IBM

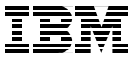
Designing and Optimizing an IBM Storage Area Network

- Using real life case studies, we show how to build a SAN
- Review SAN designs and their application
- Discover SAN best practices

Jon Tate
Gareth Coates
Ivo Gomilšek
Andy Lewis

Redbooks

ibm.com/redbooks



International Technical Support Organization

Designing and Optimizing an IBM Storage Area Network

May 2002

SG24-6419-00

Take Note! Before using this information and the product it supports, be sure to read the general information in “Notices” on page xv.

First Edition (May 2002)

This edition applies to the IBM SAN portfolio.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. QXXE Building 80-E2
650 Harry Road
San Jose, California 95120-6099

When you send information to IBM, you grant IBM a non-exclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Contents

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.

Trademarks

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

AIX®
AS/400®
CICS®
DB2®
DYNIX®
DYNIX/ptx®
Enterprise Storage Server™
ESCON®
eServer® 
FICON™
IBM®
Informix™
iSeries™
Magstar®
Netfinity®
NetView®
NUMA-Q®
OS/2®
OS/390®
OS/400®
Perform™
pSeries™
PTX®
Redbooks™
Redbooks(logo)™ 
RS/6000®
S/390®
SANergy™
Seascape®
SP™
SP2®
StorWatch™
System/390®
Tivoli®
Tivoli Enterprise™
Tivoli Enterprise Console®
TotalStorage™
Wave®
xSeries™
zSeries™

The following terms are trademarks of other companies:

ActionMedia, LANDesk, MMX, Pentium and ProShare 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.

C-bus is a trademark of Corollary, 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 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.

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.

In this photograph we show the team that wrote this redbook.

Jon Tate is a Project Leader for SAN TotalStorage 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 16 years of experience in storage software and management, services and support, and is an IBM SAN Certified Specialist.

Gareth Coates is a European Training Consultant for IBM Webserver Technical Support, Marketing, and Education, based in the UK. He has 15 years of experience in the open systems arena, including over 8 years of use and implementation of Linux, and 6 years focus on Fibre Channel. His areas of expertise include Fibre Channel technologies and fault finding at the frame level. He has extensively developed and delivered training on small, medium, and large SMP and NUMA systems. He has recently been involved in the development and rollout of training on AIX and logical partitioning for customers, business partners, and IBMers.

Ivo Gomilšek is an IT Specialist for Storage Area Networks, and storage and Linux for IBM Global Services, Slovenia, for the CEE region. His areas of expertise include Storage Area Networks (SAN), storage, IBM eServers xSeries servers, network operating systems (Linux, MS Windows, OS/2), and Lotus Domino servers. He holds certifications for IBM e-Server xSeries, and is a Red Hat Certified Engineer, Windows 2000 MSCE, and an OS/2 Warp Certified Engineer. Ivo was a member of the team that wrote the *IBM Redbook Designing*

an IBM Storage Area Network, SG24-5758, and contributed to various e-Server xSeries, and Linux integration guides. He also provides Level 2 support for SAN, IBM e-Server xSeries, and high availability solutions for IBM e-Server xSeries and Linux. Ivo has been with IBM for 5 years.

Andy Lewis is the Director of IT Architecture for PEAK Resources, Inc., a Premier IBM Business Partner specializing in storage infrastructure, pSeries, iSeries, and xSeries servers, based in Denver, Colorado. He has 15 years of experience in IT with a multifaceted background in mainframe, networking, and open systems. His areas of expertise include performance analysis, technical business reasoning, and systems architecture. Andy is one of PEAK's certified IBM and McDATA SAN specialists. Visit www.peakresources.com for further information.

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

Scott Drummond
IBM Storage Systems Group

Maritza Dubec
Emma Jacobs
Yvonne Lyon
Deanna Polm
Sokkieng Wang
International Technical Support Organization, San Jose Center

Mark Bruni
IBM Storage Systems Advanced Technical Support

Peter Thurston
Diana Tseng
Karen Ward
Ruoyi Zhou
IBM Storage Systems Group

Jim Baldyga
Chris Beauchamp
Tim Werts
Brocade Communications Systems

Dave Burchwell
Jack Consoli
Mike Naylor
INRANGE Technologies Corporation

Kirk Jenkins
Matt Martin
Brad Norton
Corin O'Connell
Jim Wild
McDATA Corporation

Notice

This publication is intended to help technical marketers, system designers and consultants evaluate and architect IBM SAN portfolio equipment. The information in this publication is not intended as the specification of any programming interfaces that are provided by any of the SAN hardware and software components contained herein. See the PUBLICATIONS section of the IBM Programming Announcement for the SAN hardware and software components contained herein for more information about what publications are considered to be product documentation.

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 the address on page ii.

Part 1

Back to basics

In this part of the book, we look at some of the basics of Fibre Channel, review the IBM SAN portfolio, and consider some of the business drivers that might lead you to a SAN deployment. Once this is done, we introduce some of the basic design considerations that apply in general to a SAN.

1

Identifying your business and technology goals

In this chapter, we discuss the various business and technological considerations for the IT Architect/Consultant who is designing the most suitable IBM Storage Area Network solution for their organization.

Consider this chapter as an outline of what makes good business sense. Various independent studies have shown that Storage Networking will be ranked by companies as one of the top ten investment areas in the next five years.

However, we must not think that we must implement the latest technology fad just because it is available. Rather, we must implement effective and efficient solutions that satisfy the business requirements while providing true economic benefits.

You may have different motivations for considering a SAN; it may be curiosity surrounding the hype, it may be that your CIO has come to an arrangement over a game of golf, or you may have a genuine business requirement that a SAN can satisfy. Either way, this section should help you in quantifying the business justification for your company.

If your business requirements include many, or even one of the following business issues (shown in Table 1-1) then the time you invest in reading this chapter will be well spent.

Table 1-1 Business issues

Disaster recovery
Resource sharing
Investment protection
Total Cost of Ownership (TCO)
Return on Investment (ROI)
Personnel costs
Service levels and application availability
Overcome cabling limitations
Application performance
Rapid deployment and ability to accommodate change
Realizing the true potential of storage consolidation
Solution confidence

A SAN may not provide the complete answer to satisfying your business requirement, but may be an intrinsic part of the total solution.

Important: Do not assume that the highest price solution is the better solution, which will satisfy your business goals.

1.1 Not another SAN versus NAS discussion

Many articles have been written that compare the benefits and drawbacks of a Storage Area Network solution versus a Network Attached Storage solution.

These redbooks are reference points for NAS, and iSCSI, comparisons, products and solutions:

- *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
- *Storage Networking Virtualization: What's it all about?*, SG24-6210

The intent of this section is not to re-state any of these arguments, but to review functionality with regards to business requirements to determine whether a SAN solution, NAS solution, or combination of both, provides the best fit for your company's needs.

So, with this in mind, we will continue this section with a view to SAN *and* NAS; and not SAN *versus* NAS.

1.1.1 SAN and NAS differentiating factors

The key differentiating factors of SAN and NAS are summarized in Table 1-2 (with the assumption that the SAN is using native Fibre Channel).

Table 1-2 High level SAN and NAS differentiators

Function	SAN	NAS
Optimized for SCSI (block) I/O	Yes	No, except iSCSI/iFCP/SOIP
Additional Host I/O Adapter required?	Yes	No, standard Network Interface Card (NIC)
Can you use existing communications network	No	Yes
Other process dependant	No	Yes, IP
Client/Server application	No	Yes
Inherent Data Sharing	No	Yes
Host overhead	Yes (minimal)	Yes (could be substantial)
Heterogeneous Environment support	Yes, with enterprise disk subsystems	Yes, with IP applications such as CIFS and NFS
All I/O must be post processed by a custodial server	No	Yes
Guaranteed I/O commit	Yes	Yes, with TCP, no with UDP

1.1.2 Exploding some of the myths

The first myth is that SANs are expensive, and NASs are cheap.

This myth is typically associated with the hardware aspects relating to the solution. Both solutions require a network topology: Ethernet has become the de facto LAN standard for the majority of organizations, which, when considering the need for storage I/O over this medium, *does* offset initial costs of connectivity infrastructure and knowledge pertaining to systems/network administration personnel.

That is, you may be able to use the existing Network Interface Card (NIC) in your server for I/O, but it may share and conflict with the existing network traffic on that card/segment. If you introduce another NIC card (for a private storage LAN), then additional network ports will be required in switches. At this point, the NAS solution starts to realize hardware costs that can be similar to that of a SAN. If we discount the hardware argument, some truth can be found in the personnel/training investment. This is covered further in 1.7 “Personnel” on page 23.

The second myth is that NASs are slow and SANs are fast.

Fast and slow are subjective relative terms; speeds and feeds are increasing at an exponential rate. If architected correctly both solutions can perform well. There is always going to be a faster solution available to you, but typically at a price. We recommend that you step back and look at your requirements before committing to the latest high speed technology. Will the end user really notice a 2 millisecond difference in the presentation of the data, or are there so many other weaker links and bottlenecks in the application infrastructure that a second or two additional delay really will not make a difference? File size characteristics also have a factor on which solution performs better; NASs can perform rather well with small files and the greater benefit can be seen in a SAN when larger files are accessed.

In Figure 1-1 we show the typical data path.

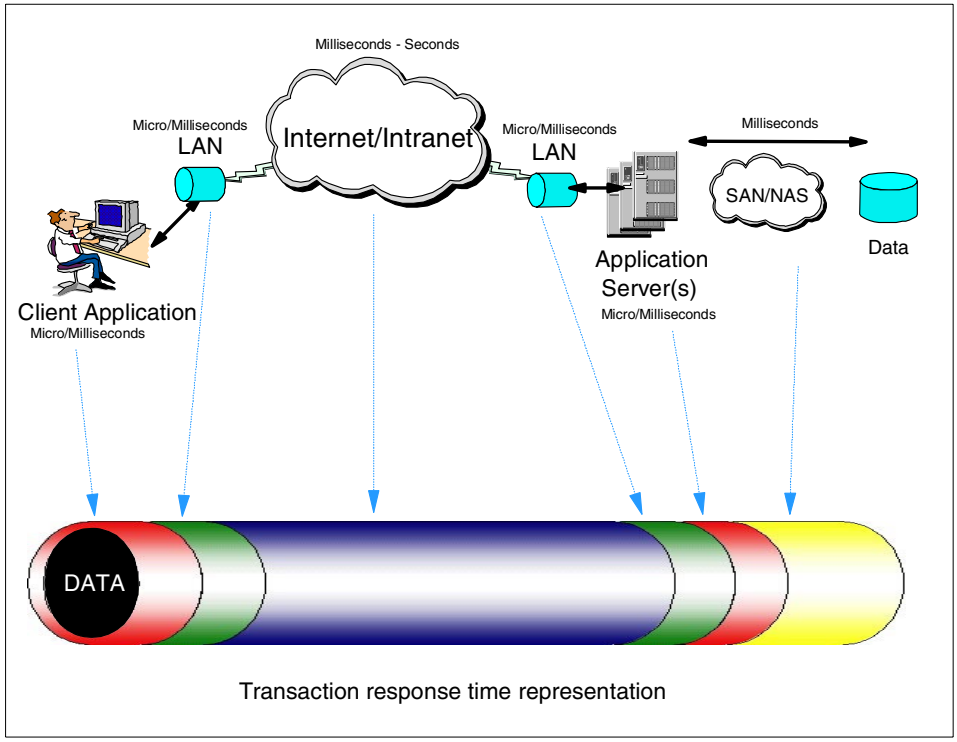


Figure 1-1 Typical data path for client/server applications

The third myth is that SANs and NASs cannot co-exist as a unified solution for storage consolidation.

It has appeared to be counter-productive to have both solutions implemented as the business goal of storage consolidation was not achieved by having separate islands of storage. NAS devices exist today that perform all of the tasks associated with a NAS device but have no direct (local) storage for file sharing attached to them. These devices are known as gateways (also known as LAN/SAN gateway), and use the SAN for access to storage devices. Introducing a gateway device overcomes the storage island issue, by allowing all storage to be consolidated.

1.2 Business and technological goals

In the topics that follow, we will describe some of the most frequently encountered reasons when considering a SAN.

1.2.1 Realizing the true potential of consolidated storage

Let us say that you understand how much disk was being wasted by using dispersed and local (to server) storage devices (typically seen around 70%). This is shown in Figure 1-2.

.

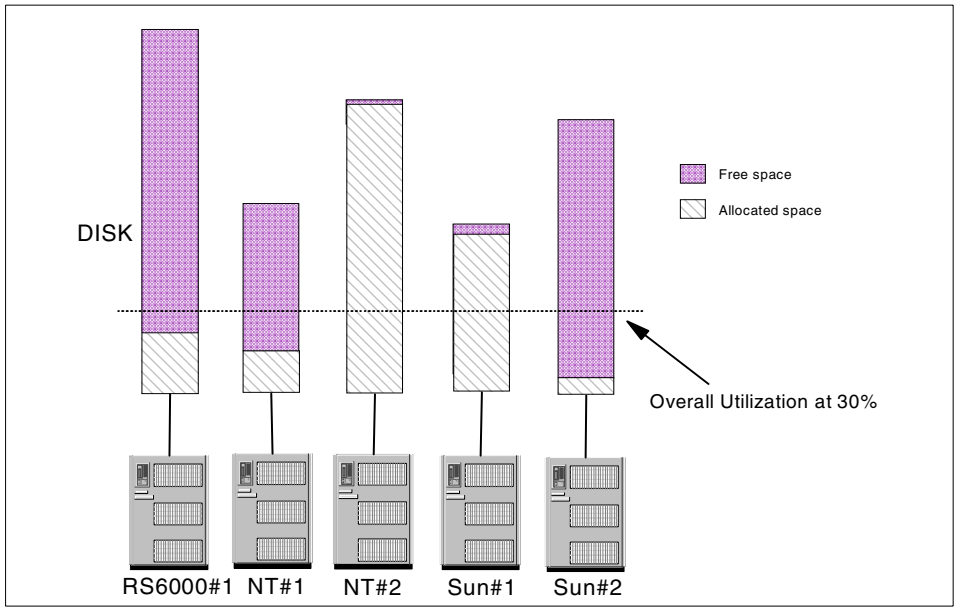


Figure 1-2 Typical server disk allocation representation

And you were questioning the amount of effort required to manage these devices, so you implemented a consolidated storage device as shown in Figure 1-3.

