports that are members of a group zone can communicate with each other but are isolated from ports in other zones. In this section we describe some zoning best practices of different SAN switch/director brands:

IBM TotalStorage SAN Switch

Upon coming across these zoning practices, it is assumed that all requirements to support them are in place and accounted for in the planning process. Here are the practices to be considered in zoning:

- It is recommended not to mix HBA vendors within a zone because they do not all handle RSCNs the same way.
- LUN masking on the storage devices must be take into consideration while doing zoning.
- It is highly recommended to make backup copies of zoning definitions.
- For ease of management, the use of aliases are recommended. So in case of replacing the port in the server or storage device, you just update the nickname, and not all of the associated zones.
- Isolate tape and disk storage using the same HBA. This will prevent performance issues when both types of storage are being accessed concurrently.
- Hardware zones are done at level of ASIC.
- When adding a new director or switch, its configuration must be cleared or there may be a conflict in zoning.
- Whenever you have problems accessing device across a SAN, you should check the zone definition first. Sometimes, problem like these are caused by zone definitions. Therefore, any changes made to the zones must be documented for future use.
- There is no need to include any E-ports in a zone.

Cisco Directors

Upon coming across these zoning practices, it is assumed that all requirements to support them are in place and accounted for in the planning process. Here are the practices to be considered in zoning:

- It is highly recommended to make backup copies of zoning definitions. Changes to the full database must be explicitly saved using **`copy running-config startup-config`** command. This will make sure changes are not lost upon reboot.
- Ensure that are no mixed vendor HBAs being used within a zone because they do not all handle RSCNs the same way.
- Isolate tape and disk storage using the same HBA. This will prevent performance issues when both types of storage are being accessed concurrently.
- The port WWN based zoning is recommended only in **`interop`** mode while connecting to non-Cisco switches. Interface based zoning is recommended for Cisco SAN.
- To avoid accidental access to LUNs, configure default facility as **`deny`**. This policy needs to be configured on each switch in the fabric and it does not get propagated automatically.

- Although a zone or zone set same name is allowed in each VSAN, it is a best practice to use a unique name. This will allow a much easier reference.
- Whenever you have problems accessing device across a SAN, you should check the zone definition first. Sometimes, problem like these are caused by zone definitions. Therefore, any changes made to the zones must be documented for future use.
- There is no need to include any E-ports in a zone.

CNT

Upon coming across these zoning practices, it is assumed that all requirements to support them are in place and accounted for in the planning process. Here are the practices to be considered in zoning:

- LUN masking on the storage devices must be take into consideration while doing zoning.
- It is highly recommended to make backup copies of zoning definitions.
- For ease of management, the use of aliases are recommended. So in case of replacing the port in the server or storage device, you just update the alias, and not all of the associated zones.
- Ensure that are no mixed vendor HBAs being used within a zone because they do not all handle RSCNs the same way.
- Isolate tape and disk storage using the same HBA. This will prevent performance issues when both types of storage are being accessed concurrently.
- When adding a new director or switch, its configuration must be cleared or there may be a conflict in zoning.
- Whenever you have problems accessing device across a SAN, you should check the zone definition first. Sometimes, problem like these are caused by zone definitions. Therefore, any changes made to the zones must be documented for future use.
- There is no need to include any E_Ports in a zone.

McDATA directors and switches

Upon coming across these zoning practices, it is assumed that all requirements to support them are in place and accounted for in the planning process. Here are the practices to be considered in zoning:

- In McDATA, default zoning comes enabled by default which allows unzoned devices to communicate with each other but not with the members of any other zone. It is recommended to disable the default zone when you first implement zone.

- LUN masking on the storage devices must be take into consideration while doing zoning.
- It is highly recommended to make backup copies of zoning definitions.
- For ease of management, the use of nicknames are recommended. So in case of replacing the port in the server or storage device, you just update the nickname, and not all of the associated zones.
- Ensure that are no mixed vendor HBAs being used within a zone because they do not all handle RSCNs the same way.
- Isolate tape and disk storage using the same HBA. This will prevent performance issues when both types of storage are being accessed concurrently.
- When adding a new director or switch, its configuration must be cleared or there may be a conflict in zoning.
- Whenever you have problems accessing device across a SAN, you should check the zone definition first. Sometimes, problem like these are caused by zone definitions. Therefore, any changes made to the zones must be documented for future use.
- There is no need to include any E-ports in a zone.

27.10 Powering a SAN up or down

After the SAN is setup and you have defined all your configurations, you should collect data about the domain IDs for each switch/director. When the fabric is first configured the first switch/director in the fabric will become the principal switch/director with domain ID 1. All the others will become subordinate switches and they will receive their domain ID from the principal switch.

When you power down a switch/director you are in the same situation as if the switch/director had failed. If the domain ID used by the switch is the same as original, then you do not need to worry about the power on and off procedures. But in a case where you have replaced the switch with a new one, you should manually setup the domain ID to be the same as the switch that was replaced.

Note: If the domain ID of the switch changes you should change all your hard zones definitions.

If the domain ID of the switches are setup manually, the order of powering the switches on does not matter. But, it is still recommended that you perform the power on sequence in the same order every time if possible.

When the principal switch is powered down, the switch with the lowest WWN in the fabric will become the principal.

Glossary

8b/10b A data encoding scheme developed by IBM, translating byte-wide data to an encoded 10-bit format. Fibre Channel's FC-1 level defines this as the method to be used to encode and decode data transmissions over the Fibre channel.

active configuration. In an ESCON environment, the ESCON Director configuration determined by the status of the current set of connectivity attributes. Contrast with *saved configuration*.

Adapter A hardware unit that aggregates other I/O units, devices or communications links to a system bus.

ADSM ADSTAR Distributed Storage Manager.

Agent (1) In the client-server model, the part of the system that performs information preparation and exchange on behalf of a client or server application. (2) In SNMP, the word agent refers to the managed system. See also: Management Agent

Aggregation In the Storage Networking Industry Association Storage Model (SNIA), "virtualization" is known as "aggregation.This aggregation can take place at the file level or at the level of individual blocks that are transferred to disk.

AIT Advanced Intelligent Tape - A magnetic tape format by Sony that uses 8mm cassettes, but is only used in specific drives.

AL See Arbitrated Loop

allowed. In an ESCON Director, the attribute that, when set, establishes dynamic connectivity capability. Contrast with *prohibited*.

AL_PA Arbitrated Loop Physical Address

ANSI American National Standards Institute - The primary organization for fostering the development of technology standards in the United States. The ANSI family of Fibre Channel documents provide the standards basis for the Fibre Channel architecture and technology. See FC-PH

APAR. See *authorized program analysis report*.

authorized program analysis report (APAR). A report of a problem caused by a suspected defect in a current, unaltered release of a program.

Arbitration The process of selecting one respondent from a collection of several candidates that request service concurrently.

Arbitrated Loop A Fibre Channel interconnection technology that allows up to 126 participating node ports and one participating fabric port to communicate.

ATL Automated Tape Library - Large scale tape storage system, which uses multiple tape drives and mechanisms to address 50 or more cassettes.

ATM Asynchronous Transfer Mode - A type of packet switching that transmits fixed-length units of data.

Backup A copy of computer data that is used to recreate data that has been lost, mislaid,

© Copyright IBM Corp. 2004. All rights reserved.

corrupted, or erased. The act of creating a copy of computer data that can be used to recreate data that has been lost, mislaid, corrupted or erased.

Bandwidth Measure of the information capacity of a transmission channel.

basic mode. A S/390 or zSeries central processing mode that does not use logical partitioning. Contrast with *logically partitioned (LPAR) mode*.

blocked. In an ESCON and FICON Director, the attribute that, when set, removes the communication capability of a specific port. Contrast with *unblocked*.

Bridge (1) A component used to attach more than one I/O unit to a port. (2) A data communications device that connects two or more networks and forwards packets between them. The bridge may use similar or dissimilar media and signaling systems. It operates at the data link level of the OSI model. Bridges read and filter data packets and frames.

Bridge/Router A device that can provide the functions of a bridge, router or both concurrently. A bridge/router can route one or more protocols, such as TCP/IP, and bridge all other traffic. See also: Bridge, Router

Broadcast Sending a transmission to all N_Ports on a fabric.

byte. (1) In fibre channel, an eight-bit entity prior to encoding or after decoding, with its least significant bit denoted as bit 0, and most significant bit as bit 7. The most significant bit is shown on the left side in FC-FS unless otherwise shown. (2) In S/390 architecture or zSeries z/Architecture (and FICON), an eight-bit entity prior to encoding or after decoding, with its least significant bit denoted as bit 7, and most significant bit as bit 0. The most significant bit is shown on the left side in S/390 architecture and zSeries z/Architecture.

Cascaded switches. The connecting of one Fibre Channel switch to another Fibre Channel switch, thereby creating a cascaded switch route between two N_Nodes connected to a fibre channel fabric.

chained. In an ESCON environment, pertaining to the physical attachment of two ESCON Directors (ESCDs) to each other.

channel. (1) A processor system element that controls one channel path, whose mode of operation depends on the type of hardware to which it is attached. In a channel subsystem, each channel controls an I/O interface between the channel control element and the logically attached control units. (2) In the ESA/390 or zSeries architecture (z/Architecture), the part of a channel subsystem that manages a single I/O interface between a channel subsystem and a set of controllers (control units).

channel I/O A form of I/O where request and response correlation is maintained through some form of source, destination and request identification.

channel path (CHP). A single interface between a central processor and one or more control units along which signals and data can be sent to perform I/O requests.

channel path identifier (CHPID). In a channel subsystem, a value assigned to each installed channel path of the system that uniquely identifies that path to the system.

channel subsystem (CSS). Relieves the processor of direct I/O communication tasks, and performs path management functions. Uses a collection of subchannels to direct a

channel to control the flow of information between I/O devices and main storage.

channel-attached. (1) Pertaining to attachment of devices directly by data channels (I/O channels) to a computer. (2) Pertaining to devices attached to a controlling unit by cables rather than by telecommunication lines.

CHPID. Channel path identifier.

CIFS Common Internet File System

cladding. In an optical cable, the region of low refractive index surrounding the core. See also *core* and *optical fiber.*

Class of Service A Fibre Channel frame delivery scheme exhibiting a specified set of delivery characteristics and attributes.

Class-1 A class of service providing dedicated connection between two ports with confirmed delivery or notification of non-deliverability.

Class-2 A class of service providing a frame switching service between two ports with confirmed delivery or notification of non-deliverability.

Class-3 A class of service providing frame switching datagram service between two ports or a multicast service between a multicast originator and one or more multicast recipients.

Class-4 A class of service providing a fractional bandwidth virtual circuit between two ports with confirmed delivery or notification of non-deliverability.

Class-6 A class of service providing a multicast connection between a multicast originator and one or more multicast recipients with confirmed delivery or notification of non-deliverability.

Client A software program used to contact and obtain data from a *server* software program on another computer -- often across a great distance. Each *client* program is designed to work specifically with one or more kinds of server programs and each server requires a specific kind of client program.

Client/Server The relationship between machines in a communications network. The client is the requesting machine, the server the supplying machine. Also used to describe the information management relationship between software components in a processing system.

Cluster A type of parallel or distributed system that consists of a collection of interconnected whole computers and is used as a single, unified **computing resource.**

CNC. Mnemonic for an ESCON channel used to communicate to an ESCON-capable device.

configuration matrix. In an ESCON environment or FICON, an array of connectivity attributes that appear as rows and columns on a display device and can be used to determine or change active and saved ESCON or FICON director configurations.

connected. In an ESCON Director, the attribute that, when set, establishes a dedicated connection between two ESCON ports. Contrast with *disconnected.*

connection. In an ESCON Director, an association established between two ports that provides a physical communication path between them.

connectivity attribute. In an ESCON and FICON Director, the characteristic that

determines a particular element of a port's status. See *allowed, prohibited, blocked, unblocked, (connected and disconnected)*.

control unit. A hardware unit that controls the reading, writing, or displaying of data at one or more input/output units.

Controller A component that attaches to the system topology through a channel semantic protocol that includes some form of request/response identification.

core. (1) In an optical cable, the central region of an optical fiber through which light is transmitted. (2) In an optical cable, the central region of an optical fiber that has an index of refraction greater than the surrounding cladding material. See also *cladding* and *optical fiber*.

coupler. In an ESCON environment, link hardware used to join optical fiber connectors of the same type. Contrast with *adapter*.

Coaxial Cable A transmission media (cable) used for high speed transmission. It is called *coaxial* because it includes one physical channel that carries the signal surrounded (after a layer of insulation) by another concentric physical channel, both of which run along the same axis. The inner channel carries the signal and the outer channel serves as a ground.

CRC Cyclic Redundancy Check - An error-correcting code used in Fibre Channel.

CTC. (1) Channel-to-channel. (2) Mnemonic for an ESCON channel attached to another ESCON channel, where one of the two ESCON channels is defined as an ESCON CTC channel and the other ESCON channel would be defined as a ESCON CNC channel (3) Mnemonic for a FICON channel supporting a CTC Control Unit function logically or physically connected to another FICON channel that also supports a CTC Control Unit function. FICON channels supporting the FICON CTC control unit function are defined as normal FICON native (FC) mode channels.

CVC. Mnemonic for an ESCON channel attached to an IBM 9034 convertor. The 9034 converts from ESCON CVC signals to parallel channel interface (OEMI) communication operating in block multiplex mode (Bus and Tag). Contrast with *CBY*.

DASD Direct Access Storage Device - any on-line storage device: a disc, drive or CD-ROM.

DAT Digital Audio Tape - A tape media technology designed for very high quality audio recording and data backup. DAT cartridges look like audio cassettes and are often used in mechanical auto-loaders. typically, a DAT cartridge provides 2GB of storage. But new DAT systems have much larger capacities.