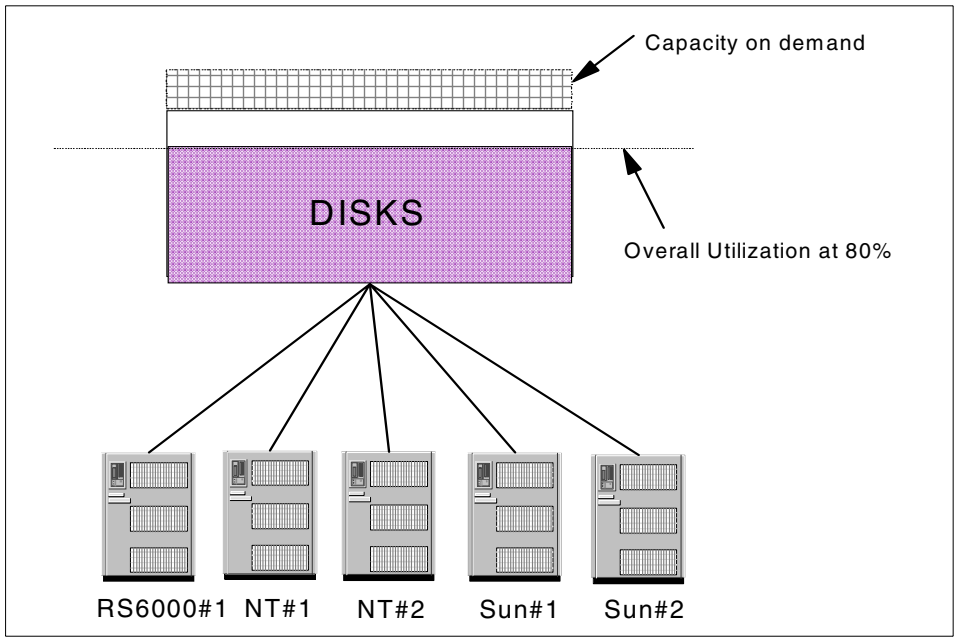


Figure 1-3 Consolidated disk storage efficiently sharing capacity

Whatever the rationale was for implementing a consolidated storage solution, now you realize that you have used up all the relatively expensive ports on your storage device, and now you want more servers to take advantage of the scalability and manageability of your storage device. This is where a SAN implementation will benefit your environment and make sound business sense.

1.2.2 Investment protection

In this section, we will discuss the implications surrounding protecting your investment.

Existing equipment

Now you are looking at a SAN solution for your chosen reason, but you have a huge investment in multiple gigabytes and maybe terabytes of disk on the data center floor today that are not SAN ready. What do you do?

There is not one answer that fits all, but here are a few approaches that you may, or should, consider:

- Introduce components that will allow these legacy solutions to become part of your SAN solution (known as bridges, gateways or routers). Merge.

- Re-deploy these resources in a less critical area of your business (maybe test or development environments). Re-Deploy.
- Just leaving them where they are and supplementing with the SAN. Separate and Add.
- Replace the old with the new. Replace.
- Analyze the greater return based on application(s) requirements and choose selective applications to use the SAN. Case-by-case.

All of these solutions will have a differing effect on your budget and provide differing long and short term goals.

Merge

The factors to consider here include:

On the down side:

- Legacy disk solutions may be more prone to failure than newer technology due to their age with respect to their life span
- Added complexity and management costs of bridges/routers
- Non-standard performance characteristics across disk complex
- Non-standard disk management techniques within the organization

On the up side:

- Initial investment in consolidated storage may be somewhat less of a hit as you may consider a partial implementation
- Existing equipment can be used further and depreciate over full-term
- Existing equipment is now part of a storage pool concept
- Reduced impact to business (no immediate migration of data off disks)

Re-deploy

On the down side:

- Non-standard disk management techniques within the organization
- Re-deployment environment will show differing performance characteristics of SAN environment
- Back-up strategy cannot take advantage of LAN-free or server-free
- Existing distance limitations still exist
- May not take advantage of storage-on-demand, and efficiently use legacy equipment

On the up side:

- Cost avoidance

Separate and add

On the down side:

- Non-standard disk management techniques within the organization and potentially on the same server
- Differing performance characteristics of devices to the application, potentially negating any performance benefit of introducing a SAN
- Back-up strategy cannot take advantage of LAN-free or server-free
- Multiple host adapter interfaces (SCSI/SSA and FC) in server, therefore, increasing complexity
- Islands of disks still present

On the up side:

- Cost avoidance
- Potential for migration path to be simplified, with less impact

Replace

On the down side:

- Higher cost option
- Greater risk and potential negative impact (if not managed correctly)
- New SAN skills needed immediately

On the up side:

- Standardized disk management policies across the enterprise
- Standard I/O performance characteristics across each application
- Ability to take advantage of LAN-free and Server-less backup solutions
- Gain full advantage of consolidated storage

Case-by-case

On the down side:

- No clear migration path to a SAN defined
- Non-standard disk management techniques within the organization and potentially on the same server
- Back-up strategy cannot take advantage of LAN-free or server-free
- Could result in long, drawn-out process

On the up side:

- Eases your introduction into SAN technology
- Less/no impact on current applications
- Sets the stage for redeployment, merge, or separate and add options

New equipment

Careful consideration of your future requirements should be taken into account when selecting your initial SAN design. SAN features and functions are constantly being introduced. Inquire and understand the road-map offered by different vendors prior to selection to determine whether these functions are of interest to you in achieving your long-term strategic goals. Look for solutions that are upgradeable rather than replaceable. Vendors that have shown a track record of adopting future standards and technologies should be preferred. This approach should assist in providing a solution that will not need to be replaced every time you wish to introduce new technology.

1.3 Service levels

The cost of your infrastructure may be dictated by operational service level requirements.

Management has a passion for service levels, as this implies a greater degree of control over availability, with the ability to seek compensation or apportion blame when up-time goals are not met, and inversely, take the credit when service levels are met and exceeded.

Vendors love to have service levels as they theoretically prove the point (based on statistics derived from the mean) that up-time goals can be met and this also assists with the sale of product: this sounds like a win situation.

The reality is somewhat different: first the customer needs to be able to identify each measurable entity, then be able to measure the pertinent metrics. The end result (if service levels are not met and the question remains, which ideally means that it can be proven with the vendor) what action will be taken?

Vendors know that only on very rare occasions will a customer actually implement a method for measuring availability and if they do, typically no contractual obligation will exist to rectify a situation.

Should you take on the task of measuring availability, the next question becomes: What needs to be measured and how do I measure it?

There are various ways to measure outages and uptime, from a manual log, automated system log collection and review, to an integrated Simple Network Management Protocol (SNMP) solution.

For the purposes of measuring SAN infrastructure uptime, the latter two options should be considered due to the granular time frame that may be associated with SAN component outages.

Director class switches are typically sold as 99.999% availability, switches (without the title of director) are known to have 99.9% availability. An IT architect should ask the question: Is it worth a 99.999% investment into the back-end infrastructure if the front end (network, servers, and applications) are designed to tolerate 99.9%? This is not to say that the answer will be no, as the effect on the availability goal could be compounded if each component within the infrastructure had a 99.9% availability.

Important: Do not fall into the trap of associating the impressiveness of the names of solutions with the aptness of the solution.The term *director* sounds more impressive than *switch*, which in turn sounds more impressive than *hub*.

If we look at the full RAS acronym (Reliability, Availability, and Serviceability), then manufacturers of equipment typically refer to the published MTTR (Mean Time To Repair) and MTBF (Mean Time Between Failure) figures.

These figures are normally in the region of tens of years for MTBF, and tens of minutes for MTTR. A key differentiation between directors and switches comes in the nature of serviceability: the ability to perform concurrent microcode/firmware upgrades with no loss of service is a standard feature in director class products. This being said the boundary between director and switch class products appears to be constantly shifting.

1.4 Disaster recovery and data protection

In this section we will attempt to relate the interruption to service caused by data not being readily available, and the necessary precautions that can be taken to prevent and/or minimize this.

Each business vertical can relate a different financial loss of revenue if data is unavailable. Some can sustain outages of differing lengths of times, which will determine what recovery plan is acceptable or suitable.

Various studies have been performed and published by independent groups that show the effect of down time to business, a sample of which can be found in Table 1-3.

Table 1-3 Example of industry loss of revenue related to downtime

Industry	Loss in Revenue per Hour
Retail Brokerage	$7,000,000
Credit Card Authorization	$3,000,000
Pay-Per-View Media	$1,300,000
Home Shopping	$130,000
Catalog Sales	$100,000
Airline Reservations	$100,000
Packaging	$40,000
ATM Service Fees	$20,000

The data represented reflects the variance in different vertical industries for one hour of downtime; it does not reflect the compounded issues relating to downtime and the resulting loss of business (revenue) due to loss of faith. We also need to note that there are many building blocks within the IT Infrastructure that could cause an outage; access to storage just being one, but access is the one we are concerned with in this book.

The cost of data protection is not directly proportional to time to restore, this is outlined in Figure 1-4. Solutions such as RAID protection may offer the least impact to the application, but if performed in the disk control unit, RAID may not offer much in the way of geographic resiliency. Other solutions such as the use of the Pickup Truck Access Method (PTAM) may offer a greater geographic resiliency, but will take considerably more time to restore from.

Note: Studies show that 20% of IT budget is allocated for hardware, 35% of this is storage related.

In Figure 1-4 we show the relationship between the cost and time to recover of the data restoration options.

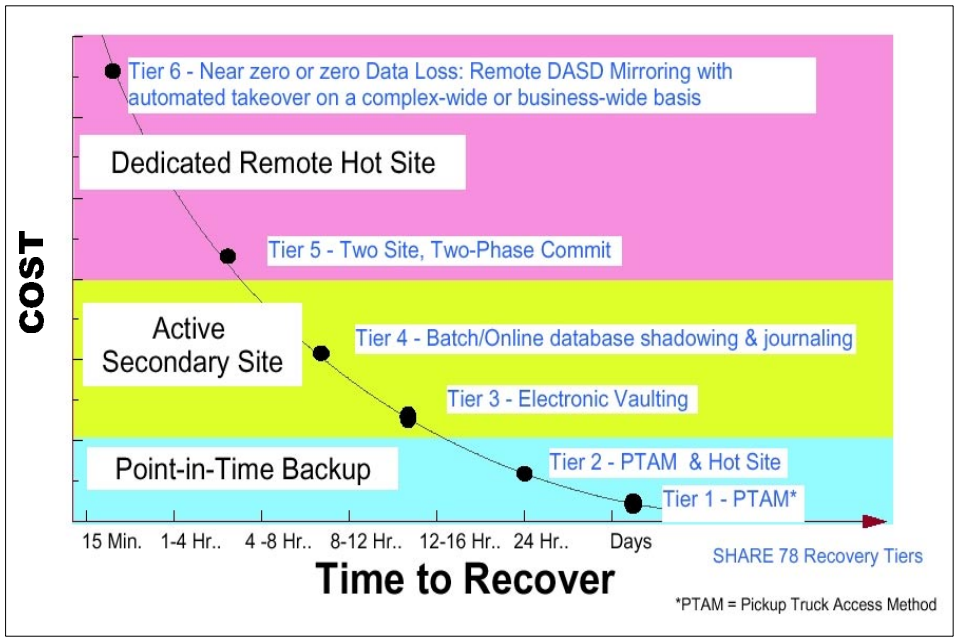


Figure 1-4 Data restoration options: Cost/time/solution

We will detail the different tiers and their history.

At SHARE 78 held in Anaheim, California in 1992, session M028, the Automated Remote Site Recovery Task Force presented seven tiers of recoverability, which were ranked based on the recovery method and recovery time. Although, this is almost ten years old now, we still feel that it has a valid place in today's society and economy.

Tier 0

This tier provides no preparation in saving information, establishing a backup hardware platform, or developing a contingency plan.

The length of time for recovery is unpredictable. Your data can be safely regarded as unprotected, and if you are in this tier, then you really do not care about your data. Perhaps we should just call it *tears*, because that is usually what it ends in — tears.

Tier 1

To be at Tier1, an installation would need to develop a contingency plan, back up required information and store it in contingency storage (an off-site location). It also should determine recovery requirements, and optionally establish a backup platform in a custom built facility, but without processing hardware. The length of time until recovery is usually more than a week.

in Figure 1-5 we show a Tier 1 recovery solution.

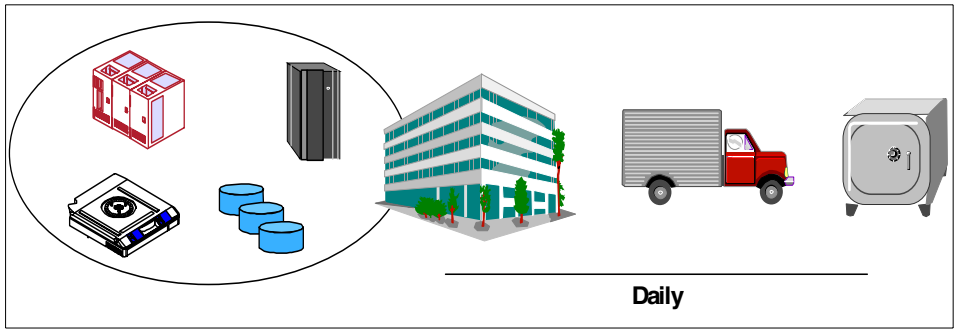


Figure 1-5 Tier 1 recovery solution

Tier 2

Tier 2 encompasses all requirements of Tier 1 and also requires a backup platform to have sufficient hardware and network to support the installation's critical processing requirements. Processing is considered critical if it must be supported on hardware that exists at the time of the disaster. The length of time for recovery is usually more than one day.

In Figure 1-6 we show a Tier 2 recovery solution.

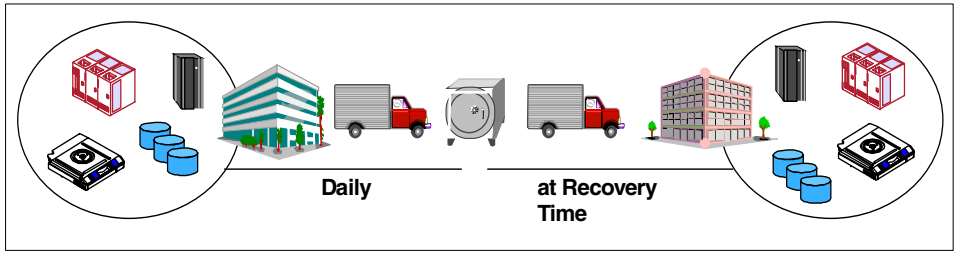


Figure 1-6 Tier 2 recovery solution

Tier 3

Tier 3 encompasses all the requirements of Tier 2, and in addition, supports electronic vaulting of some subset of the information. The receiving hardware must be physically separated from the primary platform and the data stored for recovery after the disaster. The length of time is usually about one day.

In Figure 1-7 we show a Tier 3 recovery solution.

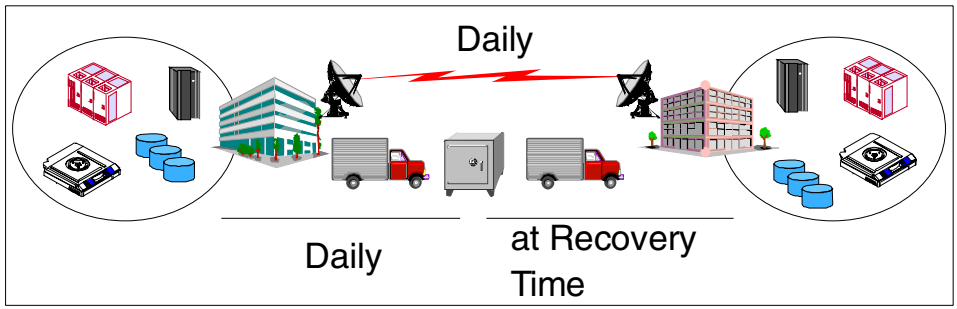


Figure 1-7 Tier 3 recovery solution

Tier 4

Tier 4 introduces the requirements of active management of the recovery data by utilizing a processor at the recovery site, and bi-directional recovery. The receiving hardware must be physically separated from the primary platform. The length of time for recovery is usually up to 24 hours.

In Figure 1-8 we show a Tier 4 recovery solution.

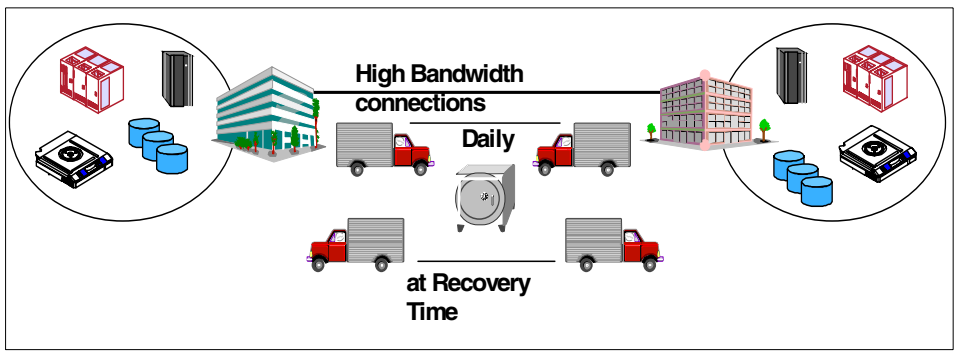


Figure 1-8 Tier 4 recovery solution

Tier 5

Tier 5 encompasses all the requirements of Tier 4 and, in addition, will maintain selected data in image status (updates will be applied to both the local and remote copies of the databases within a single commit scope).

Tier 5 requires both the primary and secondary platforms data to be updated before the update request is considered satisfied. Tier 5 requires partially or fully dedicated hardware on the secondary platform, with the capability to automatically transfer the workload to the secondary platform. The length of time for recovery is usually less than 12 hours.

In Figure 1-9 we show a Tier 5 recovery solution.

Figure 1-9 Tier 5 recovery solution

Tier 6

Tier 6 encompasses zero loss of data and immediate and automatic transfer to the secondary platform. The length of time for recovery is usually a few minutes.

In Figure 1-10 we show a Tier 6 recovery solution.

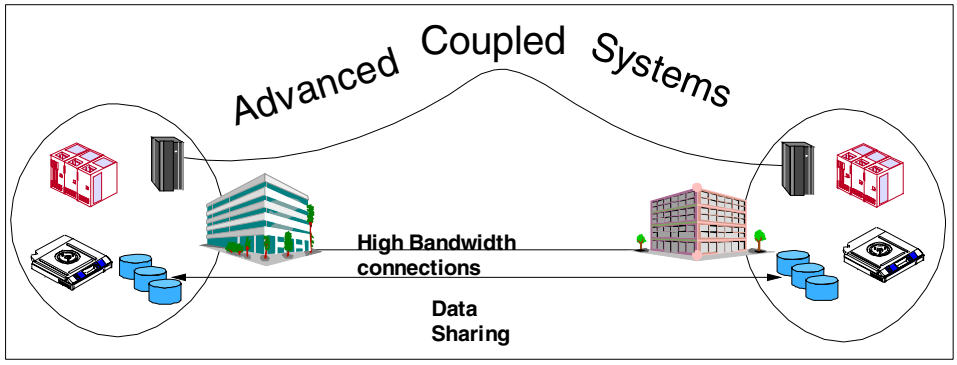


Figure 1-10 Tier 6 recovery solution

All of these functions represented can be achieved through the use of SAN components; however, whether this is the best place for all of them is debatable.

We also need to separate the notion that data availability is the same as data integrity. Although they can be somewhat related, data availability assumes that data is intact, but needs to be recovered, whereas data integrity indicates that data needs to be complete (that is not corrupt or missing).

The Fibre Channel infrastructure has enabled a more robust set of disaster recovery rules to exist. Distance limitations that have previously required very expensive solutions to overcome are now being overcome with native Fibre Channel technologies. Not only can disk be extended and mirrored with very little latency, but the ability to read and write to tape in the same manner.

1.5 Performance

Application performance should be of paramount importance to all companies. This does not imply that every company should design and implement the highest performing solution, but rather should set an expectation of what performance is acceptable and reasonable for each application.

For example, an internal e-mail system does not require the same level of performance as an Online Transaction Processing (OLTP) application.

Setting the expectations for what is acceptable is where time and consideration should be spent. If your company relies on an e-commerce site for revenue generation, then we must ask what is an acceptable response from a Web browser to present the data which has been retrieved, processed and transformed from the source. Analysts would suggest that a response time of less than 8 seconds will be acceptable for the end user, anything longer could result in end user boredom and a re-direct to another site: net result, lost business.

In this example of an open queueing model application with many unknown variables (such as network path and bandwidth, client processing capabilities), then the best that one can do is to benchmark within the confines of what is known (that is to say, within your environment) and then apply various overheads for the unknowns to determine approximate response times, such as when a client receives 10 KB of data returned within 1.2 seconds on the internal network.

A typical client with a 56 Kb modem may connect at 28 Kb, therefore, the download would take approximately 3.5 seconds. Total response time would be expected to be under 5 seconds, which is both acceptable and maybe reasonable (content dependant).

Likewise, expectations should be set when designing the I/O subsystem; the SAN being one of those components. Many variables need to be considered to ensure that utilization is maximized, capacity is sufficient, and performance is acceptable.

1.5.1 Logical scalability

Performance goals should be reviewed alongside scalability factors. If we use the e-commerce example above, we can benchmark internally what the expected response time would be for one user; however, what will be the impact of 10 users or 100 users accessing the site simultaneously? In planning this, we need to consider that 10 concurrent users accessing the site simultaneously may represent a *real* work load of considerably more users as the arrival rate of such users is undetermined, but probably will not be simultaneous.

These types of tests should not only uncover bottlenecks, but identify capacity thresholds that may need to be set and monitored. We will refer to this form of scalability as logical scalability

Again, we can draw a parallel with this example with that of a SAN under the topics of blocking, oversubscription, over commitment, trunking, throughput and concurrent operation limitations. We discuss these later.

1.5.2 Physical scalability

Physical scalability must also be considered as a business goal. If you know that your environment is somewhat static on terms of growth of servers, throughput, and data storage, then you may not need to choose a solution (and incur the expense) which offers the most flexibility in expansion.

However, industry analysts have predicted that data storage will increase at a rate of 8 Exabytes (EB) or one million gigabytes per year for the next five years, which includes personal, home, and business use. Therefore, it is expected that most industries will need to plan for some form of data growth. Plan your infrastructure to cater for expansion for the next two to three years to ensure your investment has time to pay for itself.

Some companies are attempting to reduce overall costs by consolidating servers, although not always practical or possible from an operational management standpoint, or application co-existence conflict. If this is your company's goal, then maybe a highly scalable solution is not the best fit.

Other forms of server consolidation will be more prevalent in the coming years in the form of one large scalable server logically partitioned and sharing resources (such as I/O paths). This type of technology is becoming available in servers such as IBM pSeries 690 class servers. In this situation, we must be careful not to introduce I/O oversubscription at the server or bus level.

For the majority of businesses, data I/O performance is becoming less of an issue due to the exponential increase in hardware capabilities over the past few years. In the same breath, we can also say the requirement to store more data is growing exponentially. Is there a direct parallel between the two? No, as the profile of how data is being used has changed.

If we consider an application transaction path, the key areas of latency are typically network related and/or data I/O.

With the advent, introduction, and deployment of Fibre Channel, most direct connect connectivity throughput issues have been overcome. However, the introduction of consolidated storage devices that support multiple heterogeneous storage units and servers results in consideration in the design of SANs. The most common bottleneck in a SAN environment will come from oversubscription; this will be discussed further in 2.11.3 “Oversubscription” on page 102.

This is not to be confused with the typical throughput issues associated with disk storage, which are related to access densities (number of I/Os per GB), which are compounded when large drive sizes are used (such as 180 GB), but *can* also be relieved by the efficient use of cache (workload permitting).

In Figure 1-11 we show the disproportionate manner in which disk capacity has increased with respect to speed.

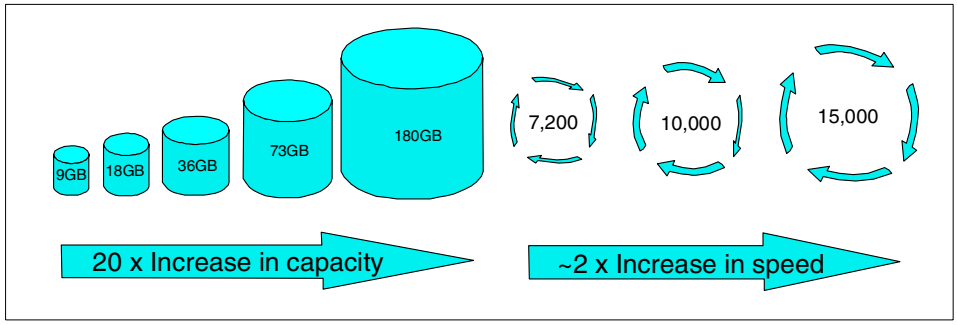


Figure 1-11 Disproportionate growth of disk capacity to speed

It is also worth noting that not all disk drives are available in all speeds. For example, 180 GB drives tend to only be available in the slower revolutions per minute (RPM) speeds at the time of writing. Even though disk speeds are increasing, accessor speeds are still bound by similar constraints for access to the platter, therefore, only the rotational delay and read/write speeds will be improved with higher speed disks.

To identify and isolate bottlenecks, performance monitors, and performance data collection tools (supplied with most storage, server and SAN devices) should be activated and monitored. Sudden performance degradation may be related to obvious factors such as component failure. These events can be captured and/or will be indicated by light emitting diode (LED) displays.

However, they can also be caused when a threshold or the knee of the curve for throughput has been exceeded. These situations may only be brought to light if historical performance data is available. Other server based I/O analysis tools (such as iostat, perfmon and iometer) can be used to view hot spots and delays to logical disk volumes).

Attention: Multiple LUNs can be created on the same physical volume. The impact of a highly-used LUN on one system can affect the performance of another LUN on another system if placed on the same physical disk. This may not be obvious through the use of system I/O monitors. Consolidated disk solution monitors should be used in this case.

Today, 1 Gb and 2 Gb Fibre Channel switched fabric SANs are readily available. These solutions satisfy the needs of most end users; in fact most SAN component vendors will state that no server and storage clusters exist today that can truly test the aggregate throughput capabilities of high port count SANs.