Data Sharing A SAN solution in which files on a storage device are shared between multiple hosts.

Datagram Refers to the Class 3 Fibre Channel Service that allows data to be sent rapidly to multiple devices attached to the fabric, with no confirmation of delivery.

DDM. See *disk drive module.*

dedicated connection. In an ESCON Director, a connection between two ports that is not affected by information contained in the transmission frames. This connection, which restricts those ports from communicating with any other port, can be established or removed only as a result of actions performed by a host control program or at the ESCD console. Contrast with *dynamic connection*.

Note: The two links having a dedicated connection appear as one continuous link.

default. Pertaining to an attribute, value, or option that is assumed when none is explicitly specified.

Dense Wavelength Division Multiplexing (DWDM). The concept of packing multiple signals tightly together in separate groups, and transmitting them simultaneously over a common carrier wave.

destination. Any point or location, such as a node, station, or a particular terminal, to which information is to be sent. An example is a Fibre Channel fabric F_Port; when attached to a fibre channel N_port, communication to the N_port via the F_port is said to be to the F_Port destination identifier (D_ID).

device. A mechanical, electrical, or electronic contrivance with a specific purpose.

device address. (1) In ESA/390 architecture and zSeries z/Architecture, the field of an ESCON device-level frame that selects a specific device on a control unit image. (2) In the FICON channel FC-SB-2 architecture, the device address field in an SB-2 header that is used to select a specific device on a control unit image.

device number. (1) In ESA/390 and zSeries z/Architecture, a four-hexadecimal character identifier (for example, 19A0) that you associate with a device to facilitate communication between the program and the host operator. (2) The device number that you associate with a subchannel that uniquely identifies an I/O device.

dB Decibel - a ratio measurement distinguishing the percentage of signal attenuation between the input and output power. Attenuation (loss) is expressed as dB/km

direct access storage device (DASD). A mass storage medium on which a computer stores data.

disconnected. In an ESCON Director, the attribute that, when set, removes a dedicated connection. Contrast with *connected*.

disk. A mass storage medium on which a computer stores data.

disk drive module (DDM). A disk storage medium that you use for any host data that is stored within a disk subsystem.

Disk Mirroring A fault-tolerant technique that writes data simultaneously to two hard disks using the same hard disk controller.

Disk Pooling A SAN solution in which disk storage resources are pooled across multiple hosts rather than be dedicated to a specific host.

distribution panel. (1) In an ESCON and FICON environment, a panel that provides a central location for the attachment of trunk and jumper cables and can be mounted in a rack, wiring closet, or on a wall.

DLT Digital Linear Tape - A magnetic tape technology originally developed by Digital Equipment Corporation (DEC) and now sold by Quantum. DLT cartridges provide storage capacities from 10 to 35GB.

duplex. Pertaining to communication in which data or control information can be sent and received at the same time, from the same node. Contrast with *half duplex.*

duplex connector. In an ESCON environment, an optical fiber component that

terminates both jumper cable fibers in one housing and provides physical keying for attachment to a duplex receptacle.

duplex receptacle. In an ESCON environment, a fixed or stationary optical fiber component that provides a keyed attachment method for a duplex connector.

dynamic connection. In an ESCON Director, a connection between two ports, established or removed by the ESCD and that, when active, appears as one continuous link. The duration of the connection depends on the protocol defined for the frames transmitted through the ports and on the state of the ports. Contrast with *dedicated connection.*

dynamic connectivity. In an ESCON Director, the capability that allows connections to be established and removed at any time.

Dynamic I/O Reconfiguration. A S/390 and z/Architecture function that allows I/O configuration changes to be made non-disruptively to the current operating I/O configuration.

ECL Emitter Coupled Logic - The type of transmitter used to drive copper media such as Twinax, Shielded Twisted Pair, or Coax.

ELS. See *Extended Link Services.*

EMIF. See *ESCON Multiple Image Facility.*

E_Port Expansion Port - a port on a switch used to link multiple switches together into a Fibre Channel switch fabric.

Enterprise Network A geographically dispersed network under the auspices of one organization.

Enterprise System Connection (ESCON). (1) An ESA/390 computer peripheral interface. The I/O interface uses ESA/390 logical protocols over a serial interface that configures attached units to a communication fabric. (2) A set of IBM products and services that provide a dynamically connected environment within an enterprise.

Enterprise Systems Architecture/390 (ESA/390). An IBM architecture for mainframe computers and peripherals. Processors that follow this architecture include the S/390 Server family of processors.

Entity In general, a real or existing thing from the Latin ens, or being, which makes the distinction between a thing's existence and it qualities. In programming, engineering and probably many other contexts, the word is used to identify units, whether concrete things or abstract ideas, that have no ready name or label.

ESA/390. See *Enterprise Systems Architecture/390*.

ESCD. Enterprise Systems Connection (ESCON) Director.

ESCD console. The ESCON Director display and keyboard device used to perform operator and service tasks at the ESCD.

ESCON. See *Enterprise System Connection.*

ESCON channel. A channel having an Enterprise Systems Connection channel-to-control-unit I/O interface that uses optical cables as a transmission medium. May operate in CBY, CNC, CTC or CVC mode. Contrast with *parallel channel.*

ESCON Director. An I/O interface switch that provides the interconnection capability of multiple ESCON interfaces (or FICON Bridge (FCV) mode - 9032-5) in a distributed-star topology.

ESCON Multiple Image Facility (EMIF). In the ESA/390 architecture and zSeries z/Architecture, a function that allows LPARs to share an ESCON and FICON channel path (and other channel types) by providing each LPAR with its own channel-subsystem image.

Extended Link Services (ELS). An Extended Link Service (command) request solicits a destination port (N_Port or F_Port) to perform a function or service. Each ELS request consists of an Link Service (LS) command; the N_Port ELS commands are defined in the FC-FS architecture.

Exchange A group of sequences which share a unique identifier. All sequences within a given exchange use the same protocol. Frames from multiple sequences can be multiplexed to prevent a single exchange from consuming all the bandwidth. See also: Sequence

F_Node Fabric Node - a fabric attached node.

F_Port Fabric Port - a port used to attach a Node Port (N_Port) to a switch fabric.

Fabric Fibre Channel employs a fabric to connect devices. A fabric can be as simple as a single cable connecting two devices. The term is most often used to describe a more complex network utilizing hubs, switches and gateways.

Fabric Login Fabric Login (FLOGI) is used by an N_Port to determine if a fabric is present and, if so, to initiate a session with the fabric by exchanging service parameters with the fabric. Fabric Login is performed by an N_Port following link initialization and before communication with other N_Ports is attempted.

FC. (1) (Fibre Channel), a short form when referring to something that is part of the fibre channel standard. (2) Also used by the IBM I/O definition process when defining a FICON channel (using IOCP of HCD) that will be used in FICON native mode (using the FC-SB-2 communication protocol).

FC-FS. Fibre Channel-Framing and Signalling, the term used to describe the FC-FS architecture.

FC Fibre Channel

FC-0 Lowest level of the Fibre Channel Physical standard, covering the physical characteristics of the interface and media

FC-1 Middle level of the Fibre Channel Physical standard, defining the 8b/10b encoding/decoding and transmission protocol.

FC-2 Highest level of the Fibre Channel Physical standard, defining the rules for signaling protocol and describing transfer of frame, sequence and exchanges.

FC-3 The hierarchical level in the Fibre Channel standard that provides common services such as striping definition.

FC-4 The hierarchical level in the Fibre Channel standard that specifies the mapping of upper-layer protocols to levels below.

FCA Fibre Channel Association.

FC-AL Fibre Channel Arbitrated Loop - A reference to the Fibre Channel Arbitrated Loop standard, a shared gigabit media for up to 127 nodes, one of which may be attached to a switch fabric. See also: Arbitrated Loop.

FC-CT Fibre Channel common transport protocol

FC-FG Fibre Channel Fabric Generic - A reference to the document (ANSI X3.289-1996) which defines the concepts,

behavior and characteristics of the Fibre Channel Fabric along with suggested partitioning of the 24-bit address space to facilitate the routing of frames.

FC-FP Fibre Channel HIPPI Framing Protocol - A reference to the document (ANSI X3.254-1994) defining how the HIPPI framing protocol is transported via the Fibre Channel

FC-GS Fibre Channel Generic Services -A reference to the document (ANSI X3.289-1996) describing a common transport protocol used to communicate with the server functions, a full X500 based directory service, mapping of the Simple Network Management Protocol (SNMP) directly to the Fibre Channel, a time server and an alias server.

FC-LE Fibre Channel Link Encapsulation - A reference to the document (ANSI X3.287-1996) which defines how IEEE 802.2 Logical Link Control (LLC) information is transported via the Fibre Channel.

FC-PH A reference to the Fibre Channel Physical and Signaling standard ANSI X3.230, containing the definition of the three lower levels (FC-0, FC-1, and FC-2) of the Fibre Channel.

FC-PLDA Fibre Channel Private Loop Direct Attach - See PLDA.

FC-SB Fibre Channel Single Byte Command Code Set - A reference to the document (ANSI X.271-1996) which defines how the ESCON command set protocol is transported using the Fibre Channel.

FC-SW Fibre Channel Switch Fabric - A reference to the ANSI standard under development that further defines the fabric behavior described in FC-FG and defines the communications between different fabric elements required for those elements to coordinate their operations and management address assignment.

FC Storage Director See SAN Storage Director

FCA Fibre Channel Association - a Fibre Channel industry association that works to promote awareness and understanding of the Fibre Channel technology and its application and provides a means for implementers to support the standards committee activities.

FCLC Fibre Channel Loop Association - an independent working group of the Fibre Channel Association focused on the marketing aspects of the Fibre Channel Loop technology.

FCP Fibre Channel Protocol - the mapping of SCSI-3 operations to Fibre Channel.

FCS. See *fibre channel standard*.

fiber. See *optical fiber*.

fiber optic cable. See *optical cable.*

fiber optics. The branch of optical technology concerned with the transmission of radiant power through fibers made of transparent materials such as glass, fused silica, and plastic.

Note: Telecommunication applications of fiber optics use optical fibers. Either a single discrete fiber or a non-spatially aligned fiber bundle can be used for each information channel. Such fibers are often called “optical fibers” to differentiate them from fibers used in non-communication applications.

Fibre Channel A technology for transmitting data between computer devices at a data rate of up to 4 Gbps. It is especially suited for connecting computer servers to shared

storage devices and for interconnecting storage controllers and drives.

fibre channel standard. An ANSI standard for a computer peripheral interface. The I/O interface defines a protocol for communication over a serial interface that configures attached units to a communication fabric. The protocol has four layers. The lower of the four layers defines the physical media and interface, the upper of the four layers defines one or more Upper Layer Protocols (ULP)—for example, FCP for SCSI command protocols and FC-SB-2 for FICON protocol supported by ESA/390 and z/Architecture. Refer to ANSI X3.230.1999x.

FICON. (1) An ESA/390 and zSeries computer peripheral interface. The I/O interface uses ESA/390 and zSeries FICON protocols (FC-FS and FC-SB-2) over a Fibre Channel serial interface that configures attached units to a FICON supported Fibre Channel communication fabric. (2) An FC4 proposed standard that defines an effective mechanism for the export of the SBCCS-2 (FC-SB-2) command protocol via fibre channels.

FICON channel. A channel having a Fibre Channel connection (FICON) channel-to-control-unit I/O interface that uses optical cables as a transmission medium. May operate in either FC or FCV mode.

FICON Director. A Fibre Channel switch that supports the ESCON-like “control unit port” (CUP function) that is assigned a 24-bit FC port address to allow FC-SB-2 addressing of the CUP function to perform command and data transfer (in the FC world, it is a means of in-band management using a FC-4 ULP).

field replaceable unit (FRU). An assembly that is replaced in its entirety when any one of its required components fails.

FL_Port Fabric Loop Port - the access point of the fabric for physically connecting the user's Node Loop Port (NL_Port).

FLOGI See Fabric Log In

Frame A linear set of transmitted bits that define the basic transport unit. The frame is the most basic element of a message in Fibre Channel communications, consisting of a 24-byte header and zero to 2112 bytes of data. See also: Sequence

FRU. See *field replaceable unit*.

FSP Fibre Channel Service Protocol - The common FC-4 level protocol for all services, transparent to the fabric type or topology.

FSPF Fabric Shortest Path First - is an intelligent path selection and routing standard and is part of the Fibre Channel Protocol.

Full-Duplex A mode of communications allowing simultaneous transmission and reception of frames.

G_Port Generic Port - a generic switch port that is either a Fabric Port (F_Port) or an Expansion Port (E_Port). The function is automatically determined during login.

Gateway A node on a network that interconnects two otherwise incompatible networks.

Gb/s Gigabits per second. Also sometimes referred to as Gbps. In computing terms it is approximately 1,000,000,000 bits per second. Most precisely it is 1,073,741,824 (1024 x 1024 x 1024) bits per second.

GB/s Gigabytes per second. Also sometimes referred to as GBps. In computing terms it is approximately 1,000,000,000 bytes per

second. Most precisely it is 1,073,741,824 (1024 x 1024 x 1024) bytes per second.

GBIC GigaBit Interface Converter - Industry standard transceivers for connection of Fibre Channel nodes to arbitrated loop hubs and fabric switches.

Gigabit One billion bits, or one thousand megabits.

GLM Gigabit Link Module - a generic Fibre Channel transceiver unit that integrates the key functions necessary for installation of a Fibre channel media interface on most systems.

half duplex. In data communication, pertaining to transmission in only one direction at a time. Contrast with *duplex.*

hard disk drive. (1) A storage media within a storage server used to maintain information that the storage server requires. (2) A mass storage medium for computers that is typically available as a fixed disk or a removable cartridge.

Hardware The mechanical, magnetic and electronic components of a system, e.g., computers, telephone switches, terminals and the like.

HBA Host Bus Adapter

HCD. Hardware Configuration Dialog.

HDA. Head and disk assembly.

HDD. See *hard disk drive.*

head and disk assembly. The portion of an HDD associated with the medium and the read/write head.

HIPPI High Performance Parallel Interface - An ANSI standard defining a channel that transfers data between CPUs and from a CPU to disk arrays and other peripherals.

HMMP HyperMedia Management Protocol

HMMS HyperMedia Management Schema - the definition of an implementation-independent, extensible, common data description/schema allowing data from a variety of sources to be described and accessed in real time regardless of the source of the data. See also: WEBM, HMMP

hop A FC frame may travel from a switch to a director, a switch to a switch, or director to a director which, in this case, is one hop.

HSM Hierarchical Storage Management - A software and hardware system that moves files from disk to slower, less expensive storage media based on rules and observation of file activity. Modern HSM systems move files from magnetic disk to optical disk to magnetic tape.

HUB A Fibre Channel device that connects nodes into a logical loop by using a physical star topology. Hubs will automatically recognize an active node and insert the node into the loop. A node that fails or is powered off is automatically removed from the loop.

HUB Topology see Loop Topology

Hunt Group A set of associated Node Ports (N_Ports) attached to a single node, assigned a special identifier that allows any frames containing this identifier to be routed to any available Node Port (N_Port) in the set.

ID. See *identifier.*

identifier. A unique name or address that identifies things such as programs, devices or systems.

In-band Signaling This is signaling that is carried in the same channel as the information. Also referred to as in-band.

In-band virtualization An implementation in which the virtualization process takes place in the data path between servers and disk systems. The virtualization can be implemented as software running on servers or in dedicated engines.

Information Unit A unit of information defined by an FC-4 mapping. Information Units are transferred as a Fibre Channel Sequence.

initial program load (IPL). (1) The initialization procedure that causes an operating system to commence operation. (2) The process by which a configuration image is loaded into storage at the beginning of a work day, or after a system malfunction. (3) The process of loading system programs and preparing a system to run jobs.

input/output (I/O). (1) Pertaining to a device whose parts can perform an input process and an output process at the same time. (2) Pertaining to a functional unit or channel involved in an input process, output process, or both, concurrently or not, and to the data involved in such a process. (3) Pertaining to input, output, or both.

input/output configuration data set (IOCDS). The data set in the S/390 and zSeries processor (in the support element) that contains an I/O configuration definition built by the input/output configuration program (IOCP).

input/output configuration program (IOCP). A S/390 program that defines to a system the channels, I/O devices, paths to the I/O devices, and the addresses of the I/O devices.The output is normally written to a S/390 or zSeries IOCDS.

interface. (1) A shared boundary between two functional units, defined by functional characteristics, signal characteristics, or other characteristics as appropriate. The concept includes the specification of the connection of two devices having different functions. (2) Hardware, software, or both, that links systems, programs, or devices.

Intermix A mode of service defined by Fibre Channel that reserves the full Fibre Channel bandwidth for a dedicated Class 1 connection, but also allows connection-less Class 2 traffic to share the link if the bandwidth is available.

Inter switch link A FC connection between switches and/or directors. Also known as ISL.

I/O. See *input/output*.

I/O configuration. The collection of channel paths, control units, and I/O devices that attaches to the processor. This may also include channel switches (for example, an ESCON Director).

IOCDS. See *Input/Output configuration data set*.

IOCP. See *Input/Output configuration control program.*

IODF. The data set that contains the S/390 or zSeries I/O configuration definition file produced during the defining of the S/390 or zSeries I/O configuration by HCD. Used as a source for IPL, IOCP and Dynamic I/O Reconfiguration.

IPL. See *initial program load*.

I/O Input/output

IP Internet Protocol

IPI Intelligent Peripheral Interface

ISL See Inter switch link.

Isochronous Transmission Data transmission which supports network-wide timing requirements. A typical application for isochronous transmission is a broadcast environment which needs information to be delivered at a predictable time.

JBOD Just a bunch of disks.

Jukebox A device that holds multiple optical disks and one or more disk drives, and can swap disks in and out of the drive as needed.

jumper cable. In an ESCON and FICON environment, an optical cable having two conductors that provides physical attachment between a channel and a distribution panel or an ESCON/FICON Director port or a control unit/device, or between an ESCON/FICON Director port and a distribution panel or a control unit/device, or between a control unit/device and a distribution panel. Contrast with *trunk cable.*

laser. A device that produces optical radiation using a population inversion to provide *light amplification by stimulated emission of radiation* and (generally) an optical resonant cavity to provide positive feedback. Laser radiation can be highly coherent temporally, or spatially, or both.

L_Port Loop Port - A node or fabric port capable of performing Arbitrated Loop functions and protocols. NL_Ports and FL_Ports are loop-capable ports.

LAN - A network covering a relatively small geographic area (usually not larger than a floor or small building). Transmissions within a Local Area Network are mostly digital, carrying data among stations at rates usually above one megabit/s.

Latency A measurement of the time it takes to send a frame between two locations.

LC Lucent Connector. A registered trademark of Lucent Technologies.

LCU. *See Logical Control Unit.*

LED. See *light emitting diode*.

licensed internal code (LIC). Microcode that IBM does not sell as part of a machine, but instead, licenses it to the customer. LIC is implemented in a part of storage that is not addressable by user programs. Some IBM products use it to implement functions as an alternate to hard-wire circuitry.

light-emitting diode (LED). A semiconductor chip that gives off visible or infrared light when activated. Contrast *Laser*.

link. (1) In an ESCON environment or FICON environment (fibre channel environment), the physical connection and transmission medium used between an optical transmitter and an optical receiver. A link consists of two conductors, one used for sending and the other for receiving, thereby providing a duplex communication path. (2) In an ESCON I/O interface, the physical connection and transmission medium used between a channel and a control unit, a channel and an ESCD, a control unit and an ESCD, or, at times, between two ESCDs. (3) In a FICON I/O interface, the physical connection and transmission medium used between a channel and a control unit, a channel and a FICON Director, a control unit and a fibre channel FICON Director, or, at times, between two fibre channels switches.

link address. (1) On an ESCON interface, the portion of a source or destination address in a frame that ESCON uses to route a frame through an ESCON director. ESCON

associates the link address with a specific switch port that is on the ESCON director. See also *port address.* (2) On a FICON interface, the port address (1-byte link address), or domain and port address (2-byte link address) portion of a source (S_ID) or destination address (D_ID) in a fibre channel frame that the fibre channel switch uses to route a frame through a fibre channel switch or fibre channel switch fabric. See also *port address.*

Link_Control_Facility A termination card that handles the logical and physical control of the Fibre Channel link for each mode of use.

LIP A Loop Initialization Primitive sequence is a special Fibre Channel sequence that is used to start loop initialization. Allows ports to establish their port addresses.

local area network (LAN). A computer network located in a user's premises within a limited geographic area.

logical control unit (LCU). A separately addressable control unit function within a physical control unit. Usually a physical control unit that supports several LCUs. For ESCON, the maximum number of LCUs that can be in a control unit (and addressed from the same ESCON fiber link) is 16; they are addressed from x'0' to x'F'. For FICON architecture, the maximum number of LCUs that can be in a control unit (and addressed from the same FICON fibre link) is 256; they are addressed from x'00' to x'FF'. For both ESCON and FICON, the actual number supported, and the LCU address value, is both processor- and control unit implementation-dependent.

logical partition (LPAR). A set of functions that create a programming environment that is defined by the ESA/390 architecture or zSeries z/Architecture. ESA/390 architecture or zSeries z/Architecture uses the term LPAR when more than one logical partition is established on a processor. An LPAR is conceptually similar to a virtual machine environment except that the LPAR is a function of the processor. Also, LPAR does not depend on an operating system to create the virtual machine environment.

logical switch number (LSN). A two-digit number used by the I/O Configuration Program (IOCP) to identify a specific ESCON or FICON Director. (This number is separate from the director's "switch device number" and, for FICON, it is separate from the director's "FC switch address").

l**ogically partitioned (LPAR) mode**. A central processor mode, available on the Configuration frame when using the PR/SM facility, that allows an operator to allocate processor hardware resources among logical partitions. Contrast with *basic mode.*

Login Server Entity within the Fibre Channel fabric that receives and responds to login requests.

Loop Circuit A temporary point-to-point like path that allows bi-directional communications between loop-capable ports.

Loop Topology An interconnection structure in which each point has physical links to two neighbors resulting in a closed circuit. In a loop topology, the available bandwidth is shared.

LPAR. See *logical partition*.

LVD Low Voltage Differential

Management Agent A process that exchanges a managed node's information with a management station.

Managed Node A managed node is a computer, a storage system, a gateway, a

media device such as a switch or hub, a control instrument, a software product such as an operating system or an accounting package, or a machine on a factory floor, such as a robot.

Managed Object A variable of a managed node. This variable contains one piece of information about the node. Each node can have several objects.

Management Station A host system that runs the management software.

MAR Media Access Rules. Enable systems to self-configure themselves is a SAN environment

Mb/s Megabits per second. Also sometimes referred to as Mbps. In computing terms it is approximately 1,000,000 bits per second. Most precisely it is 1,048,576 (1024 x 1024) bits per second.

MB/s Megabytes per second. Also sometimes referred to as MBps. In computing terms it is approximately 1,000,000 bytes per second. Most precisely it is 1,048,576 (1024 x 1024) bytes per second.

Metadata server In Storage Tank, servers that maintain information ("metadata") about the data files and grant permission for application servers to communicate directly with disk systems.

Meter 39.37 inches, or just slightly larger than a yard (36 inches)

Media Plural of medium. The physical environment through which transmission signals pass. Common media include copper and fiber optic cable.

Media Access Rules (MAR).

MIA Media Interface Adapter - MIAs enable optic-based adapters to interface to copper-based devices, including adapters, hubs, and switches.

MIB Management Information Block - A formal description of a set of network objects that can be managed using the Simple Network Management Protocol (SNMP). The format of the MIB is defined as part of SNMP and is a hierarchical structure of information relevant to a specific device, defined in object oriented terminology as a collection of objects, relations, and operations among objects.

Mirroring The process of writing data to two separate physical devices simultaneously.

MM Multi-Mode - See Multi-Mode Fiber

MMF See Multi-Mode Fiber - - In optical fiber technology, an optical fiber that is designed to carry multiple light rays or modes concurrently, each at a slightly different reflection angle within the optical core. Multi-Mode fiber transmission is used for relatively short distances because the modes tend to disperse over longer distances. See also: Single-Mode Fiber, SMF

Multicast Sending a copy of the same transmission from a single source device to multiple destination devices on a fabric. This includes sending to all N_Ports on a fabric (broadcast) or to only a subset of the N_Ports on a fabric (multicast).

Multi-Mode Fiber (MMF) In optical fiber technology, an optical fiber that is designed to carry multiple light rays or modes concurrently, each at a slightly different reflection angle within the optical core. Multi-Mode fiber transmission is used for relatively short distances because the modes tend to disperse over longer distances. See also: Single-Mode Fiber

Multiplex The ability to intersperse data from multiple sources and destinations onto a single transmission medium. Refers to delivering a single transmission to multiple destination Node Ports (N_Ports).

N_Port Node Port - A Fibre Channel-defined hardware entity at the end of a link which provides the mechanisms necessary to transport information units to or from another node.

N_Port Login N_Port Login (PLOGI) allows two N_Ports to establish a session and exchange identities and service parameters. It is performed following completion of the fabric login process and prior to the FC-4 level operations with the destination port. N_Port Login may be either explicit or implicit.

Name Server Provides translation from a given node name to one or more associated N_Port identifiers.

NAS Network Attached Storage - a term used to describe a technology where an integrated storage system is attached to a messaging network that uses common communications protocols, such as TCP/IP.

ND. See *node descriptor*.

NDMP Network Data Management Protocol

NED. See *node-element descriptor.*

Network An aggregation of interconnected nodes, workstations, file servers, and/or peripherals, with its own protocol that supports interaction.

Network Topology Physical arrangement of nodes and interconnecting communications links in networks based on application requirements and geographical distribution of users.

NFS Network File System - A distributed file system in UNIX developed by Sun Microsystems which allows a set of computers to cooperatively access each other's files in a transparent manner.

NL_Port Node Loop Port - a node port that supports Arbitrated Loop devices.

NMS Network Management System - A system responsible for managing at least part of a network. NMSs communicate with agents to help keep track of network statistics and resources.

Node An entity with one or more N_Ports or NL_Ports.

node descriptor. In an ESCON and FICON environment, a node descriptor (ND) is a 32-byte field that describes a node, channel, ESCON Director port or a FICON Director port, or a control unit.

node-element descriptor. In an ESCON and FICON environment, a node-element descriptor (NED) is a 32-byte field that describes a node element, such as a disk (DASD) device.

Non-Blocking A term used to indicate that the capabilities of a switch are such that the total number of available transmission paths is equal to the number of ports. Therefore, all ports can have simultaneous access through the switch.

Non-L_Port A Node or Fabric port that is not capable of performing the Arbitrated Loop functions and protocols. N_Ports and F_Ports are not loop-capable ports.

OEMI. See *original equipment manufacturers information.*

open system. A system whose characteristics comply with standards made available throughout the industry and that therefore can be connected to other systems complying with the same standards.