1.6 Resource sharing

To understand our business goal of resource sharing, we need to determine which resources we need to share. Resources relating to SAN are defined as being:

- Disk: Typically, a consolidated disk solution that offers flexibility in configuration and multiple heterogeneous server support. Sharing this resource with more than a few servers can make a good business case for implementing a SAN in itself. Small to mid-size consolidated disk solutions typically only provide capacity for a small number of direct host attachment ports, to connect additional servers, and effectively use the space provided with these disk solutions requires the implementation of a SAN.

- Tape: Enterprise level libraries aside, most tape drives and smaller libraries have been directly attached and owned by a server. Using software to manage the access to the library (such as Tivoli Storage Manager Tape Library sharing), multiple servers can access drives and tapes through a SAN fabric. Other forms of tape library sharing can be achieved using Network Data Management Protocol (NDMP) clients in conjunction with managing software, such as Tivoli Storage Manager.
- Connectivity: Network resources inherently provide resource sharing capabilities in the form of switches and directors. That is, multiple servers can connect to the same device and use common components (i.e. backplane, ASICs). From a logical stance, connectivity solutions in Storage Networking also support multiple concurrent protocols.
- Data sharing: This is not to be confused with disk sharing. Data sharing should be defined as the ability to access the same data from two or more sources (or hosts). This capability is available in a SAN environment in a limited fashion and they are usually of a homogenous nature. Software based solutions are available (such as Tivoli's SANergy), which do allow for heterogeneous support through a managing server. This managing server performs enqueueing on the data source and manages the security aspects of the clients and storage (file systems).

1.7 Personnel

For this topic, we look at some of the roles and responsibilities of the new breed of SAN administrator.

1.7.1 Areas of responsibility and ownership

There are a number of factors to consider when reviewing the personnel requirements to implement and support a SAN. The first and foremost we should consider before studying the costs related to this is *who* (that is to say, which department(s)) should support the SAN environment.

Many small to medium size companies will have no choice as their IT support structure will more than likely consist of a few people only. Other larger companies will have groups that consist of: network administration, system administration, database administration, capacity planning and architecture, hardware implementation, storage management and operations support.

Core skills

If we consider the core skills necessary in understanding the SAN environment then most (if not all) of the above groups have key knowledge and skills that could be used to manage the various items that constitute a SAN. These may include:

- The system administrator understands the configuration of the disks and the host bus adapters.
- The network administrator understands the concepts of networking and paths.
- The storage manager understands I/O requirements related from people such as the database administrator.
- The hardware implementor knows which cables to lay and ensures power, cooling and space are available for all components.
- The capacity planner/architect has conceptually designed the solution and knows which components should be hooked-up to which and why.
- The operations support personnel need to manage the environment on a day to day basis and need to know how to perform first level support.

So, is there one place that best takes ownership or responsibility for this infrastructure? We feel that a distributed approach is probably best at this time, but with clearly defined roles and responsibilities for each effort required in managing the SAN.

Table 1-4 should assist in defining these responsibilities.

Table 1-4 Areas of SAN infrastructure responsibility

Group/Department	Function
Capacity planning and architecture	Business case Define performance requirements Design Configuration management Ongoing performance monitoring
Storage management	Collect users disk requirements Ensure backup of data is performed Data restoration procedures Allocation of LUNS Disaster recovery plan

Group/Department	Function
System administration	Load server drivers and firmware Install host bus adapters Define logical volumes and file systems Ensure monitoring is in place for operations support Provide first level support documentation
Network administration	Zone management Logical asset management (WWNs)
Hardware implementation	Monitor and manage environmentals Cable management Labelling Hardware asset management (serial numbers) Ensure maintenance (hardware and software is current)
Operations support	Ensure events are being responded to. Manage end user issues Identify and escalate problems to assigned area of responsibility

Without defining the roles and responsibilities prior to installing a SAN, much effort will be duplicated, frustration within the support teams will surface and the potential for organized chaos will transpire.

Tip: It is worth considering the cost benefits of engaging with certified professionals for the implementation and ongoing maintenance of your SAN.

1.7.2 Training

As companies have accepted SAN solutions over the past few years, some may have neglected to look at the investment (cost) required to successfully educate the correct personnel with the correct skills.

Your IT department has made considerable investment into hiring and training the most suitable candidates for the job with attributes that reflect their job title. As an example, a network expert may fully understand concepts associated with protocols, routing mechanisms and network layers, as the SAN itself (as the name suggests) is a network and therefore inherits similar properties.

If we look at the skills required in Table 1-4, we must consider the effectiveness of training the most suitable people for the task at hand.

Cross pollination of knowledge is an area to consider to ensure that one individual does not become indispensable.

1.7.3 Setting the standard

If responsibilities are designated to the correct organizational structure the first time around you will save considerable time and costs. Do not overlook such mundane tasks as inventory management and maintenance renewal, as these tasks tend to absorb huge amounts of time if abandoned at the outset.

1.8 Solution confidence

In the following topic, we cover some of the business considerations surrounding the vendors that will provide the solution you choose.

1.8.1 Supported and certified solutions

Most components within the SAN will co-operate with each other in this new era of new open standards. However, whether they work or not, you need to ensure that solutions are certified, and therefore, supported by the component vendor.

If you implement a non-supported environment, you may well open the doors to extended outages when each vendor in-turn refuses to fix or take ownership of the problem.

If your solution incorporates a multi-vendor requirement, then you need to ensure that each product is cross-certified. This can become very complex; not only must the infrastructure components of your SAN be certified with the storage and server platforms (and typically the O/S level), but you also need to review the compliance of the solution with each application (such as your backup/restore software, database, clustering software, and so on).

If any one of these infrastructure products is not supported by your SAN design, you may have to review the overall effectiveness of the business solution. These situations are typically met when organizations run back-level software or have not implemented the most prevailing (sometimes referred to as Tier 2 and Tier 3) host solutions for their application.

1.8.2 Best-of-breed

SANs have assisted the architect by further enabling the decision for infrastructure components to be considered independently. Within the scope of supported and certified solutions, do not limit yourself to one vendor that has one global solution offering. Solutions exist that may satisfy your requirements more suitably by inter-mixing components from other vendors.

1.9 Rapid deployment, ability to accommodate change

If your business dictates the need for fast adaptability, then it is probable that your infrastructure may also. If you review the projects encountered within your organization over the past year and find that most of them were out of plan, then you should consider a solution that can scale. You may also want to consider increasing your margin of spare components (such as ports and/or blades) to accommodate change without having to go through the business process bureaucracy and fulfilment duration.

A different approach may be to consider the net impact to your business, such as missed opportunities and market leadership, when you do not implement a solution in a timely manner by following the routine processes. Review the project implementation life-cycle and determine the key delay areas to decide whether you have the control and authorization to change the mechanisms for these tasks.

1.10 Is it all worth it?

The previous sections have all provided input in determining the Total Cost of Ownership (TCO) and/or Return on Investment (ROI). This section describes the process that will eventually lead to determining whether it makes good business sense to invest in a SAN.

At this point it may be worth reading through the following chapters and returning to this section once you have identified what solution (topology and components) works best for your business needs.

It is worthwhile asking your selected vendor(s) to perform a TCO analysis for you. When reviewing vendor supplied TCO analysis, which has a direct comparison to a competing product, you must make a concerted effort to ensure that you understand the assumptions made. Does the vendor presenting the analysis have accurate competitive information on hand, or will it make assumptions (based on rumor, hearsay, or gossip?) in their favor?

Ensure that the TCO analysis fit's your need and not the vendors. An example of this might be to consider a three year TCO rather than five years.

1.10.1 Gather all your input, and then gather some more

At this time you should be prepared to perform the effort of due diligence and distribute your Request For Information (RFI) or Request For Proposal (RFP) to selected vendors.

From the RFP you will gather pricing information that will be used in the ROI and TCO study. Other questions that you should present to the selected vendors should include:

- How will the vendor assist with the installation and migration, and what is the associated cost?
- How many installations similar to the one being proposed has the vendor been successful with?
- What are the ongoing maintenance and support costs?
- Ensure that all RFP/RFI responses cater for and explain how they will accommodate your growth over the next *X* years.

Note: Ensure that all vendors provide support equivalent to each other that satisfy your service level requirements. Also, ensure that hardware *and* software maintenance is accounted for.

1.10.2 Focus on the identified business goals

You have determined why you want to employ a SAN, so make sure that you stay focused on that goal in determining what it is worth to your company to have that specific goal or goals. By investing in your solution for a selected business goal, you may improve the ROI of other related business goals down the path and protect the initial investment.

1.10.3 Cost avoidance

If your company has a need to procure additional storage arrays, whether or not a SAN is used, then this would have been deemed an expense, and can therefore, be included in the expense side of the analysis.

If the SAN is expected to prevent you from having to buy an array (refer to Figure 1-2 on page 8, and Figure on page 9) then the cost savings would be added to the benefit side of the analysis.

Staffing expenditure must be realized as a long-term cost benefit, but remember to offset the cost of staffing for time spent *beyond* the normal duties (SAN related). Long-term cost savings should prevail for personnel, if at the onset, the correct people are trained and managed.

Tip: Include only those items that relate directly to the SAN implementation when determining costs.

1.10.4 Calculating ROI

There are several standard ROI calculations in common use, such as net present value (in currency), internal rate of return (as a percentage), and payback period (in months). These are defined as:

- Net Present Value (NPV): A method used in evaluating investments where the net present value of all cash flows is calculated using a given discount rate.
- Internal Rate of Return (IRR): A discount rate at which the present value of the future cash flows of an investment equal the costs of the investment.
- Payback period: The length of time needed to recoup the cost of a capital investment on a non-discount basis.

Consultation with your financial department will probably determine that one of the above methods is used within your organization.

2

Constituent parts of a SAN

In this chapter, we introduce various items that may be used to build an IBM networked storage solution. This includes *physical* building blocks, protocols and concepts such as topologies. We also have a look at what can be used to manage the infrastructure.

More details of particular examples that are introduced in this chapter will be covered in Chapter 3, “SAN fabric products” on page 145.

2.1 Hardware

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.

2.1.1 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 2-1.

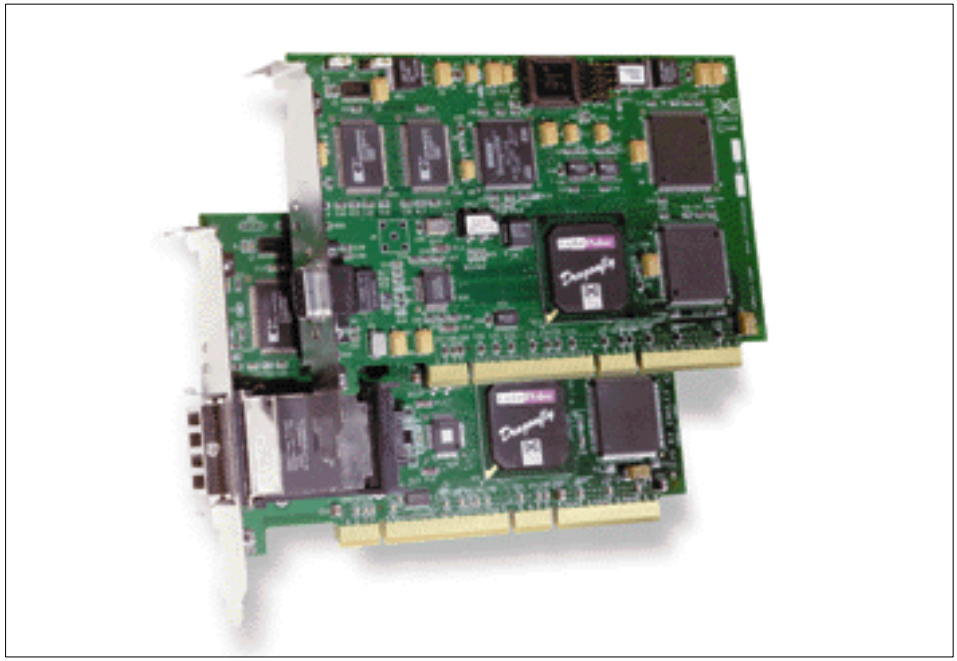


Figure 2-1 Host Bus Adapters

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

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

We cover these in greater detail in 2.1.9 “Cabling” on page 39.

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.

2.1.2 Bridges and SAN Data Gateways

A bridge is a device which converts signals and data 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.

2.1.3 Arbitrated Loop hubs

There are various ways to connect Fibre Channel devices together using different topologies, as shown in 2.2.2 “Topologies” on page 58. 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.

For a diagram showing the way that a hub connects ports together, see Figure 2-17 “Arbitrated Loop topology” on page 60.

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, but 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.

2.1.4 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.

2.1.5 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.

For a diagram showing the way that a switch connects ports together, see Figure 2-18 “Fibre Channel Switched topology” on page 61.

It is possible to connect switches together in cascades and meshes using Inter Switch links or ISLs, see “Switched Fabric” on page 60. It should be noted that devices from different manufacturers may not inter-operate fully (or even partially).

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

Features which are commonly 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
- Other protocols as well as Fibre Channel

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.

2.1.6 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 2.11.2 “Frame filtering” on page 102
- 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.

2.1.7 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.

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.

Features

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

Blocking

To support highly performing fabrics, the fabric components (hubs, switches, directors) must be able to move data around without any impact to other ports, targets or initiators that are on the same fabric. 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 100 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 are the Ferraris on the SAN racetrack. Non-blocking switches provide for multiple connections travelling through the internal components of the switch concurrently.

We illustrate this in Figure 2-2.