Operation A term defined in FC-2 that refers to one of the Fibre Channel *building blocks* composed of one or more, possibly concurrent, exchanges.

optical cable. A fiber, multiple fibers, or a fiber bundle in a structure built to meet optical, mechanical, and environmental specifications. See also *jumper cable*, *optical cable assembly*, and *trunk cable.*

optical cable assembly. An optical cable that is connector-terminated. Generally, an optical cable that has been connector-terminated by a manufacturer and is ready for installation. See also *jumper cable* and *optical cable*.

optical fiber. Any filament made of dialectic materials that guides light, regardless of its ability to send signals. See also *fiber optics* and *optical waveguide*.

optical fiber connector. A hardware component that transfers optical power between two optical fibers or bundles and is designed to be repeatedly connected and disconnected.

optical waveguide. (1) A structure capable of guiding optical power. (2) In optical communications, generally a fiber designed to transmit optical signals. See *optical fiber.*

Ordered Set A Fibre Channel term referring to four 10 -bit characters (a combination of data and special characters) providing low-level link functions, such as frame demarcation and signaling between two ends of a link.

original equipment manufacturer information (OEMI). A reference to an IBM guideline for a computer peripheral interface. More specifically, it refers to IBM S/360 and S/370 Channel to Control Unit Original Equipment Manufacturer Information. The interface uses ESA/390 logical protocols over an I/O interface that configures attached units in a multi-drop bus environment. This OEMI interface is also supported by the zSeries 900 processors.

Originator A Fibre Channel term referring to the initiating device.

Out of Band Signaling This is signaling that is separated from the channel carrying the information. Also referred to as out-of-band.

Out-of-band virtualization An alternative type of virtualization in which servers communicate directly with disk systems under control of a virtualization function that is not involved in the data transfer.

parallel channel. A channel having a System/360 and System/370 channel-to-control-unit I/O interface that uses bus and tag cables as a transmission medium. Contrast with *ESCON channel*.

path. In a channel or communication network, any route between any two nodes. For ESCON and FICON this would be the route between the channel and the control unit/device, or sometimes from the operating system control block for the device and the device itself.

path group. The ESA/390 and zSeries architecture (z/Architecture) term for a set of

channel paths that are defined to a controller as being associated with a single S/390 image. The channel paths are in a group state and are on-line to the host.

path-group identifier. The ESA/390 and zSeries architecture (z/Architecture) term for the identifier that uniquely identifies a given LPAR. The path-group identifier is used in communication between the system image program and a device. The identifier associates the path-group with one or more channel paths, thereby defining these paths to the control unit as being associated with the same system image.

Peripheral Any computer device that is not part of the essential computer (the processor, memory and data paths) but is situated relatively close by. A near synonym is input/output (I/O) device.

Petard A device that is small and sometimes explosive.

PLDA Private Loop Direct Attach - A technical report which defines a subset of the relevant standards suitable for the operation of peripheral devices such as disks and tapes on a private loop.

PCICC. (IBM) PCI Cryptographic Coprocessor.

PLOGI See N_Port Login

Point-to-Point Topology An interconnection structure in which each point has physical links to only one neighbor resulting in a closed circuit. In point-to-point topology, the available bandwidth is dedicated.

Policy-based management Management of data on the basis of business policies (for example, "all production database data must be backed up every day"), rather than technological considerations (for example, "all data stored on this disk system is protected by remote copy").

port. (1) An access point for data entry or exit. (2) A receptacle on a device to which a cable for another device is attached. (3) See also *duplex receptacle*.

port address. (1) In an ESCON Director, an address used to specify port connectivity parameters and to assign link addresses for attached channels and control units. See also *link address.* (2) In a FICON director or Fibre Channel switch, it is the middle 8 bits of the full 24-bit FC port address. This field is also referred to as the “area field” in the 24-bit FC port address. See also *link address.*

Port Bypass Circuit A circuit used in hubs and disk enclosures to automatically open or close the loop to add or remove nodes on the loop.

port card. In an ESCON and FICON environment, a field-replaceable hardware component that provides the optomechanical attachment method for jumper cables and performs specific device-dependent logic functions.

port name. In an ESCON or FICON Director, a user-defined symbolic name of 24 characters or less that identifies a particular port.

Private NL_Port An NL_Port which does not attempt login with the fabric and only communicates with other NL Ports on the same loop.

processor complex. A system configuration that consists of all the machines required for operation; for example, a processor unit, a processor controller, a system display, a

service support display, and a power and coolant distribution unit.

program temporary fix (PTF). A temporary solution or bypass of a problem diagnosed by IBM in a current unaltered release of a program.

prohibited. In an ESCON or FICON Director, the attribute that, when set, removes dynamic connectivity capability. Contrast with *allowed.*

protocol. (1) A set of semantic and syntactic rules that determines the behavior of functional units in achieving communication. (2) In fibre channel, the meanings of and the sequencing rules for requests and responses used for managing the switch or switch fabric, transferring data, and synchronizing the states of fibre channel fabric components. (3) A specification for the format and relative timing of information exchanged between communicating parties.

PTF. See *program temporary fix*.

Public NL_Port An NL_Port that attempts login with the fabric and can observe the rules of either public or private loop behavior. A public NL_Port may communicate with both private and public NL_Ports.

Quality of Service (QoS) A set of communications characteristics required by an application. Each QoS defines a specific transmission priority, level of route reliability, and security level.

Quick Loop is a unique fibre-channel topology that combines arbitrated loop and fabric topologies. It is an optional licensed product that allows arbitrated loops with private devices to be attached to a fabric.

RAID Redundant Array of Inexpensive or Independent Disks. A method of configuring multiple disk drives in a storage subsystem for high availability and high performance.

Raid 0 Level 0 RAID support - Striping, no redundancy

Raid 1 Level 1 RAID support - mirroring, complete redundancy

Raid 5 Level 5 RAID support, Striping with parity

Repeater A device that receives a signal on an electromagnetic or optical transmission medium, amplifies the signal, and then retransmits it along the next leg of the medium.

Responder A Fibre Channel term referring to the answering device.

route. The path that an ESCON frame takes from a channel through an ESCD to a control unit/device.

Router (1) A device that can decide which of several paths network traffic will follow based on some optimal metric. Routers forward packets from one network to another based on network-layer information. (2) A dedicated computer hardware and/or software package which manages the connection between two or more networks. See also: Bridge, Bridge/Router

SAF-TE SCSI Accessed Fault-Tolerant Enclosures

SAN A Storage Area Network (SAN) is a dedicated, centrally managed, secure information infrastructure, which enables any-to-any interconnection of servers and storage systems.

SAN System Area Network - term originally used to describe a particular symmetric

multiprocessing (SMP) architecture in which a switched interconnect is used in place of a shared bus. Server Area Network - refers to a switched interconnect between multiple SMPs.

SANSymphony In-band block-level virtualization software made by DataCore Software Corporation and resold by IBM.

saved configuration. In an ESCON or FICON Director environment, a stored set of connectivity attributes whose values determine a configuration that can be used to replace all or part of the ESCD's or FICON's active configuration. Contrast with *active configuration*.

SC Connector A fiber optic connector standardized by ANSI TIA/EIA-568A for use in structured wiring installations.

Scalability The ability of a computer application or product (hardware or software) to continue to function well as it (or its context) is changed in size or volume. For example, the ability to retain performance levels when adding additional processors, memory and/or storage.

SCSI Small Computer System Interface - A set of evolving ANSI standard electronic interfaces that allow personal computers to communicate with peripheral hardware such as disk drives, tape drives, CD_ROM drives, printers and scanners faster and more flexibly than previous interfaces. The table below identifies the major characteristics of the different SCSI version.

SCSI Version	Signal Rate MHz	BusWidth (bits)	Max. DTR (MBps)	Max. Num. Devices	Max. Cable Length (m)
SCSI-1	5	8	5	7	6
SCSI-2	5	8	5	7	6
Wide SCSI-2	5	16	10	15	6
Fast SCSI-2	10	8	10	7	6
Fast Wide SCSI-2	10	16	20	15	6
Ultra SCSI	20	8	20	7	1.5
Ultra SCSI-2	20	16	40	7	12
Ultra2 LVD SCSI	40	16	80	15	12

SCSI-3 SCSI-3 consists of a set of primary commands and additional specialized command sets to meet the needs of specific device types. The SCSI-3 command sets are used not only for the SCSI-3 parallel interface but for additional parallel and serial protocols, including Fibre Channel, Serial Bus Protocol (used with IEEE 1394 Firewire physical protocol) and the Serial Storage Protocol (SSP).

SCSI-FCP The term used to refer to the ANSI Fibre Channel Protocol for SCSI document (X3.269-199x) that describes the FC-4 protocol mappings and the definition of how the SCSI protocol and command set are transported using a Fibre Channel interface.

Sequence A series of frames strung together in numbered order which can be transmitted over a Fibre Channel connection as a single operation. See also: Exchange

service element (SE). A dedicated service processing unit used to service a S/390 machine (processor).

SERDES Serializer Deserializer

Server A computer which is dedicated to one task.

SES SCSI Enclosure Services - ANSI SCSI-3 proposal that defines a command set for soliciting basic device status (temperature, fan speed, power supply status, etc.) from a storage enclosures.

Single-Mode Fiber In optical fiber technology, an optical fiber that is designed for the transmission of a single ray or mode of light as a carrier. It is a single light path used for long-distance signal transmission. See also: Multi-Mode Fiber

Small Computer System Interface (SCSI). (1) An ANSI standard for a logical interface to computer peripherals and for a computer peripheral interface. The interface uses an SCSI logical protocol over an I/O interface that configures attached targets and initiators in a multi-drop bus topology. (2) A standard hardware interface that enables a variety of peripheral devices to communicate with one another.

SMART Self Monitoring and Reporting Technology

SM Single Mode - See Single-Mode Fiber

SMF Single-Mode Fiber - In optical fiber technology, an optical fiber that is designed for the transmission of a single ray or mode of light as a carrier. It is a single light path used for long-distance signal transmission. See also: MMF

SNIA Storage Networking Industry Association. A non-profit organization comprised of more than 77 companies and individuals in the storage industry.

SN Storage Network. See also: SAN

SNMP Simple Network Management Protocol - The Internet network management protocol which provides a means to monitor and set network configuration and run-time parameters.

SNMWG Storage Network Management Working Group is chartered to identify, define and support open standards needed to address the increased management requirements imposed by storage area network environments.

SSA Serial Storage Architecture - A high speed serial loop-based interface developed as a high speed point-to-point connection for peripherals, particularly high speed storage arrays, RAID and CD-ROM storage by IBM.

Star The physical configuration used with hubs in which each user is connected by communications links radiating out of a central hub that handles all communications.

Storage Tank An IBM file aggregation project that enables a pool of storage, and even individual files, to be shared by servers of different types. In this way, Storage Tank can greatly improve storage utilization and enables data sharing.

StorWatch Expert These are StorWatch applications that employ a 3 tiered architecture that includes a management interface, a StorWatch manager and agents that run on the storage resource(s) being managed. Expert products employ a StorWatch data base that can be used for saving key management data (e.g. capacity or performance metrics). Expert products use the agents as well as analysis of storage data saved in the data base to perform higher value functions including -- reporting of capacity, performance, etc. over time (trends),

configuration of multiple devices based on policies, monitoring of capacity and performance, automated responses to events or conditions, and storage related data mining.

StorWatch Specialist A StorWatch interface for managing an individual fibre Channel device or a limited number of like devices (that can be viewed as a single group). StorWatch specialists typically provide simple, point-in-time management functions such as configuration, reporting on asset and status information, simple device and event monitoring, and perhaps some service utilities.

Striping A method for achieving higher bandwidth using multiple N_Ports in parallel to transmit a single information unit across multiple levels.

STP Shielded Twisted Pair

Storage Media The physical device itself, onto which data is recorded. Magnetic tape, optical disks, floppy disks are all storage media.

subchannel. A logical function of a channel subsystem associated with the management of a single device.

subsystem. (1) A secondary or subordinate system, or programming support, usually capable of operating independently of or asynchronously with a controlling system.

SWCH. In ESCON Manager, the mnemonic used to represent an ESCON Director.

Switch A component with multiple entry/exit points (ports) that provides dynamic connection between any two of these points.

Switch Topology An interconnection structure in which any entry point can be dynamically connected to any exit point. In a switch topology, the available bandwidth is scalable.

T11 A technical committee of the National Committee for Information Technology Standards, titled T11 I/O Interfaces. It is tasked with developing standards for moving data in and out of computers.

Tape Backup Making magnetic tape copies of hard disk and optical disc files for disaster recovery.

Tape Pooling A SAN solution in which tape resources are pooled and shared across multiple hosts rather than being dedicated to a specific host.

TCP Transmission Control Protocol - a reliable, full duplex, connection-oriented end-to-end transport protocol running on top of IP.

TCP/IP Transmission Control Protocol/ Internet Protocol - a set of communications protocols that support peer-to-peer connectivity functions for both local and wide area networks.

Time Server A Fibre Channel-defined service function that allows for the management of all timers used within a Fibre Channel system.

Topology An interconnection scheme that allows multiple Fibre Channel ports to communicate. For example, point-to-point, Arbitrated Loop, and switched fabric are all Fibre Channel topologies.

T_Port An ISL port more commonly known as an E_Port, referred to as a Trunk port and used by INRANGE.

TL_Port A private to public bridging of switches or directors, referred to as Translative Loop.

trunk cable. In an ESCON and FICON environment, a cable consisting of multiple fiber pairs that do not directly attach to an active device. This cable usually exists between distribution panels (or sometimes between a set processor channels and a distribution panel) and can be located within, or external to, a building. Contrast with *jumper cable*.

Twinax A transmission media (cable) consisting of two insulated central conducting leads of coaxial cable.