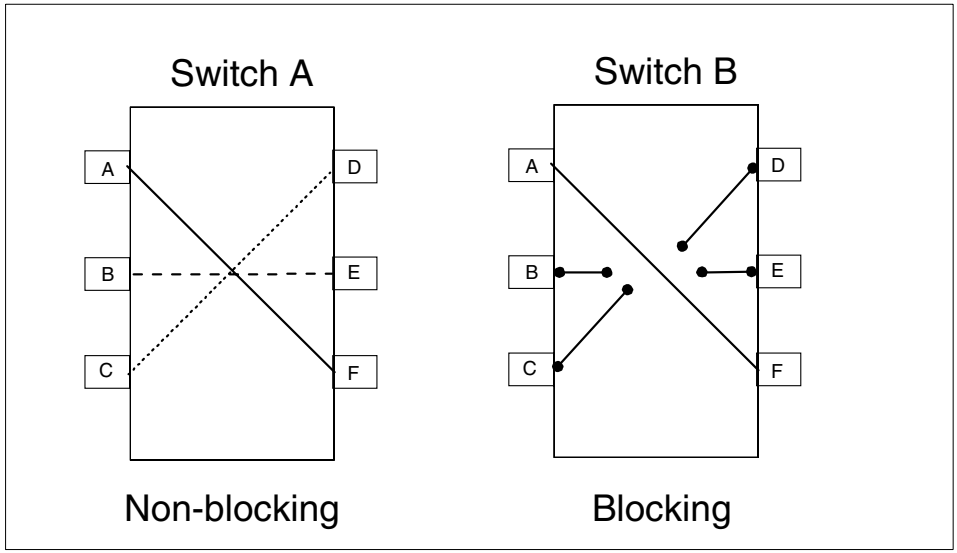


Figure 2-2 Non blocking and blocking switch designs

In the non-blocking switch, 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. In other words, the communication is not blocked. In the blocking switch (Switch B) while Port A is speaking to F, all other communication has been stopped or blocked.

2.1.8 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.

2.1.9 Cabling

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 telecommunications or open systems environments. The major difference between a Fibre Channel and an open systems LAN/WAN environment is the reduced cable distance between devices and associated attenuation loss. 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.

Shortwave or longwave

Every data communications fiber falls into 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 Fibre Channel physical layer working subcommittee (orange for multi-mode and yellow for single-mode). There may not be a 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.

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.

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 500meters from transmitter to receiver (or the reverse).

50/125 micrometers or 62.5/125 micrometers

Optical fiber for telecommunications consists of three components:

- Core
- Cladding
- Coating

In Figure 2-3, we show the physical characteristics of fiber optic cable.

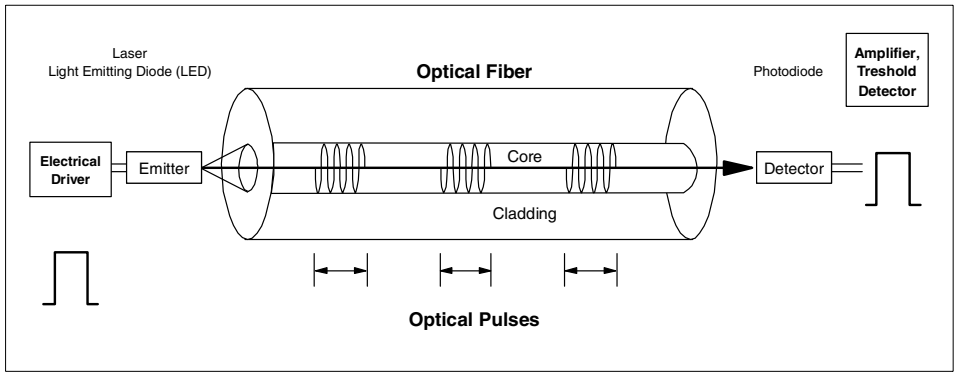


Figure 2-3 The structure of a fiber optic cable

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 micrometers (µm) to 62.5 micrometers. The standard telecommunications core sizes in use today are 8.3 µm (single-mode), 50 µm (multi-mode), and 62.5 µm (multi-mode). Single-mode and multi-mode will be discussed in “Single mode and multi-mode distances” on page 41.

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.

Most enterprises today use the 62.5 micron core fiber due to its high proliferation in the local area networks (LAN). The Fibre Channel SAN standard is based on the 50 micron core fiber and is required to achieve distances specified in the ANSI Fibre Channel standards. Customers should not use the 62.5 micron fiber for use in SAN applications. It is wise to check with any SAN component vendor to see if 62.5 is supported.

Single mode and multi-mode distances

Typical supported combinations are:

- 50 Micron Multimode Shortwave <= 500 meters
- 62.5 Micron Multimode Shortwave <= 175 meters
- 9 Micron Singlemode Longwave =< 10 Km

In Figure 2-4 we show single mode and multi-mode.

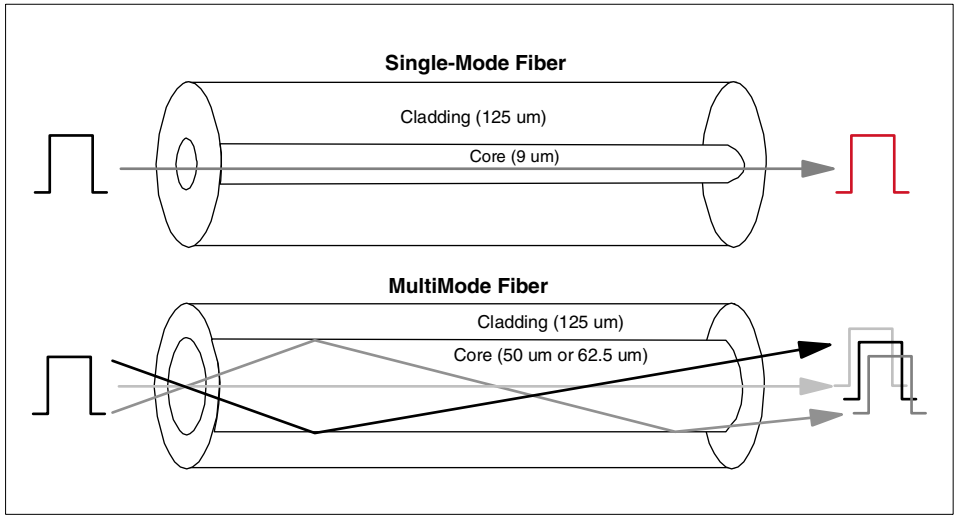


Figure 2-4 Single-mode and multi-mode differences

Copper

The Fibre Channel standards also allows for fibers made out of copper. There are different standards available:

- 75Ω Video Coax
- 75Ω Mini Coax

- 150Ω shielded twisted pair

The maximum supported speeds and distances using copper are lower than when using fiber optics, and are shown in Figure 2-8 on page 128. This is due to the transmission line and skin effect physics, which has a great significance at high frequencies.

Plenum rating

A term that is sometimes used when describing cabling is whether a particular cable is *plenum rated* or not.

In this case, a plenum is an air filled duct, usually forming part of an air conditioning or venting system. If a cable is to be laid in a plenum, there are certain specifications which need to be met. In the event of a fire, some burning cables emit poisonous gasses. If the cable is in a room, then there could be a danger to people in that room. If on the other hand, the cable is in a duct which carries air to an entire building, clearly, there is a much higher risk of endangering life.

For this reason, cable manufacturers will specify that their products are *plenum rated* or *not plenum rated*.

2.1.10 Dark Fiber

In order to connect one optical device to another, some form of fiber optic link is required. If the distance is short, then a standard fiber *cable* will suffice. Over a slightly longer distance, for example from one building to the next, then a fiber link may need to be laid. This may need to be laid underground or through a conduit, but it will not be as simple as connecting two switches together in a single rack.

If the two units which need to be connected are in different cities, then the problem is much larger.

Dark Fiber generically refers to a long, dedicated fiber optic link that can be used without the need for additional equipment. It can be as long as the particular technology supports, for example 10 Km for Long Wave laser GBICs, see 2.14.1 “Limits” on page 127.

Some forward thinking services companies have laid fiber optic links along their pipes and cables. For example, a water company might be digging up a road to lay a mains pipe; or an electric company might be taking a High Tension power cable across a mountain range using pylons, or a cable TV company might be laying cable to all the buildings in a city. While carrying out the work to support their core business, they also lay fiber optic links.

These cables are simply cables. They are not used in anyway by the company who has own/laid them. Hence, the name *Dark Fiber*. At some point in the future, another company may need to connect a computer room in one city to an off-site storage facility in another. Rather than laying their own cable, at vast expense, they can purchase bandwidth from the company has already laid the Dark Fiber. So, in this case, there will just be the need to connect their site into the Dark Fiber.

Dark Fiber should be one of the options considered for joining remote sites optically, but bear in mind the maximum distance limits discussed “Limits” on page 127.

2.1.11 Connectors

It hardly needs to be said that the particular connectors used to connect a fiber to a component will depend upon the receptacle into which they are being plugged. There are, however, some generalization that can be made. It also is useful to mention some guidelines for best practices when dealing with connectors or cables.

As far as we are aware at the time of writing, any 2 Gb/s devices will be using the SFF or SFP technology, and therefore, use Lucent Connector (LC) connectors. Most GBICs and GLMs use industry standard Subscriber Connector (SC) connectors.

SC connectors

The duplex SC connector is a low loss, push/pull fitting connector. It is easy to configure and replace. The two fibers each have their own part of the connector. The connector is keyed to ensure correct polarization when connected, that is transmit to receive and vice-versa.

See the diagram of an SC connector in Figure 2-5.

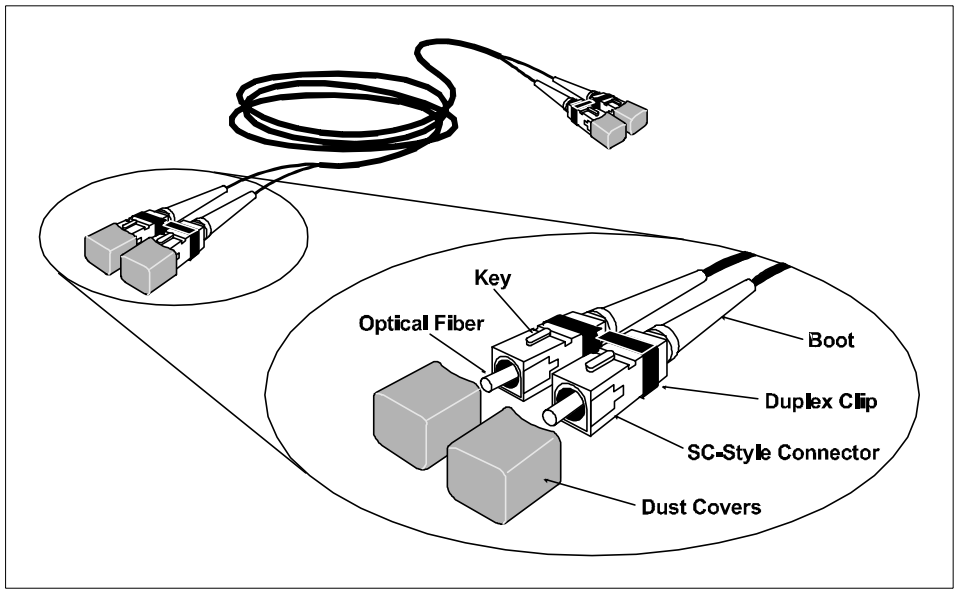


Figure 2-5 A fiber optic cable with an SC connector

In the diagram, there is a dust cover for each half of the connector.

Important: The dust covers should always be placed on the connector when it is not connected into a fibre channel device.

Using the dust covers actually has a number of important benefits:

- Because dust particles tend to be in the order of 10 to 100 microns across, and the actual fiber has a diameter of between 9 and 62.5 microns, it is clear that a dust particle could cause a problem if it rested on the light path.
- Another thing that can happen is that dust can be transferred from the connector to any device that it is plugged into. It is, therefore, necessary to keep the connectors as clean as possible.
- A third benefit to using a dust cover is that light cannot pass through it. They, therefore, protect unsuspecting people from eye damage.

Note: You should never look into a cable while removing its dust cover.

The SC connector above is actually constructed as two separate parts, which are held together by a clip. Another single piece design is shown in Figure 2-6.

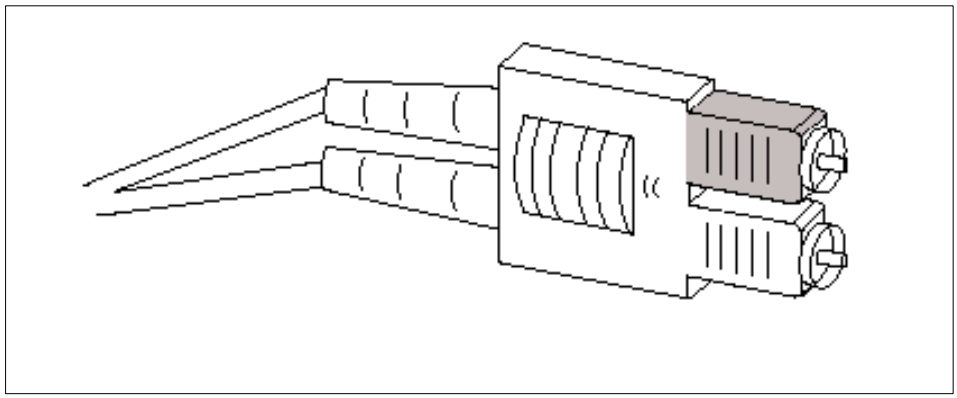


Figure 2-6 A single piece SC connector

LC connectors

The type of connectors which plug into SFF or SFP devices are called LC connectors. Again a duplex version is used so that the transmit and receive are connected in one step.

The main advantage that these LC connectors have over the SC connectors is that they are of a smaller form factor and so manufacturers of Fibre Channel components are able to provide more connections in the same amount of space.

We show a picture of an LC connector in Figure 2-7.