Twisted Pair A transmission media (cable) consisting of two insulated copper wires twisted around each other to reduce the induction (thus interference) from one wire to another. The twists, or lays, are varied in length to reduce the potential for signal interference between pairs. Several sets of twisted pair wires may be enclosed in a single cable. This is the most common type of transmission media.

ULP Upper Level Protocols

unblocked. In an ESCON and FICON Director, the attribute that, when set, establishes communication capability for a specific port. Contrast with *blocked.*

unit address. The ESA/390 and zSeries term for the address associated with a device on a given controller. On ESCON and FICON interfaces, the unit address is the same as the device address. On OEMI interfaces, the unit address specifies a controller and device pair on the interface.

UTC Under-The-Covers, a term used to characterize a subsystem in which a small number of hard drives are mounted inside a higher function unit. The power and cooling are obtained from the system unit. Connection is by parallel copper ribbon cable or pluggable backplane, using IDE or SCSI protocols.

UTP Unshielded Twisted Pair

Virtual Circuit A unidirectional path between two communicating N_Ports that permits fractional bandwidth.

Virtualization An abstraction of storage where the representation of a storage unit to the operating system and applications on a server is divorced from the actual physical storage where the information is contained.

Virtualization engine Dedicated hardware and software that is used to implement virtualization.

WAN Wide Area Network - A network which encompasses inter-connectivity between devices over a wide geographic area. A wide area network may be privately owned or rented, but the term usually connotes the inclusion of public (shared) networks.

WDM Wave Division Multiplexing - A technology that puts data from different sources together on an optical fiber, with each signal carried on its own separate light wavelength. Using WDM, up to 80 (and theoretically more) separate wavelengths or channels of data can be multiplexed into a stream of light transmitted on a single optical fiber.

WEBM Web-Based Enterprise Management - A consortium working on the development of a series of standards to enable active management and monitoring of network-based elements.

Zoning In Fibre Channel environments, the grouping together of multiple ports to form a virtual private storage network. Ports that are members of a group or zone can communicate

with each other but are isolated from ports in other zones.

z/Architecture. An IBM architecture for mainframe computers and peripherals. Processors that follow this architecture include the zSeries family of processors.

zSeries. A family of IBM mainframe servers that support high performance, availability, connectivity, security and integrity.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

IBM Redbooks

- *Introduction to SAN Distance Solutions*, SG24-6408
- *Introducing Hosts to the SAN fabric*, SG24-6411
- *Implementing an Open IBM SAN,* SG24-6116
- *Implementing the Cisco MDS 9000 in an Intermix FCP, FCIP, and FICON Environment*, SG24-6397
- *Introduction to Storage Area Networks*, SG24-5470
- *IP Storage Networking: IBM NAS and iSCSI Solutions*, SG24-6240
- *The IBM TotalStorage NAS 200 and 300 Integration Guide*, SG24-6505
- *Implementing the IBM TotalStorage NAS 300G: High Speed Cross Platform Storage and Tivoli SANergy!*, SG24-6278
- *iSCSI Performance Testing & Tuning*, SG24-6531
- *Using iSCSI Solutions' Planning and Implementation*, SG24-6291
- *IBM Storage Solutions for Server Consolidation*, SG24-5355
- *Implementing the Enterprise Storage Server in Your Environment*, SG24-5420
- *Implementing Linux with IBM Disk Storage*, SG24-6261
- *Storage Area Networks: Tape Future In Fabrics*, SG24-5474
- *IBM Enterprise Storage Server*, SG24-5465

Other resources

These publications are also relevant as further information sources:

- *Building Storage Networks*, ISBN 0072120509

These IBM publications are also relevant as further information sources:

© Copyright IBM Corp. 2004. All rights reserved.

Referenced Web sites

These Web sites are also relevant as further information sources:

- IBM TotalStorage hardware, software and solutions:

  `http://www.storage.ibm.com`

- IBM TotalStorage Storage Networking:

  `http://www.storage.ibm.com/snetwork/index.html`

- Brocade:

  `http://www.brocade.com`

- CNT:

  `http://www.inrange.com`

- McDATA:

  `http://www.mcdata.com`

- QLogic:

  `http://www.qlogic.com`

- Emulex:

  `http://www.emulex.com`

- Finisar:

  `http://www.finisar.co`

- Veritas:

  `http://www.veritas.co`

- Vixel:

  `http://www.vixel.com`

- Tivoli:

  `http://www.tivoli.co`

- JNI:

  `http://www.Jni.com`

- IEEE:

  `http://www.ieee.org`

- Storage Networking Industry Association:

  `http://www.snia.org`

- Fibre Channel Industry Association:

  `http://www.fibrechannel.com`

- SCSI Trade Association:

  `http://www.scsita.org`

- Internet Engineering Task Force:

  `http://www.ietf.org`

- American National Standards Institute:

  `http://www.ansi.org`

- Technical Committee T10:

  `http://www.t10.org`

- Technical Committee T11:

  `http://www.t11.org`

- IBM @server™ xSeries® 430 and NUMA-Q Information Center:

  `http://webdocs.numaq.ibm.com`

How to get IBM Redbooks

You can order hardcopy Redbooks, as well as view, download, or search for Redbooks at the following Web site:

ibm.com/redbooks

You can also download additional materials (code samples or diskette/CD-ROM images) from that site.

IBM Redbooks collections

Redbooks are also available on CD-ROMs. Click the CD-ROMs button on the Redbooks Web site for information about all the CD-ROMs offered, as well as updates and formats.

Index

Numerics

128-port fabric 299
1394b 23
1x9 Transceivers 19, 27
2032-001 446
2042-001 366
2042-128 367
2042-256 368
2042-C40 366, 386
2062-D01 490–491
2062-D07 494–495
2062-T07 494, 496
2109-F16 291
2109-F32 291
2109-M12 294
2109-S08 287
2109-S16 287
24 bit addressing 77
24-bit addressing 74, 77, 419
24-bit port address 55, 79
32064
    H__h3
        4.3.5 Quick Loop 134
6227 256
64-bit address 74
8B/10B 103
8b/10b 116
8b/10b code 419
8b/10b encoder 80
8-bit addressing 419–420

A

AAA 527
absolute port number 298
AC module 443
ACC 86
acceptance test 285
access 503, 505, 525–527, 541, 543–547
access control 541, 547
Access Control List 332
Access Control Lists 308, 310
access controls 308
access fairness mechanism 51, 126
accurate performance data 584, 608, 632, 660
ACL 308–310, 332
activate 524, 535
active backplane 38
active clients 398
active connections 571
active CP 297
active supervisor 494, 496, 504
active supervisor module 541
active zone set 531, 545
active zoneset 549
adapter cable 28
adding a switch 1001
adding ISL's 1002
Address Resolution Protocol 542
address translation 338
addressing scheme 74
adds 194
adjacent ISLs 313
adjacent ports 414
Admin 529, 543
administrator 527, 529, 543
administrator access 399
Advanced Performance Monitoring 306, 321
advanced SAN features 331
AFS 167
Agent 493, 505
aggregate 65, 514
aggregate bandwidth 505–506, 515–516
air 378
air flow 373
airflow 444
AL 83
AL_PA 54, 86, 130, 332, 337–338, 419
    priority 55, 131
AL_PA monitoring 321
alerts 474, 519
Alias 332
alias 546
aliases 336
American National Standards Institute 99
amplify 193
analyze switch fabric configuration 531
analyzing end-to-end connectivity 530

© Copyright IBM Corp. 2004. All rights reserved.

ANSI 17, 138
APD 192
API 137, 540
application availability 570
application programming interface 540
Application Programming Interfaces 137
application specific integrated circuit 436
application type 244
application-specific integrated circuit 18
ARB 130
arbitrated loop 48, 176, 373, 447, 501, 627
Arbitrated Loop Physical Address 54, 130
arbitration 50, 54, 129–130
arbitration protocol 50
area 78–79, 83
ARP 542
AS 489–491, 493, 495–497, 500–503, 505, 507–508, 515–517, 519–520, 523–525, 527–528, 530–535, 537–540, 542–543, 545–549
AS/400 168, 698, 742
ASIC 18, 263, 280, 297, 379, 385, 436
asynchronous transfer mode 209
ATM 209
attached 493, 500–501, 508, 519–520, 522, 530, 544, 546–547
attenuation 39
attenuation meters 39
audit log 475
audit trail 222, 402
authentication process 309
authentication, authorization, and accounting 527
authority 399
authorized personnel 484
authorized switches 309
auto mode 500
auto-detecting 468
automated scripts 476
automatic call home 392
automatic failover 568
auto-negotiating 431, 433
auto-sensing 431, 433, 491, 494–495, 500
autosensing 505
availability 440, 448, 568, 588, 742, 746, 750, 755, 764, 770, 777, 782, 787, 791, 797, 808, 815, 819, 825, 828, 833, 842, 849, 856, 860, 865, 869, 875
availability criteria 331
availability solutions 636
available addresses 79
avalanche photodiode 192
avoid downtime 1000

B

B_Port 627
backbones 435
backplane 18, 37, 374, 379, 385, 441, 443, 448, 496, 727, 761, 806, 808, 818, 839
backplane module 379, 385
backup 326, 520
backup copy 401
backup CTP2 442, 449
backup FRU 448
backup SBAR 444
backup window 165
balanced 266
balancing 65
bandwidth 31, 57, 73, 113, 184, 189, 260, 691, 820
bandwidth requirements 719
bandwidth utilization 65, 514
BB_Credit 99, 461, 617
beaconing 433, 435, 452
benchmark 137
BER 115
binding 257, 310
bit error rate 115
blade 761, 839
blade-based architecture 379
bladed architecture 294
blades 37, 296, 803
blocked port 481
blocking 45–46, 252, 454, 481
bottleneck 114, 575, 580
bound 308, 484
bridge 37
bridging 47, 1035
broadcast 57, 523
broadcast frames 480
broadcast messages 410
broadcast transfer 334
broadcast zone 334, 644
broadcast zoning 410
broadcast-storms 330
Brocade 89
Brocade SilkWorm 288
browser sessions 263
Buffer Credits 411
buffer credits 282, 316, 357, 373, 381, 415, 503, 505–506, 736, 798, 834

buffers 316, 482
buffer-to-buffer 536
building blocks 567
bus 28, 542
bus arbitration 176
business continuance 568
business continuity 5
business recovery 4
byte-encoding scheme 116

C

cabinet 222, 386
cabinet key 222
cabinet protection 232
cabinets 469
cable connections 457
cable distance limitations 702
cable identification tag 229
cable lengths 469
cable management 222
cable protection 233
cable routing 213
cable supports 222
cable tag naming standard 229
cable ties 222
cable types 213
cables 281
cabling blueprint 413
Call home 493, 511, 519
call home 271
call home feature 457
call-home 469
campus 250
canvas 322
cascade 411
cascaded directors 410
cascades 32
cascading 58, 65, 465, 498, 514
cascading switches 575
CDR 116
central arbiter 503
central zoning 320
centralized management 15, 691
certificate exchange 311
certified 283
change record 229
change zones 1010
changes 500, 546, 548
channel expansion 191
channel extender 186
  WAN extension 901, 930, 960, 990
chassis 405, 494–495
circuit board 37
circuit breaker 450
Cisco 489–491, 494, 496, 506, 510, 514, 522, 526, 539, 550–551
Cisco Fabric Manager 508, 540–541, 543–544
Cisco IOS CLI 540
Cisco MDS 9000 490, 493, 498, 501–502, 505–506, 515–519, 523, 527–528, 530–531, 537, 540–544, 546–549
Cisco MDS 9216 Multilayer Fabric Switch 539
Cisco MDS 9506 Multilayer Director 539
Cisco MDS 9509 Multilayer Director 498, 539
Cisco MDS9000 499
cladding 216
Class F 280
CLI 476, 492, 505, 527, 539–540, 543–544
Client 511–512, 527
clock 115–116, 496
clock and data recovery 116
clock module 496
clocking circuitry 119
cloning 326
closed-ring topology 56
closing a loop circuit 56
cluster 13
clustering software 570
clustering solution 618
clustering solutions 594, 672
clusters 616
coarse wave division multiplexer 188
coating 216
color coding schemes 213
combined solution 814, 823
Comma characters 118
command 527, 531, 535
command line interface 476, 531, 540
common interface model 148
common services 103
communication 493, 501, 520, 527
communication channels 269
CompactFlash 492, 505
complex design 567
complex switched fabric 35
complexity 270
compression 186

concepts 43
configuration 489, 500–501, 505, 514, 520, 523–525, 527, 529–531, 534, 536, 540, 542–543, 549, 1009
configuration changes 529
configuration option 397
configuration options 503
configure 492, 500, 505–506, 528–529, 534–536, 543
configuring zones 478
congestion 115, 252, 517–518, 734, 1010
connection 257
connection authentication 173
connectivity 430, 496, 505, 530–531, 541, 575
consolidate 271
continuous alarm 324
control 492, 494–495, 503, 517–519, 541, 547
control blades 372
convert 37
copy 502, 532
copy services 695
core 215, 411, 1000
core design 250
core switch 33
core switch design 726, 738
core switches 32, 578, 723, 730
core—edge 724, 730–731, 739
core-edge design 578
core—edge design 723
core-to-edge 430
cost 489, 505–506, 509, 539
cost function 461–462
coupling switches 47
CP cards 296–297
credit starvation 461
cross point chip 374
crossbar 379
crossbar switches 506
crossbar switching fabric 504
cryptographic authentication 311
CTP 436
CTP Card 448
CTP2 card 441
CUP 458, 466
cut-through logic 35
CWDM 550–551