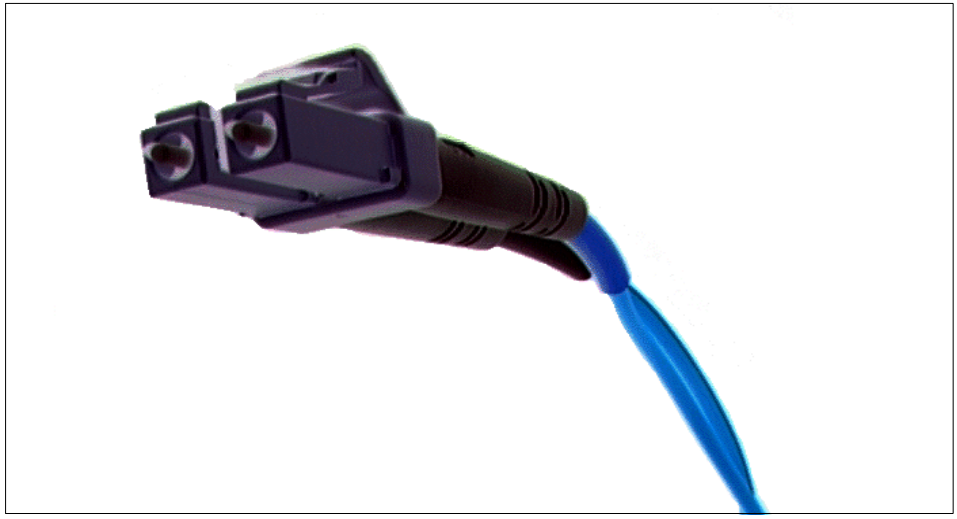


Figure 2-7 An LC connector

Copper connectors

A variety of connectors have been used for copper Fibre Channel connections, but the most popular has been the 9 pin D-Type connector. The familiar connector can clearly be seen as part of the Media Interface Adapter in Figure 2-12 on page 50. This is the connector most commonly used for twisted pair connections.

Other connectors that have been used for the co-axial copper options are the BNC and TNC connectors. The BNC connector will probably be familiar from 10Base2/thinnet ethernet usage. The TNC is very similar, but rather than having a bayonet fixing it is threaded.

2.1.12 GBICs, GLMs, and transceivers

These different components carry out the same task.

The task is that of converting electrical signals on HBAs, switches, and other building blocks to and from optical or electrical signals for transmission over the Fibre Channel media.

The format of the devices is, however, different.

GBICs

Gigabit Interface Converters (GBICs) are laser-based, hot-pluggable, data communications transceivers. They are suitable for a wide range of networking applications requiring high data rates. Designed for ease of configuration and replacement, they are well-suited for Gigabit Ethernet, Fibre Channel, and 1394b applications.

GBICs are available in copper, and both short wavelength and long wavelength versions, which provide 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. They provide a high-speed serial interface for connecting servers, switches and peripherals through an optical fiber cable. In SANs, they can be used for transmitting data between physical Fibre Channel ports.

The optical GBICs use lasers that enable cost-effective data transmission over optical fibers at various distances (depending on the type) of up to distances of around a 100 km. The transfer rates depend on the type. For distances and rates, see 2.14.1 “Limits” on page 127.

These compact, hot-pluggable, field-replaceable modules are designed to connect easily to a system card or other device through an industry-standard connector. On the media side, single-mode or multi-mode optical fiber cables, terminated with industry-standard connectors, can be used.

GBICs are usually easy to configure and replace. If they are optical, they use low-loss, push-pull, optical connectors. They are mainly used in hubs, switches, directors, and gateways.

A GBIC is shown in Figure 2-8.

Figure 2-8 A Gigabit Interface Converter (GBIC)

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.

GLMs

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.

A GLM is shown in Figure 2-9.

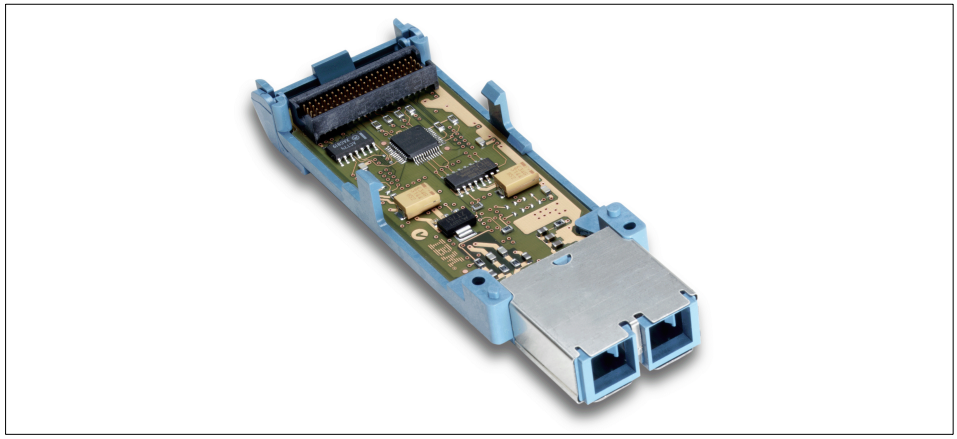


Figure 2-9 A Gigabit Link Module (GLM)

Transceivers

Another form of component to convert between electrical and optical signals is the transceiver.

Small Form Factor Optical Transceivers

Small Form Factor (SFF), or sometimes called Small Form Pluggable (SFP), and Transceivers serial optical converters are the next generation of laser-based, optical transceivers for a wide range of networking applications requiring high data rates. The transceivers, which are designed for increased densities, performance, and reduced power, are well-suited for gigabit Ethernet, Fibre Channel, and 1394b applications (see Figure 2-10).

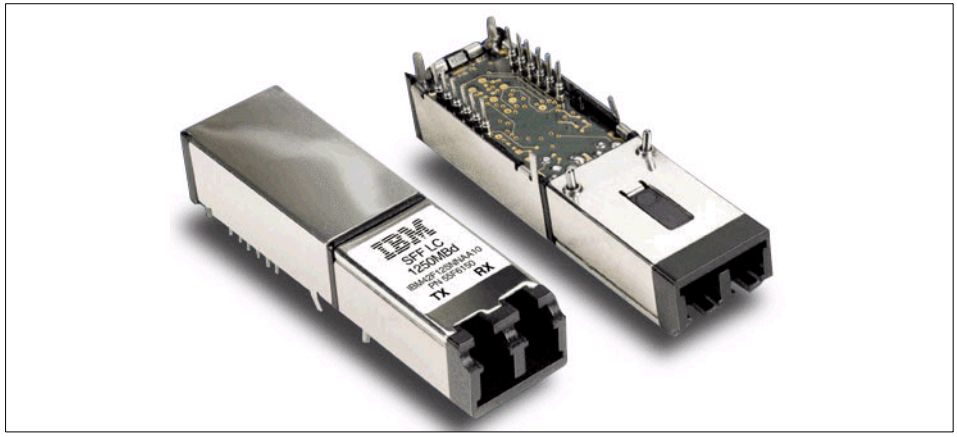


Figure 2-10 A Small Form Factor Transceivers (SFF)

The SFF optical transceivers use short wavelength and long wavelength lasers and are available in pin through hole (PTH) or hot-plugable versions.

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

The SFF serial optical transceivers are integrated fiber-optic transceivers that provide 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.

1 x 9 Transceivers

Some manufacturers have used 1x 9 transceivers in their products rather than GBICS or GLMs for providing optical connection to their devices. Arguably, 1x 9 transceivers have some advantages over GBICs, which are the most widely used in switch implementations:

- Easier to cool
- Better air flow

Figure 2-11 shows a pair of 1x 9 transceivers.

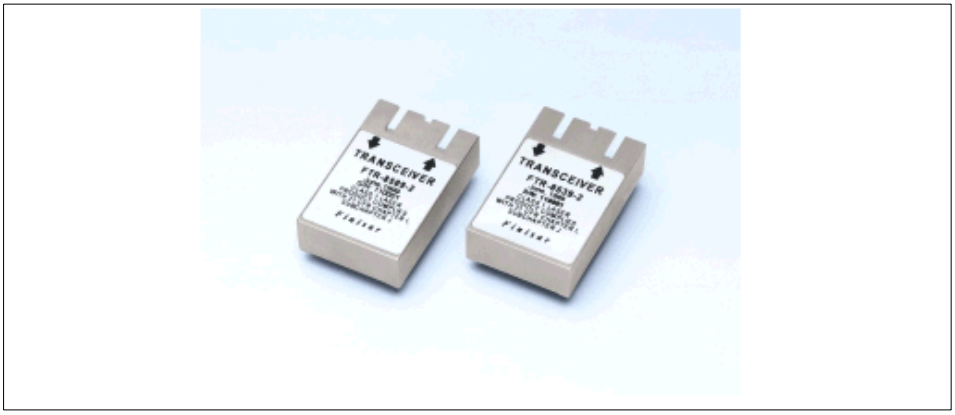


Figure 2-11 A pair of 1 x 9 transceivers

In the late 1990s, it was considered that 1 x 9 transceivers were more reliable than GBICs (2.5 times that of a GBIC). This is now a dubious factor, due to the advances in GBIC technology over the last few years, and the trend towards the use of GBICs.

MIA

Media Interface Adapters (MIA) can be used to facilitate conversion between two forms of media, usually 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.

A MIA is shown in Figure 2-12.

Figure 2-12 A Media Interface Adapter (MIA)

2.1.13 ASICs

An ASIC is an Application Specific Integrated Circuit. A device that was designed to do one particular function. One function may include several smaller things, but fundamentally it does a very specific job. Thus, it is not like a single chip microprocessor, which can be programmed to do many different jobs depending on the application.

This focus on a particular application means that an ASIC can be designed with the highest possible level of optimization. There may be some compromise, however, as optimizing the speed may influence the size of the device as more logic gates may be required and there may be an increase in the amount of heat which the device generates, as data travels through the device. There will also need to be a balance between the complexity of the device, and the time taken to develop it, debug it, and produce it in sufficient quantities for the product.

The design of an ASIC can heavily influence the way that a device carries out its job internally. So, manufactures will put a lot of emphasis on features of their particular ASIC design. The device from one manufacturer may carry out the tasks slightly differently from the way that another manufacturer does it, but the net, external result will comply to the standard. For this reason, ASIC design may influence the way that a SAN is designed.

In the case of switches and directors, ASICs are used to:

- Interface between ports
- Interface between I/O blades containing ports
- Interface between I/O blades and switching/control blades

In fact, ASICS actually implement the F_Ports or FL_Ports themselves and provide the memory for the port buffers. Depending on the particular design, those buffers may be allocated equally amongst all ports on the ASIC, or there may be the possibility to have different numbers of buffers on different ports on the ASIC. This would be of use particularly if particular ports were going to be used for longer distance links. These buffers equate to the BB_Credit described in “Flow control” on page 114.

The specific design will, of course, depend on the manufacturer, but typically there will be between one and eight ports being controlled by an ASIC. That ASIC will be involved in the routing of frames. The ASIC will be able to interpret the headers in the frame and route it to the appropriate port. Routing tables are set up so that the ASIC knows the best path to send the frame along. That path may be:

- To a port on the same ASIC
- A port on a different ASIC on the same blade
- A port on a different ASIC on a different blade
- A port on another switch/director

If there are two (or more) ports to an ASIC, then there will be less latency between those two ports than if the frame needs to travel from ASIC to ASIC in order to reach its destination. On the other hand if the ASIC fails then all the ports it controls may be put out of action.

It should be stated here that the latency at this level is in the order of a few microseconds or less. This may not seem like a great deal of time, but if there are several ASIC to ASIC hops (potentially in each of several switches in a cascaded or meshed fabric) then this latency can become more significant.

2.1.14 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.

2.1.15 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.

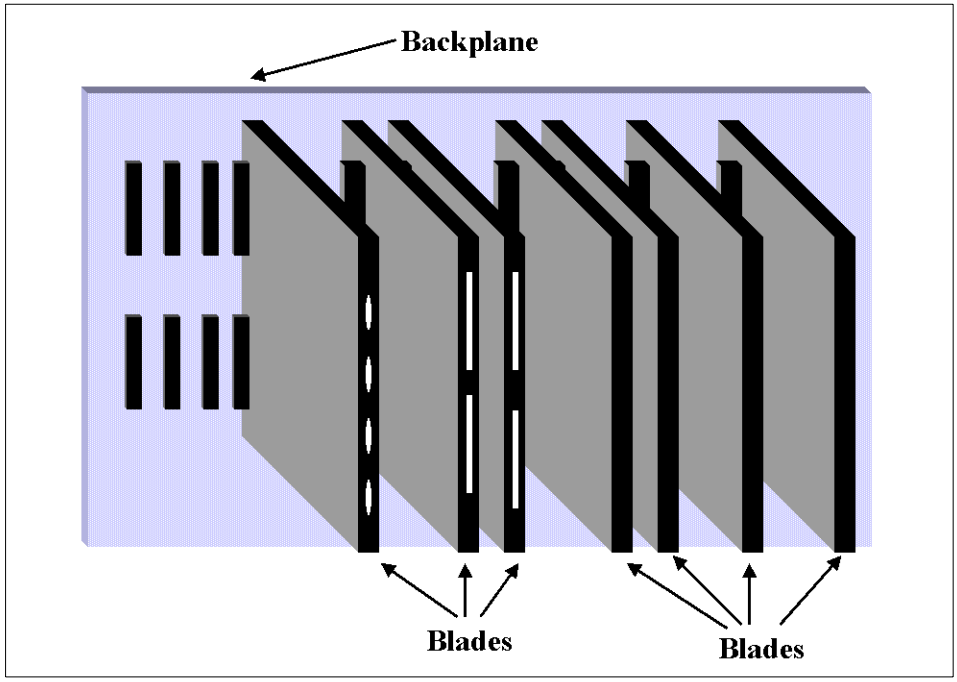


Figure 2-13 A diagram of a backplane and blades architecture

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.