D

d balancing 571
D_ID 123
daemon 324
dark 222
dark fiber 219
data bandwidth 693
data communications fiber 213
data encryption 319
data exchange 331
data integrity 135, 724
data migration 14
data movements 247
data objects 247
data replication 696, 735, 824
data sharing 10
data traffic 306, 321, 525, 532, 536
data transfer rate 63, 114
data transmission range 315
database 499, 527, 543
database synchronization 64, 92
Datacenter Server 2000 598, 677
DB-9 24
DC-balance 103
deactivated 525, 545
debug 510
decibel 201
decrypt 172
decrypts 309
Dedicated Simplex 100
Defacto Standards 138
default 500, 517–518, 523–524, 527, 529, 543, 549
default values 543
default VSAN 523–524
default zone 479, 545–548
default zone policy 546
defined 523–526
degraded performance 63, 114
delay factor 56
Dell 706
demultiplexers 192
Dense Wave Division Multiplexer 189
Dense Wave Division Multiplexers 183
design 250, 744, 748, 752, 760, 766, 773, 780, 785, 790, 794, 802, 810, 817, 821, 826, 830, 838, 845, 852, 858, 863, 867, 871
destination ID 35, 123
destination port address identifier (D_ID) 80
detail bar 401

device 493, 499, 501, 514, 520, 530–531, 536–537, 539–540, 545, 547, 549
Device Connection Controls 308
device level zoning 114
device restrictions 457
device sharing 453
Device View 540
device WWN 410
DF_CTL 123
DFB lasers 192
diagnostic 530
diagnostics 445, 489, 493, 510
dial home 224, 397
DID 515, 518
digital certificate 172, 310, 319
digital certificates 309
digital data 209
director 34, 493–495, 497, 503, 520, 522, 526, 548–549
director class 494, 496, 833
director class product 262
directors 34, 490–491, 498, 503, 505–506, 530
disable 518, 544, 548
disaster planning 1
disaster recovery 2, 245, 687, 689, 691
disaster tolerance 14, 568
disciplines 211
discovery method 80
disk resource sharing 691
dispersed networks 184
disruptive 494, 496, 504, 549
distance 281, 447, 736, 743, 747, 751, 756, 764, 771, 778, 783, 788, 792, 798, 809, 816, 819, 825, 829, 834, 843, 850, 857, 861, 865, 870, 876
distance limitations 212
distance option 439, 447
distribute fabrics 315
distributing traffic 515
documentation 230, 1009
domain 78–79, 83, 518, 549
domain ID 328, 405, 455
domain ID conflicts 456
domain IDs 280, 315, 548–549
Domain Manager 520, 522, 549
domain number routing 35
domains 308
downstream 52, 128
downtime 2, 34, 687, 811
drops 194
dual connections 790, 868
dual connectivity 311
dual director 808
dual directors 453–454
dual fabric 311, 592, 802, 828
dual fabrics 613, 636, 665
dual fabrics over distance 640, 668
dual LAN connections 470
dual pathing software 656
dual redundant supervisor modules 494–495
dual SAN fabrics 666
dual site solution 615, 668
dual switches 299
DWDM 183, 189, 713, 736, 746–747, 754, 756–757, 778, 787–788, 795, 798, 800, 822, 825, 831, 833–835, 857, 864, 866, 873, 876

E

E_D_TOV 60, 315, 482, 548, 617
E_Port 18, 35, 46–47, 70, 280, 370, 411, 418, 436, 439, 447, 500, 516, 523
E_Port mode 500
E_Port segmentation 456
E_Ports 482, 500–501, 534, 548–549, 736, 798
ED-6064 Director management software 466
edge 1000
edge switch 730
edge switches 32, 270, 739
education 234
EFC Manager 472, 766, 773, 780, 784, 789, 794, 800
EFC Server 467, 469
EISL 501, 508, 515–516, 536
element 323
elements 323, 542
e-mail 519
embedded web server 263
Emulex 256
enable 489, 499, 501, 510, 518–519, 527–528, 541, 544
encoder 116
encrypt 172
encrypted 309, 528
encrypted tunnel 174
encryption 172
encrypts 308, 601
End-of-Frame 120
end-of-frame 55, 80, 131

End-to-End monitoring 321
end-to-end visibility 306, 321
end-user scenarios 685
enforcement 333
enhanced graphical user interface 322
enterprise manager 323
entry switch 434
EOF 55, 80, 120, 131
ERP 2
error 548
error detect time-out value 60
error detection 432, 434, 437, 445, 452
error messages 324
errors 493, 505
ES-1000 436, 627
ES-3016 436
ES-3032 436
ESCON 166, 210, 220, 645, 746, 787, 864
ethernet 492–493, 499, 505, 507, 509–510, 519, 541–542
ethernet cables 222
ethernet network 264
event log 475
Event View 320
events 392, 519–520
exchange 121
existing resources 247
expandable fabric 47
expansion 717
expansion port 500–501
extended distances 315, 482
Extended Fabric 307
Extended ISL 501
extended link 482
extending device 186
Extensible Core Architecture 372

F

F_BSY 60
F_CTL 123
F_Port 18, 46, 370, 436, 439, 447, 500–501
F_Ports 79, 500, 534
fabric 248, 489–493, 499–501, 503–505, 517–520, 523–524, 539, 542, 545–548
fabric aware 421
fabric behavior 330
Fabric Binding 514
Fabric Configuration Servers 308
fabric design 574
Fabric Event View 319
Fabric Login 81–82, 84
Fabric login 86
fabric login 85
fabric management 114, 318, 500, 540
Fabric Management Policy Set 308
Fabric Manager 307, 325, 508, 510–512, 530, 540–541, 543–544
fabric Name Server 80
Fabric planning 453
fabric port layout plan 229
Fabric pWWN 546
fabric services 520
Fabric Shortest Path First 88, 252, 454
Fabric Topology View 319
Fabric View 540
fabric-authorized switch ACL 310
Fabricenter 438
fabrics 509, 522, 540
fabric-wide resource 336
failover 259, 495–496, 519, 541, 569, 592
fairness algorithm 54, 130
Fan 491, 493–494, 496
fan assembly 384
fan modules 378, 443
fan out 723, 732, 740, 742, 746, 749, 755, 763, 770, 776, 782, 787, 791, 797, 815, 819, 828, 833, 842, 849, 855, 860, 864, 869, 875
fan-in 253
fan-out 252, 579
fan-out ratio 453
fans 437, 491, 496, 530
fastest path 96
fastest route 92
FAStT200 287
fault tolerance 570
fault tolerant 761, 839
faults 1010
FC 5020 387
FC 5920 387
FC cabling 17
FC ports 497
FC/9000 Control Module Enhanced 376
FC_AL 373, 498
FC-0 102
FC-1 103
FC-2 102–103
FC-AL 50, 176, 270, 602

FCAP 310
FCC 517
FC-GS-3 418
FCIP 499, 507–508, 511, 519, 539, 550–551
FCIP tunnels 508
FC-LE 177
FCM 379, 383, 387
FCME 376, 396
FCP 498, 510
FC-PH 81, 83, 98, 100, 103
FC-PH-2 98, 100
FC-PH-3 101
FC-SW 57, 70, 251, 280
FC-SW-2 411
FC-SW2 70, 280, 283, 418
feature codes 508, 549
fiber optic interconnects 19
fiber-optic transceivers 21
Fibre Alliance 138
fibre cable routes 646
Fibre Channel 490–492, 494–495, 499–502, 507–510, 515, 517–520, 530, 534, 536–538, 541–542, 546, 548
  FC-0 102
  FC-1 103
  FC-2 103
  FC-3 103
  FC-4 104
  point-to-point 49
Fibre Channel analyzer 40, 502
Fibre Channel Arbitrated Loop 50
Fibre Channel Authentication Protocol 310
Fibre Channel cabling 17
Fibre Channel Control Module 383
Fibre Channel HBAs 249
Fibre Channel IDs 499
Fibre Channel Industry Association 138
Fibre Channel interface 502, 531
Fibre Channel protocol 542
Fibre Channel Shortest Path First 88
Fibre Channel Standards 43, 424
Fibre Channel Switch Module 374, 381
Fibre Channel Switched Fabric 57
Fibre Channel topologies 48
Fibre I/O Module 380
fibre network environment 183
fibre saving devices 184
FICON 210, 220, 370, 431, 498, 514, 539
FICON cascaded 411, 514
FICON cascading 59, 393
FICON Cascading Activation 393
FICON Intermix 465
FICON Management Server 466
FICON management server 466
FICON Manager Server 392
FICON support 465
field programmable gate arrays 379
field replaceable units 441
field-replaceable modules 24
Filler Module 373
filter 535
Filter-based monitoring 321
filtering 535
filters 535
FIO 372, 379–380, 387, 804, 819
FIO blade 383
Firewalls 174
firmware 35, 395, 491–492
firmware download 326
firmware levels 465
firmware version 421
FL_Port 18, 46, 53, 370, 500–501
FL_Ports 79, 501, 534
flash memory 493
FlashCopy 679
flexibility 489, 499, 519
flexible fabric switch 491, 493
FLOGI 86, 133
floor plan 212
flow control 517–518
flow statistics 461
FMPS 308
FMS 466
Forward Congestion Control 517
forwarding tables 461
FPGA 379
FPM cards 449
FPM numbering scheme 439, 447
Fractional Bandwidth 100
frame 122, 501, 510, 515, 517–518, 547
frame delimiter 120
Frame Filtering 322
frame filtering 114, 284
frame header 74
frame holdtime 60
frame limit 418
frame transmission 56
Frames 121

frames 40, 247, 500–501, 509–510, 516–519, 532, 536, 538, 542, 547
framing and signaling protocol 103
free port addresses 131
FRU 441
FRU beaconing 445
FSPF 88–89, 113, 252, 268, 311, 454, 460, 520, 522, 734, 778, 816, 857, 1010
FSW 374, 379, 381–382, 387
FSW module 383
full bandwidth 314
full duplex protocol 49
full mesh 576
fusion 218
future needs 248
FX_Port 501

G

G_Port 46, 436, 447
G_Ports 47, 436, 439, 442, 447
gateway 37, 508
GBIC 19, 381
general concepts 567
generic port 439, 447
geographically dispersed 244
Gigabaud Link Modules 26
Gigabit Ethernet 23, 209
Gigabit Interface Converters 19
Gigabit Link Modules 19, 26
gigabit transport 19
GLM 19, 26
global integration 325
globally capture 325
GO/NOGO testers 39
goal throughput 98
grant 527
graphical user interface 306, 319
gridlock 580
group of zones 479
groups 527
growth 717, 724, 727
GUI 306, 319
GUI tools 1007

H

hackers 4
HACMP 618, 701, 703
hard zoning 547, 736
hardware 17, 489, 494, 496, 507, 525, 547
hardware designs 284
hardware enforced 65, 519
hardware enforced zones 333
hardware enforcement 332
hardware zoning 106, 135
hardware-assisted authentication 334
HBA 28, 239, 256
HBAs 257
HCD 82, 84
health 307, 514, 530
heat damage 224
heavily loaded ISLs 461
heterogeneous 125, 249, 690
heterogeneous operating systems 331
heterogeneous solution 654
heterogeneous support 283
high availability 244, 370, 431, 494, 496, 504, 508, 515, 519, 717, 803
high availability option 387
high performance profile 585, 609, 633, 661
high priority status 517
High-Availability Option 366, 370–371, 376
higher throughput 183
highly available 694
highly available networks 881, 910, 941, 970
highly available SAN 644
high-speed switching 74
hold time value 315
Holdtime 315
holdtime 60
homogeneous 249
hop 251, 412, 724, 734, 736, 778, 857
hop count 315, 416
hop count cost 64, 92
hop counts 454
Hoppy 59, 416
hops 59–60, 89, 92, 315, 416
Host Bus Adapter 28
Host Bus Adapters 239, 256
host bus adapters 568
hot plugabble 271
hot-spare 383
hot-swappable 434
hot-swappable fan 494, 496
HSSDC 24
Hub Cascading 128
hub cascading 52, 128
human error 227

Hunt Groups 100
Hunt groups 103

I

I/O circuits 18
I/O Priority Queuing 14
IBM TotalStorage SAN Switch 879
identify 515
identity 74, 527
Idle 120
IEEE 75, 78
IETF 138
image pair 87
implicit transfer 523
in order frame delivery 68–69
Inband 465
in-band 541–542
inband 152, 765, 844
inband fabric service 85
inband management 465, 810
increased availability 23
independent fabrics 255
independent power sources 457
in-depth monitoring 320
industry associations 137
Information Unit 115
information units 121
Informix 690
infrastructure 281
initial machine load 441
initialization 500–501
initiator 31, 328
initiators 508
in-order delivery 314
input fibres 192
INRANGE fabric 405
INRANGE FC/9000 939
INRANGE FC/9000 Fibre Channel Director 801
INRANGE firmware 421
INRANGE hard zoning 649
INRANGE Name Server zoning 407
INRANGE zoning methods 406
installing 398
insurance brokerage company 697
intangible damages 4
Inter Switch Links 32, 251
interconnected using ISLs 481
interconnection topologies 48
interface 28
interface module 492
interoperability 136, 339, 418, 458, 498
Interoperability Labs 285
interoperability matrix 424, 549
interoperability mode 548–549
interrupt 251
intersite cable route 221
Inter-Switch Links 411
inter-switch links 59, 311, 454
inventory 1000
investment protection 489
IN-Vision Enterprise Manager 390
IN-VSN 387, 390, 645
IN-VSN data base 401
IN-VSN Enterprise Manager Console 390
IN-VSN Enterprise Manager Server 391
IN-VSN event log 404
IN-VSN installation 397
IN-VSN management 390
IN-VSN Management Server 369
IN-VSN management software 650
IOCP 82, 84
IP address 541, 543–544
IP addresses 542
IP line card 508
IP services 519
IP storage services 494, 496
IPSec 174
IRIX 690
iSCSI 499, 507–508, 510, 519
iSeries 168
ISL 47, 59, 62, 112–113, 311, 439, 447, 482, 500–501, 515–516, 548, 719, 1010, 1035
ISL connections 505
ISL segmented 61
ISL statistical data-rate 462
ISL synchronization process 61
ISL traffic 671
ISL Trunking 313
islands of information 15
ISLs 32, 251, 411, 506, 512, 514–515, 695, 734
ISO 138
ISO/OSI 101
isolated 489, 523, 548
isolated VSAN 523–524
ISP 4
IU 115, 121