2.1.16 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:

- Fibre Channel is a far less widely known and understood protocol. So, fewer people would be able to interpret the data gathered by a Fibre Channel analyzer than for an ethernet analyzer.
- 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.

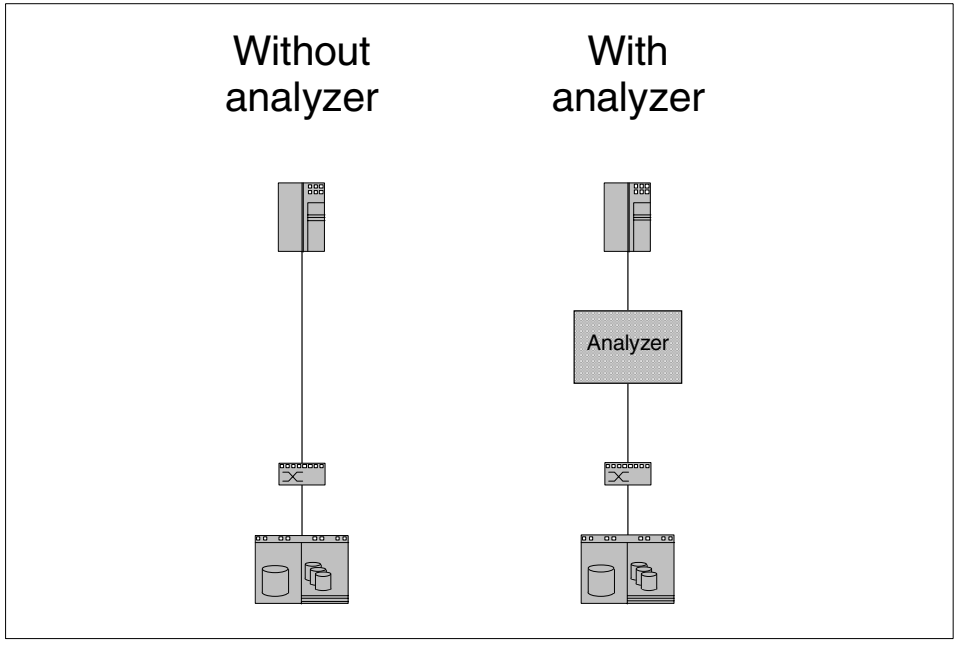


Figure 2-14 Connecting an FC analyzer

- 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.

2.2 Concepts

As well as the physical hardware there are some important concepts which need to be defined.

2.2.1 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 available bandwidth exploitation.

Based in 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:

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 to the source device. This class of service ensures that 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.

Class 2

In a Class 2 service there is no dedicated connection, each frame is sent separately using switched connections that allow 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. Class 2 makes a better use of available bandwidth since it allows the fabric to multiplex several messages in 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. The upper layer protocol should take care of frame sequence. It is up to the switch manufacturer to include design characteristics that ensure in order delivery of frames.

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 common option for SCSI.

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

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.

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.

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.

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.

2.2.2 Topologies

Fibre Channel provides three distinct interconnection topologies. By having more than one interconnection option available, a SAN designer can choose the topology that is best suited to the customer's requirements. The three Fibre Channel topologies are:

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

The topologies are shown in Figure 2-15.

No matter what topology is in use, from a node's perspective, Fibre Channel communications are always point-to-point, between the Initiator to the Responder.

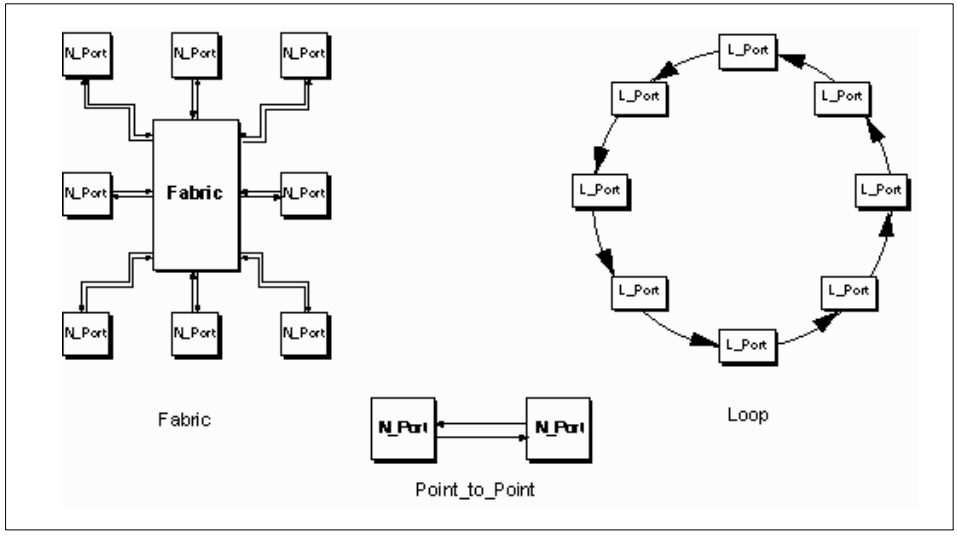


Figure 2-15 Fibre Channel topologies

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.

We illustrate a simple point-to-point connection in Figure 2-16.

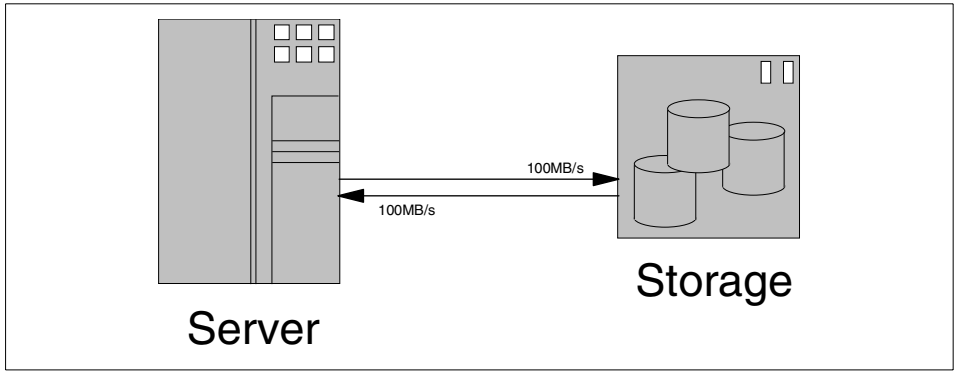


Figure 2-16 Point-to-point topology

A Point-to-point topology can also be seen as a (very simple) Arbitrated Loop, the important factor is whether or not FC-AL protocol is in use. The two nodes must both be using either FC-AL or standard Fibre Channel protocols.

Arbitrated Loop

The second topology is Arbitrated Loop. A slight variation on the standard Fibre Channel protocol is used, namely: Fibre Channel Arbitrated Loop (FC-AL). FC-AL is commonly used 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 control frames around the loop with a total bandwidth of 100 MB/s. Using an arbitration protocol, a single connection is established between an initiator and a responder, and data frames are 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. The actual distance between nodes (or between a node and a hub) depends on the link between them but may be many kilometers. Latency on the Arbitrated Loop, and therefore, for *all* nodes on the loop, is, however, affected by the loop size.

There is a diagram of an Arbitrated Loop implemented with a hub in Figure 2-17.

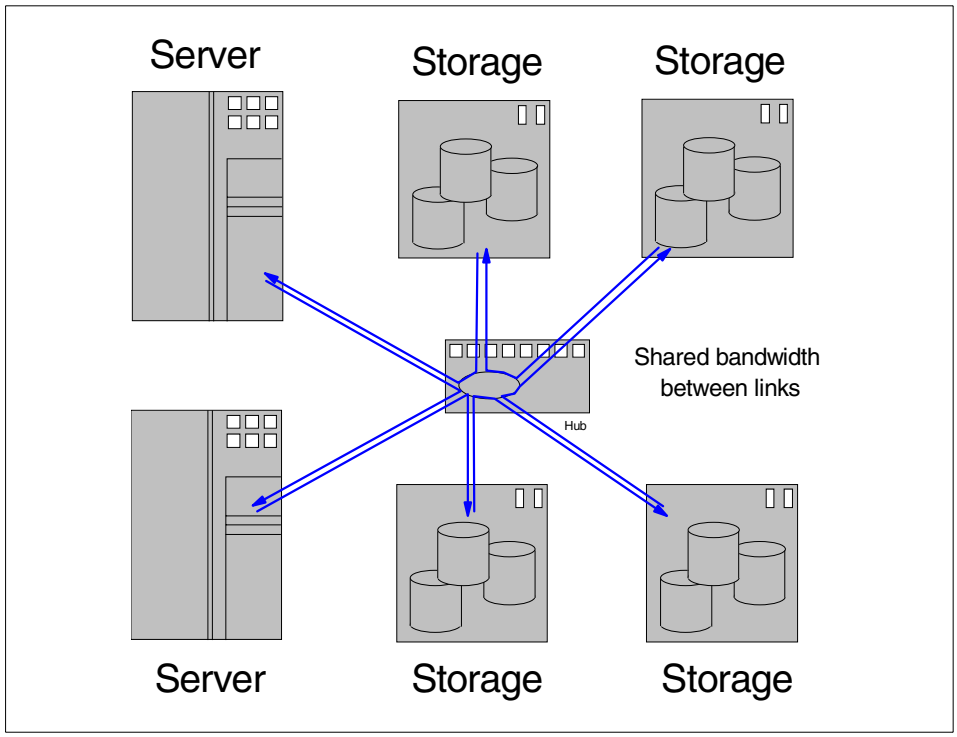


Figure 2-17 Arbitrated Loop topology

Although up to 126 devices may be configured in an Arbitrated Loop, it is usual for the number to be limited to one or two servers and a maximum of ten or so storage devices. If these numbers are exceeded, it is very likely that the bandwidth will become restrictive on data flow.

Switched Fabric

The third topology, and the one most commonly used in SAN implementations is Fibre Channel Switched Fabric (FC-SW). In this case, the fabric is one or more fabric switches (or directors, or even a combination of switches and directors) in a single, sometimes extended, configuration. Switched fabrics provide full bandwidth per port, compared to the bandwidth shared by all ports in Arbitrated Loop implementations.

If you add a new device into an Arbitrated Loop, you further divide the shared bandwidth. In a switched fabric, however, adding a new device or a new connection between existing ones actually increases the bandwidth. For example, an switch populated with 1 Gb/s GBICS with three initiators and three responders can support three concurrent 100 MB/s conversations or a total 300 MB/s throughput (600 MB/s if full-duplex applications were available).

A switched fabric configuration is shown in Figure 2-18.

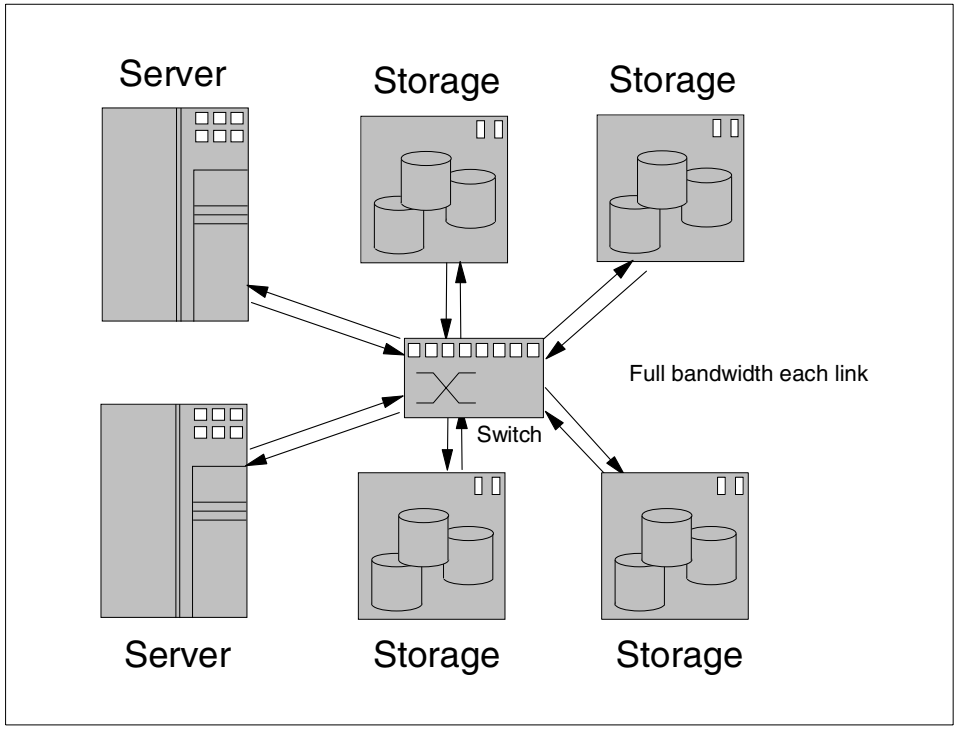


Figure 2-18 Fibre Channel Switched topology

Extending a switch fabric

As the demand for the storage grows, a switched fabric can be expanded to service these needs. There *is* a total limit of around 15 million devices on a fabric (see 2.4 “Addressing” on page 71). Not all storage requirements can be satisfied with fabrics alone. For some applications, the full bandwidth 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.

Cascade

One way to expand the fabric is to *cascade* switches. Cascading is basically interconnecting Fibre Channel switches. 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.
- 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. (See 2.12.8 “Time-outs” on page 113).
- When the fabric grows, the name server is fully distributed across all the switches in fabric.

A cascade of three switches is shown in Figure 2-19.