J

Java 514, 540
Java-capable 306, 319
JBODs 249
JFS 167
Jiro 138
jitter 115
jumper cable 216, 218
just-in-time storage 149

K

K28.5 118, 120
kangaroo 59, 416
key access 222
keys 172

L

L_Port 46
labelling 229
LAN 137
LAN addresses 457
LAN free backup 245
LAN free data movement 12
LAN interface 469
Laser safety 103
laser source 39
lasers 24, 192
latency 34, 50, 72, 96, 124, 221, 251, 517, 531, 713, 734, 742, 746, 749, 777, 787, 797, 815, 819, 855, 865, 875
layers 101
LC-SC 28
least-cost data transfer 456
legacy device 418
legacy equipment 37, 283
library 544, 549
license key 514
license key management 326
light 213
light loss 217
light spectrum 188
limit access 331, 406
limited addressing 176
limited distance 176
limits 134
link budget 201
link cost 63, 90
Link Reset 120
Link Reset Response 120
link service 86
link state change 64, 92
link states 92
link utilization 483
links 89
Linux 706, 828
LIP 51, 86–87, 127, 131, 251, 682
LIP's 51, 127
LISM 52, 127
load balance 259, 515
load balancing 454, 494, 496, 515, 524
load-balancing period 463
load-sharing power supply 443
location 540–541
log 529
log events 402
logic control 494–495
logic functions 18
logical consolidation 9
logical ISL 314
logical switches 296
Logical Unit Number 112, 136
logical unit number 111
login 86
long distance ISL 316
long wavelength 21
longwave 213
loop 248, 501
  private 53, 128
  public 53, 128
loop addressing 55
loop circuit 56
loop device 682
loop devices 499, 654
loop identifier 86
loop initialization 51, 127
Loop Initialization Master 52, 127
Loop Initialization Primitive 51, 127
loop protocols 51
loop solution 626, 682
loop switch 781, 859
loop switches 270
loop tenancy 54, 418
looplets 337
Loops 128
loops 53
loop-specific protocols 51
low bandwidth LAN 177

low performance server 596, 674
low-loss 25
LR 120
LRR 120–121
LUN 111–112, 136, 258, 511
LUN level zoning 114
LUN masking 111, 480, 619, 645, 683, 889, 918, 948, 978
LVM 713

M

MAC 541
MAC address 74
Magstar 14
maintenance port 433, 435, 437, 452
MAN 185, 221, 225
manage 491, 496, 527, 541–543
manage SAN fabrics 306, 319
management xlv, 319, 475, 492, 494, 496, 500, 503, 505, 510, 512, 514, 519, 524, 526–529, 540–544, 752, 757, 765, 772, 779, 784, 788, 793, 799, 810, 817, 820, 826, 830, 835, 844, 851, 858, 861, 866, 870, 877
Management Access Controls 308
management capabilities 725
management disciplines 261
Management Information Base 324
management information base 157
management network 808
management overhead 263
management software 398
managing 263
managing SANs 325
mappings 500
market development 138
mask 543
masking 112, 136
master 52, 127
McDATA 759, 837, 909, 969
McDATA ED-6064 764, 777, 842, 856
McDATA ES-1000 781, 859
McDATA ES-3016 746, 782, 787, 794, 860, 865, 871
McDATA ES-3032 770, 849
McDATA Intrepid 6064 Director 446
McDATA Intrepid 6140 438
McDATA Open Fabric 458
McDATA Sphereon 3232 435
McDATA Sphereon 4500 fabric switch 431, 433
MDS 9000 489–490, 493, 498, 500–502, 505–506, 509–510, 512, 515–519, 523, 527–528, 530–531, 537, 540–544, 546–549
MDS 9216 490–493, 503, 505, 508, 550
MDS 9506 490, 494, 505, 508
MDS 9509 490, 495–497, 505, 526, 549
mechanical splices 218
Media Interface Adapters 19, 27
member 478
members 109
memory 109, 381, 493
memory arrays 18
menu bar 400
merging fabrics 481
merging zone information 481
mesh 411
meshed fabric 269
meshed network topology 576
meshed tree topologies 311
meshes 32
message integrity 319
Metropolitan Area Network 221
metropolitan area network 185
MIA 19, 27
MIB 153, 157, 324, 477
Microsoft Exchange 698
minimize-interleave 418
mirror ports 373
mirroring 135, 571
misbehaving 334
mixed fabrics 334
mixed fiber diameters 218
modal dispersion 115
Mode 500–501, 503, 505, 510, 531, 536, 541, 548–549
modem 505
modular chassis 506
modulation techniques 191
monitor 493, 505, 531, 535–539
monitor bidirectional traffic 538
monitor performance 322
monitored 532, 535
monitoring 496, 502, 504, 509, 531, 535, 538, 540
MTBF 34
Multicast 100, 104
multicast 69, 523
multilayer network 489
multimedia 177

multi-mode 213
multipath software 260
multipathing software 258, 266
Multiple Allegiance 14
multiple IN-VSN clients 398
multiple paths 259, 568–569
multiple VSANs 517, 520, 522
multiple zones 335
multiplexers 187, 192
multiplexing 187
multi-stage switch interconnect 592
mux/demux 193

N

N+1 redundancy 365
N_Port 46, 55, 79
N_Ports 18, 499–500
name server 419, 525, 531, 547, 549
name server enforcement 334
name server table 77, 407
Name Server View 319
name server zoning 652
Name Server zoning rules 407
names 547
naming conventions 227, 330
naming standards 223, 227
native FICON 166
navigation tree 401
negative disparity 118
Netware 690
network traffic 502
neutral disparity 118, 419
new zone set 547
nickname 332
nicknames 410, 478
NL_Port 46, 55, 79
NL_Ports 499–501
noise 195
non heterogeneous 125
non resilient 254
non-blocking 267, 414
non-blocking architecture 366
non-blocking device 647
non-broadcast 56
non-disruptive upgrades 263, 727
non-disruptively 546, 549
non-high-availability option 387
non-interleaved option 417
non-intrusive 502
non-structured cables 216
non-trunking ports 523
non-volatile memory 436, 442
NOS 120
Not Operational 120
notification thresholds 323
NT 828
NT servers 690
numbering scheme 440
NVRAM 384

O

OADM 193
object grouping 325
object management platform 325
OEM/IBM mixes 284
OFC 103, 125
off-campus 219
Offline state 120
OLS 120
Open Fiber Control 125
Open Fibre Control 103
Open Standards Fabric Initiative 138
Open Systems Adapters 220
Open Systems Management Server 465
Open Trunking 460
open-init 51, 127
operating platforms 249
operational state 524
Operator 529
optical add/drop multiplexer 193
optical amplifiers 192, 195
optical bits 96
optical connectors 25
optical filters 191
optical pulses 195
optical signals 188
Optical Time Domain Reflectometer 40
optimize 280
optoelectronics 192
Ordered Set 120
Ordered Sets 103
ordered sets 51, 127
orphan zone 407, 409
OS/390 466
OSA 220
OSFI 138

OSMS 465
OTDR 40
outband 152, 465
outband management 467
out-of-band 541
out-of-boundary values 307
overlap zones 332, 1010
overlapping zones 478
oversubscribed 94, 483
oversubscribing ISLs 65
over-subscription 252, 267, 269
oversubscription 94, 115, 454, 579–580, 584, 608, 632, 660
OX_ID 123

P

pager alerts 392
pair of E_Ports 60
Parallel Access Volumes 14
parallel busses 37
parallel electrical signals 28
parallel ISLs 313
Parallel Sysplex 167
Parameter 124
parameters 493, 501, 505, 524, 535
partial mesh 576
partition 281, 328
passive backplane 38
passive component 385
password control 326
passwords 231, 391, 527, 543
patch panels 217
path 518–520, 524, 528, 531, 548
Path and dual-redundant HBA 569
path failover 569
path selection protocol 89
paths 90
pathway 213
PAV 14
payload 96
PCI 249
Peer to Peer Remote Copy 645
Peer-to-Peer Remote Copy 14
performance 251, 440, 448, 483, 489–490, 505–506, 508–509, 540, 742, 746, 749, 755, 763, 770, 776, 782, 787, 791, 797, 808, 815, 819, 824, 828, 833, 842, 849, 855, 860, 864, 869, 875
performance graphs 322
performance management 512
performance monitoring 514, 1002
Performance View 320
Peripheral Component Interconnect 249
persistent 499
phantom 133
phantom address 133
phantom mode 338
phone line 224
phone line bottlenecks 224
phone sockets 224
photo detector 192
physical access 484
physical consolidation 8
physical dimensions 386
Physical Hardware 100
physical interface and media 102
Physical layer 102
physical location 211
physical port location 298
physical specifications 498
pilot test 285
PIN 192
PKI 311
planned growth 688
platform independence 146
PLDA 337
PLOGI 85, 87, 133
point-to-point 48–49
point-to-point DWDM solution 893, 922, 952, 982
point-to-point DWDM with ESS PPRC 897, 926, 956, 986
policies 308
policy based management 336
pooling 149
port 78–79
port address 78
Port addressing 499
Port Analyzer Adapter 509–510
port area numbering 298
port binding 484
port blades 264, 728, 762, 840
port capacity 34
port cards 386
port connectivity 248
port count 249, 439
port density 506
port expansion 267
port failure 437, 478

port formula 580
port groups 490, 506
Port ID 546
port layout plan 229
port level zoning 114
Port login 87
port maintenance 445
port modes 47, 499–500
port numbering 298
port numbers 479
port placement 414, 1003
port tuning options 417
port types 370, 500, 502
port usage 719
Port VSAN membership 524
Port World Wide Node Name 546
port-based zoning 107
PortChannel 500, 515, 534–535
Port-Channeling 671
Port-Channels 495, 497, 514
ports 490–492, 494–497, 499–500, 502–503, 505–508, 515, 522–523, 526, 530–532, 534–536, 543–544, 546, 687
ports pool 282
positional map 51, 127
positive disparity 118
positive-intrinsic-negative 192
power 550–551
power down 231
power module assembly 450
power on Sequence 231
power outlets 223
power supplies 224, 434, 436, 493–494, 496, 498, 530
power supply 493
power supply assemblies 377, 384
power up 422
powering the switches 1013
PPRC 616, 645
Preferred Domain ID 458
preferred domain ID 456
preferred domain IDs 482
pre-planning activity 211
prevent unauthorized access 680
primitive sequence 120
Primitive Signal 130
primitive signals 120
principal switch 455, 482, 1013
principal switch selection 455
priority 55, 131, 517
private device support 418
private FC-AL 742
Private Fibre Channel devices 131
private host 134
private hosts 603
private initiator 421
private key 172, 309
Private Loop 498, 501
private loop 53, 55, 79, 128, 130, 419, 602
private loop migration 338
private-loop servers 337
private-loop storage 337
privileges 543
PRLI 87
probe 502
problem determination 474
problems 530
Process login 87
process login 87
profile 326
proof of concept 285
propagate zones 1010
propagation delay 56
proprietary management solutions 146
protecting data 652, 678
protection 489, 495, 497, 520
protocol 247
protocol conversion 249
protocol definition 209
protocol level zoning 114
public arbitrated loop 501
public device support 418
public host 603
public key 172, 308
Public Key Infrastructure 310
public loop 53, 55, 79, 129–130, 419–420, 501
public target 421
public/private key pair 310
push-pull 25
pWWN 546

Q

QoS 517
quality of service 517
quench control 518
quench message 518
queues 517

queuing logic 18
QuickLoop 134, 320, 332, 337, 602–603, 742

R

R_A_TOV 60, 315, 482, 548, 617
R_CTL 123
R_RDY 71
RADIUS 174, 527, 530, 543, 679
RAID 249
range monitoring 323
ranges 323
read/write ratio 691
reboots 499
Receiver Ready 120
reconfigure WWN 478
recovery protection 401
Redbooks Web site 1041
  Contact us xlvii
Redhat Linux 706
redundancy 253, 441, 494, 496, 503, 520, 568, 592, 687, 717, 803, 822
redundant 248, 262, 694, 735, 824
redundant connections 618
redundant core 299
redundant fabric 594, 673
redundant fabrics 299
redundant fans 437
redundant hardware 615, 668
redundant paths 263
redundant power supplies 365, 491, 494, 496
redundant power supply 293
redundant SAN 688, 717, 741
redundant SAN design 723
redundant serial SBAR 451
reference clock 115
regenerative repeater 195
registered devices 66
Registered State Change Notifications 480
reliability 818
Remote Authentication Dial-In User Service 174
remote disk consolidation 881, 911, 941, 971
remote management 372
remote management station 263
Remote Switch 307
remote tape vaulting
  solution 905, 935, 965, 995
  with disaster tolerance 908, 937, 967, 997
remote workstations 483
repeaters 439, 447
replicate 696
reporting 432, 434, 437, 445, 452, 496, 514
Request Product Quotation 240
rerouting 461
rerouting decisions 462
rerouting selection 462
research department 705
resilience 65, 253, 514, 687, 811, 822
resilient 255
resource allocation time-out value 60
resource management 149
resources 519, 544
responder 31
restart 263
restore 166
restricted 548
restricting access 483
RFI shield 451
rights 527
ring DWDM solution 898, 927, 957, 987
rodent control measures 233
role-based authorization 529
roles 511, 527, 529, 543
round robin 93
round-robin 312
router 37
routes 314
routing algorithms 65
routing control 123
routing decision 35–36
routing information 92
routing logic 36, 74
routing path 112
routing tables 515
RPQ 240, 257
RS/6000 698
RSCN 86, 480
running configuration 524
running disparity 55, 80, 117, 131
RX_ID 124

S

S_ID 123
SAN appliance 185
SAN disciplines 211
SAN fabric cabinets 212
SAN Fabric View 319

SAN implementation 568
SAN inventory 212
SAN islands 251, 490, 645
SAN management 146
SAN server 185
SAN software management 145
SAN standards 137
SANergy 11
SANlets 250
SANmark 139
SANmark Qualified Program 139
SANpilot 475
SBAR 444, 451
SBAR assemblies 449
SBAR assembly 444
SC connectors 24
SC type 25
scalability 251, 743, 747, 751, 756, 764, 771, 778, 783, 788, 792, 798, 809, 816, 820, 825, 829, 834, 843, 850, 857, 861, 865, 870, 876
scale 251
scale easily 1000
scaling 1000
scheduled downtimes 688
screen layout 399
scripting 476
SCSI 184, 209, 499, 519
SCSI attached storage 283
SCSI disk arrays 37
SCSI limitations 176
SCSI target 111
SCSI Trade Association 138
SCSI-3 protocol 327
SCSITA 138
SD_Port 502, 536, 538
SDD 569, 593, 595, 598–599, 645, 676, 760, 838
SDH 209
secondary supervisor 504
secure 519, 527, 543
secure fabric 173
Secure Fabric Operating System 306
secure machine room 234
Secure Management Communications 308
secure sharing 329
Secure Shell 174, 527
Secure Sockets Layer 174, 319
secure solution 600, 622
secure switch-to-switch 309
security 105, 110, 135, 222, 231, 281, 329, 332, 391, 393, 396, 421, 479, 490, 507, 511, 514, 527, 539, 622, 649, 678, 723, 735, 743, 746, 750, 755, 764, 771, 777, 782, 787, 792, 798, 808, 815, 819, 825, 828, 833, 842, 850, 856, 860, 865, 869, 876
security architecture 308
security breaches 169
security features 483
security layers 169
security policy configuration 309
security solution 649, 678
segmentation 104
segmented 482
segregation 66, 519
self-configuring 381
self-negotiation 296
sending frames 71
separate sites 615
SEQ_CNT 123
SEQ_ID 123
sequence 121
sequences 121
SerDes 37
serial crossbar 451
serial crossbars 444
serial optical converters 20
serial port interface 264
serial signal 28
serializer/deserializer 37
Server 500, 511–514, 525–529, 531, 543–544, 547, 549
server authentication 319
server clustering 13
server free backup 245
server free data movement 12
server less backup 245
server to storage ratio 453
servers 240
server-to-storage ratio 453
service 469
serviceability 432, 434, 437, 445, 452
SES 327
SES switch management 327
setup 396
setup routine 542
setup validation 396
SFF 19–20
SFP 490–491, 494–495, 500, 503, 550–551
SFP transceivers 503
SFTP 510

SGI 690
shared bus 50
short wavelength 21
shortest path 96
shortwave 213
SID 515, 518
signal power 193
signaling layers 102
signaling protocol 103
Signalling layer 102
signalling layers 100
significant system events 445
silica glass 216
SilkWorm 2800 753
simple fabric 567
Simple Name Server 78
Simple Network Management Protocol 174
single director approach 805
single fabric 456
single point of failure 503, 612
single-console 325
single-level storage 168
single-mode 213
SLAP/FC-SW-3 310
slot number 499
slots 492, 494–497, 505
Small Computer System Interface 209
small form factor transceivers 19–20
SNIA 138–139
SNMP 323–324, 441, 477, 493, 505, 527, 540, 779, 793, 799, 858, 871, 877
SNMP agent 156, 233
SNMP trap 324
SNS 78, 87, 131, 133
SOF 120
soft port zoning 334
Soft Zone 74
soft zoning 231, 547, 652
software 792, 870
software capabilities 390
software enforced zones 335
software upgrade 494, 496
software zoning 109, 736, 751, 778, 798, 808, 816, 825, 828, 876
SONET 209
source ID 123
source interface types 534
source port address identifier (S_ID) 80
sources 247
space management 149
SPAN 502, 509–510, 515, 523, 531–532, 534–539
Span Destination 531
SPAN session 535
SPAN source 532
spare ports 437, 671
speed 98, 504, 536, 543
SSA 689, 691, 695
SSA devices 37
SSH 174, 527, 601
SSH Secure Shell 309
SSL 174, 319
stable monitoring mode 51, 127
standard solutions 285
standards 283
standards bodies 137
standby mode 297
standby supervisor 494–496, 541
Start-of-Frame 120
static balance 454
statistics 460
status 491, 493, 496, 517, 530, 543
status line 401
storage consolidation 244, 250, 702
storage intelligence 489
Storage Network Industry Association 138
Storage Networking Industry Association 139
storage pool 815
storage sizing 260
StorWatch 15
Striping 104
structured cables 217
structured cabling system 218
Subsystem Device Driver 569, 593, 595, 598–599, 645, 676, 760, 838
Summary View 540
supervisor 504, 510, 534
supervisor module 492, 494, 496, 503–504, 517, 534, 541–544
supervisor modules 493–497, 503, 530
support 469
supported attachments 369
supported HBAs 240, 257
switch access 320
switch authentication 173
switch cascading 58
Switch Connection Controls 308
switch count 315
switch design 716

switch fabric 531–534
switch frame buffering 36
switch health option 530
switch identification 327
switch interoperability 548
Switch Link Authentication Protocol 310
switch management 527
switch maximum
239 switches in fabric 59
7 hops allowed 59
switch port 502, 531, 543, 546
switch port addresses 78
switch port analyzer 502
switch port capability 47
switch ports 518
Switch Priority 458
switch technology 262
Switch View 320
switch views 320
switched fabric 48, 57, 251
switched fabric topology 31
switches 31
switchover 504
symbolic names 109
synchronize 396
Synchronous Digital Hierarchy 209
Synchronous Optical Network 209
synchronous TDM devices 187
syslog 505
System Automation 466
system availability 331
system bus 249
system outage 3
System/38 168

T

T_Port 47, 370, 418
T_Ports 411, 413, 645
T11 99, 176
Tachyon 417
tape attachments 270
tape consolidation 626
tape pooling 11
targets 247, 508
TCO 306, 321
TDM 183, 187–188
TE_Port 500–501, 516, 534–535
TE_Ports 501, 534, 549, 668
technical requirements 687, 692
technology goals 243
telnet 527–528, 531, 543
temperature 491, 493, 496, 498
temporary loop master 52, 127
TERM 516
terminology 43
test gear 39
test sites 285
tested 283
TFTP 421
The director contains two fan 450
threshold 531
threshold alerts 452
thresholds 307, 491, 496
throughput 99, 247, 687–688
throughput capability 414
tier 256
tier-layer design 576
time 514, 517, 519, 528, 531, 535, 540, 542, 544–545, 547–548
time slice 187
time-division multiplexing 183, 187
time-out values 482
title bar 400
Tivoli 12
Tivoli SAN Manager 15
Tivoli SANergy 709, 828
Tivoli Storage Manager 15, 709, 828
Tivoli Storage Network Manager 15, 800, 835
TL_Port 47, 370, 408, 420, 501
TL_Ports 500–501, 534
TLS 174
tool bar 401
topology 250, 508, 512, 520, 540
topology changes 66
topology database 64, 92
topology management 737
total cost of ownership 306, 321
traffic 494–496, 501–504, 508–509, 514–515, 517, 519–520, 522–525, 531–539, 546–547, 718, 734
traffic load balancing 460
traffic patterns 63, 114, 454
transceivers 19, 21, 27
transducer 96
translative 420
translative loop 418, 501
translative loop mode 420
translative mode 134, 338, 603

transmission protocol 103
transmission unit 122
transmission word 122
transparent failover 571
Transport Layer Security 174
transport protocols 499
traps 477
triggered alarm 324
Tri-rate SFPs 507
Trivial File Transfer Protocol 421
troubleshoot 530
troubleshooting 474
troubleshooting features 530
troubleshooting tips 553
troubleshooting tools 531
truncate modes 510
trunk cables 217
trunking 62, 516, 535, 543, 548, 734, 1002
trunking E_Port 501
trunking group 112, 289, 291
trunking ports 112, 516, 523–524
trust 310
trusted SAN switches 309
trusted switch 601
tunnels 316
two sites up to 100km apart 891, 920, 950, 980
two sites up to 10km apart 886, 915, 945, 975
two-node clustering with dual switch 594, 673

U

U_Port 46
ULP 104
ultraviolet 216
unauthorized access 231, 653
unicast 523
unidirectional 216
unique domain ID 482
unique domain IDs 482
Universal Port Modules 439
UNIX 700
UNIX Servers 167
unlimited distances 244
unused bandwidth 454
unzoned nodes 407
update mechanism 64, 92
Upgrade Path/Extensible Core Architecture 385
UPM card 442
upper layer protocol 104
upstream 52, 128
uptime 34
URL 549
user levels 399
users 514, 519–520, 527, 543
UV 216

V

velcro strips 222
vendor identifiers 75
virtual ISL connections 508
Virtual SAN 489, 519, 522, 666
Virtual Tape Server 14
virtualization 174, 185
VSAN 65–66, 489, 499, 511, 516–520, 523–526, 531–532, 534–536, 539, 543–544, 549, 666, 671
VSAN 4094 523–524
VSAN attributes 524
VSAN manager 524
VSAN trunking 501, 516
VSANs 62, 501, 510, 514, 516–520, 522–526, 529, 534–535, 540, 549
VTS 14
vulnerabilities 169

W

WAN 137, 185
wavelength 188
wavelength division multiplexing 183, 188
wavelengths 189
WDM 183, 188
web browser 544
Web Tools 292, 306, 319
Wide Area Networks 185
Windows 2000 688
Windows NT 698, 709
workload peaks 313
workloads 1002
World Wide Name 74, 131, 294
world wide name 75
world wide name zoning 114
World Wide Node Name 332
World Wide Port Name 332
world wide port name 76
World-Wide Name (WWN) 80
WWN 74–75, 294, 308, 336, 455–456, 478, 484, 499, 511, 523, 546, 889, 918, 948, 978
WWN zoning 410

WWNN 332
WWNs 1006
WWPN 75–76, 332, 548

X

X3T11 17
XCA 372, 385
XCA architecture 386
XFIO2 373

Z

Zip Drive 480
zone 104
Zone Admin 320
zone changes 546
zone configuration 332, 335–336
zone configurations 336
zone member 332
zone members 478, 545–546
zone naming standards 229
zone set 531, 545, 547
zone sets 479, 546–547
zoned 489, 544
zones 281, 511, 519, 523, 545–548, 743, 751
zoning 266, 328, 406, 420, 437, 508, 510–511, 520, 522–523, 525, 544–548, 724, 764, 771, 778, 792, 816, 828, 834, 843, 850, 856, 870
zoning administration 336
zoning by port number 479
zoning by WWN 478
zoning configuration 320
zoning definitions 525
zoning information 546
zoning process 330
zoning restrictions 547
zoning security solution 649

(1.5" spine)
1.5"<-> 1.998"
789 <->1051 pages

(1.0" spine)
0.875"<->1.498"
460 <-> 788 pages

(0.5" spine)
0.475"<->0.875"
250 <-> 459 pages

(0.2"spine)
0.17"<->0.473"
90<->249 pages

(0.1"spine)
0.1"<->0.169"
53<->89 pages

To determine the spine width of a book, you divide the paper PPI into the number of pages in the book. An example is a 250 page book using Plainfield opaque 50# smooth which has a PPI of 526. Divided 250 by 526 which equals a spine width of .4752". In this case, you would use the .5" spine. Now select the Spine width for the book and hide the others: **Special>Conditional Text>Show/Hide>SpineSize(-->Hide:)>Set** . Move the changed Conditional text settings to all files in your book by opening the book file with the spine.fm still open and **File>Import>Formats** the Conditional Text Settings (ONLY!) to the book files.

(2.5" spine)
2.5"<->nnn.n"
1315<-> nnnn pages

(2.0" spine)
2.0" <-> 2.498"
1052 <-> 1314 pages

Draft Document for Review December 28, 2004 6:01 pm

IBM TotalStorage: SAN Product, Design and Optimization Guide

Utilizing real-life case studies we show a selection of SAN designs

Understand DWDM and distance solutions

Introduces some SAN best practices

In this IBM Redbook, we revisit some of the core components and technologies that underpin a storage area network (SAN). We cover, to varying degrees, some of the latest additions to the IBM SAN portfolio, discuss general SAN design considerations, and build these considerations into a selection of real world case studies.

We have also consolidated material from other SAN redbooks to create a complete overview of the depth and breadth of the IBM TotalStorage SAN portfolio.

We realise that there are many ways to design a SAN and put all the components together. In our examples, we have incorporated the major considerations that need to be taken into account, but still left room to manoeuvre on the SAN field of play.

This redbook focuses on the SAN products that are generally considered to form the backbone of the SAN fabric today: switches and directors. With this backbone, developing it has prompted discrete approaches to the design of a SAN fabric. The bespoke vendor implementation of technology that is characteristic in the design footprint of switches and directors, means that we have an opportunity to answer challenges in different ways.

We will show examples where strength can be built in to our SAN using the network and the features of the components themselves. Our aim is to show that you can cut your SAN fabric according to your cloth.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-6384-00 ISBN