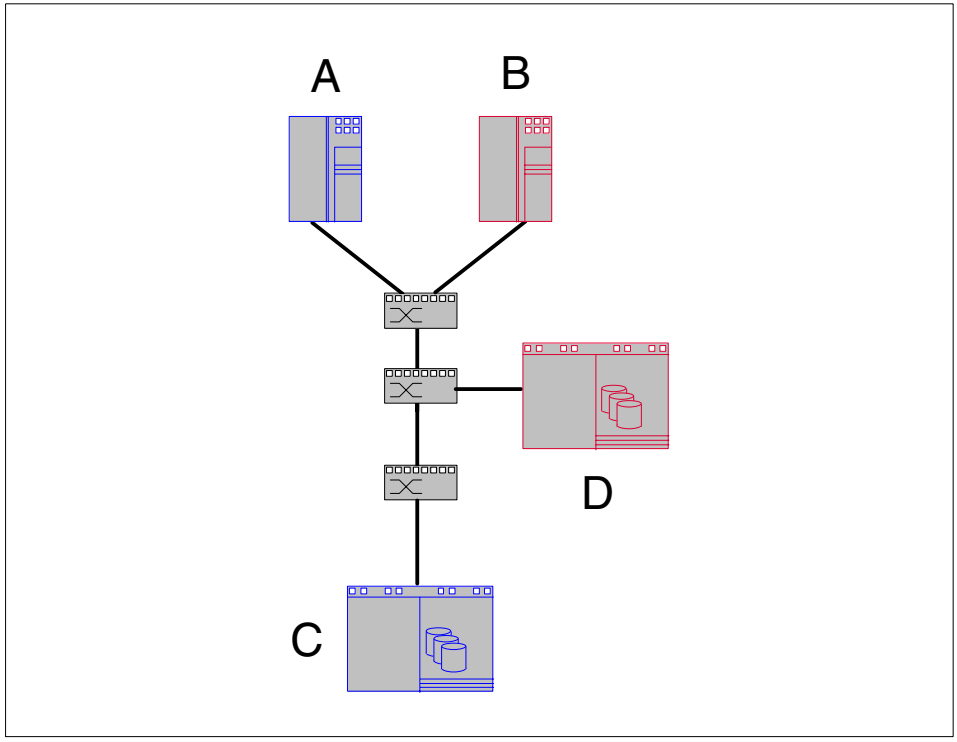


Figure 2-19 A cascaded fabric

Hops

The concept of cascading introduces the concept of hops. When frames travel across the fabric, they are said to “hop” from switch to switch. So in Figure 2-19 there are two hops between Server A and Storage C, one hop between Server B and Storage D. There are no hops between the two servers and there is one hop between the two storage devices.

Note: While it is possible to have more than one switch in a cascade, there is a limit. This limit will depend on the equipment in use. Some manufacturers may allow more hops than others.

Mesh

It is sometimes useful to create a more complicated fabric than either a single switch or a cascade of more than one switch.

In this case we connect switches in what is called a *mesh*. There are two kinds of mesh which are commonly in use as we see below:

- Full mesh
- Partial mesh

There are some benefits to creating a meshed fabric:

- Resilience due to adding alternative paths
- Possible load balancing over multiple paths
- Increase port count on the fabric (up to a point - see below)

There is some down side, though.

- As the switch count increases, so does the complexity of the mesh.
- The increase in complexity is not linear - it is worse than that.
- As the complexity of the fabric increases, so the ease of management decreases.

An alternative to using a meshed fabric of switches might be to use a Director which has more ports. This will ease the load of managing the fabric as it will be a single device, however, the initial cost may be increased.

Full mesh

In this case, all the switches in the mesh are interconnected by at least one Inter Switch link. There is a simple example of a meshed fabric in Figure 2-20. In this case, we have four switches and they are all connected to each other.

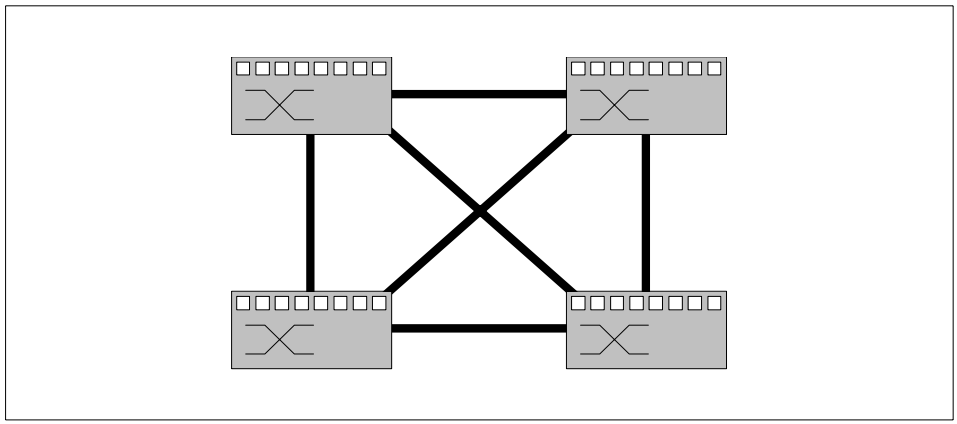


Figure 2-20 A simple meshed fabric

As all switches are connected to each other, there is a good deal of resilience. There are several paths between each pair of switches, see Figure 2-21.

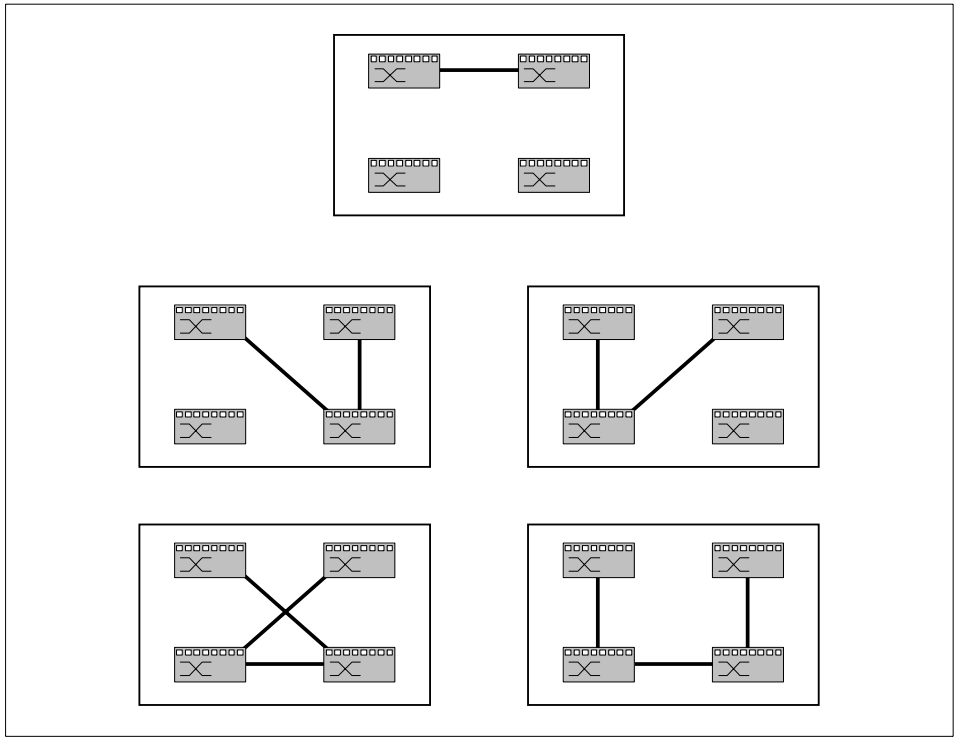


Figure 2-21 Alternative paths between meshed switches

As we can see, in this example of a simple mesh of four switches, there are five possible paths between each pair of switches. One path (the one at the top in our diagram) takes just one hop, two of the paths (the ones in the middle) take two hops. The other two paths take three hops.

There are special steps that the fabric can take to ensure that the frames will be routed over the shortest path (see 2.7 “Fabric Shortest Path First” on page 78). In the event of a failure of the shortest path, the fabric can automatically reconfigure to allow one of the next best paths to be used.

There is a big drawback. As all switches are connected to all other switches, full meshes use up switch ports rapidly. If we assume that we are using 16-port switches, we can use Table 2-1 to see how switch ports are used up by ISLs.

Table 2-1 Switch ports in a full mesh using 16-port switches

Switches	Total ports	Available ports/switch	ISLs	Available Ports
1	16	16	0	16
2	32	15	1	30
3	48	14	3	42
4	64	13	6	52
5	80	12	10	60
6	96	11	15	66
7	112	10	21	70
8	128	9	28	72
9	144	8	36	72
10	160	7	45	70
11	176	6	55	66
12	192	5	66	60
13	208	4	78	52
14	224	3	91	42
15	240	2	105	30
16	256	1	120	16
17	272	0	136	0

As we can see, using 16 port switches and ISLs in a full mesh topology will increase the available port count until we get to a maximum number of ports with eight switches. From then on, if we add extra switches, we actually decrease the number of available ports. The most cost effective configurations of fully meshed 16 port switches can be seen to be below six or seven switches.

Note: In Table 2-1:

- A one switch mesh is not really a mesh at all.
- A two switch mesh is really a cascade of two switches.
- A three port switch is just a ring of three switches.

Partial mesh

In this case, not all of the switches have connections to all of the other switches in the fabric. Or put another way, it is like a full mesh with some of the ISLs removed. This gives less resilience, and less alternative paths, but this is not necessarily a problem. The solution also allows a much larger fabric to be created.

Principal and subordinate switches

In any multi-switch fabric, there will be one principal switch and all other active switches will be subordinate switches. It is the principal switch which assigns the domain IDs to the subordinate switches. See 2.4.3 “24-bit port addresses” on page 72, for more details about what a Domain ID is.

When a switch is connected to a fabric, it carries out an exploration of the fabric. If it finds any other switches connected to it, they try to determine which one should be the principal switch. This may or may not be possible. The mechanism is implementation specific and this leads to problems of interoperability. It is even possible for two identical devices to have problems. It is important to check with the manufacturer of all switching devices in a fabric that they will interoperate, and if so, whether there are any caveats to consider.

2.2.3 Dense Wavelength Division Multiplexing (DWDM)

Dense Wavelength Division Multiplexing (DWDM) allows several fiber optical signals to be multiplexed and sent over the same fiber optic cable at long distances reducing cabling requirements.

For example, a device implementing DWDM may use, for example, 32 individual wavelengths or optical channels. The 32 optical channels are converted into electrical signals. The electrical signals are then converted back to different wavelength fiber optic signals, and transmitted over a common fiber link to the remote site. At the receiving end, the signals are optically filtered, converted back to the original signal type, and sent to the connecting device.

Because of the fact that this technology is used over long distances, it is implicit that DWDM utilizes long wave lasers and single mode fibers.

A comprehensive description of DWDM and the components that are likely to be encountered in the SAN environment can be found in:

- *Introduction to SAN Distance Solutions*, SG24-6408

2.3 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. 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.

2.3.1 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 2-22.

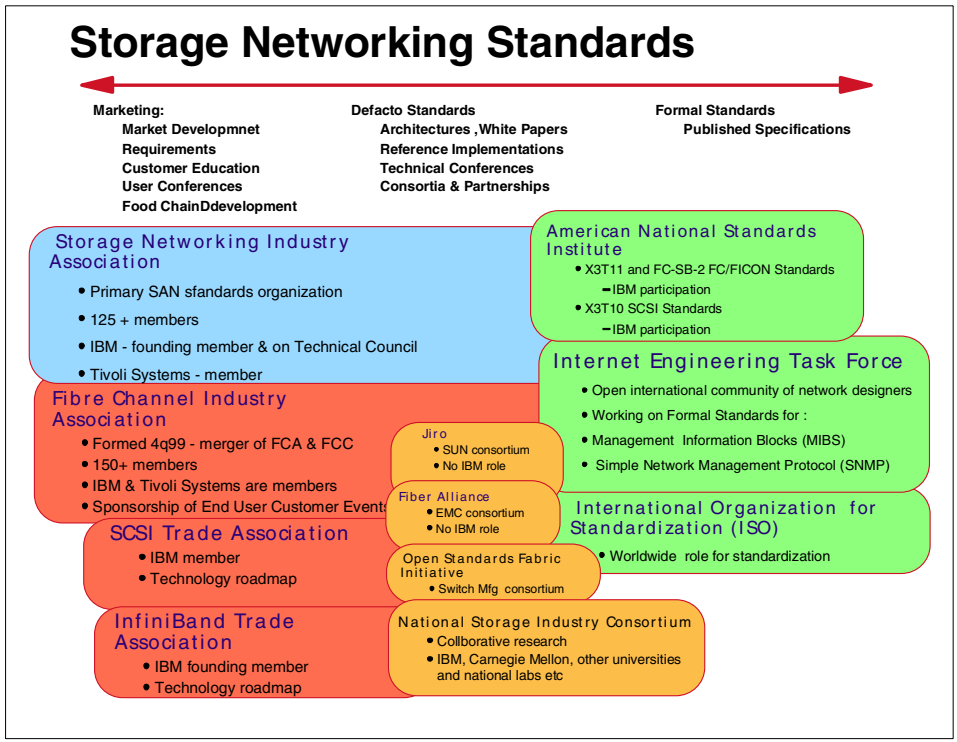


Figure 2-22 Groups involved in setting Storage Networking Standards

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.

2.3.2 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

2.3.3 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. For additional information on the various activities of FCIA, see its Web site at:

www.fibrechannel.com

2.3.4 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

2.3.5 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

2.3.6 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

2.3.7 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

2.3.8 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.org

2.4 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.

2.4.1 World Wide Name

All Fibre Channel devices have a unique identity. This is 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.

2.4.2 Port address

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. There needs to be some relationship between this 24-bit address and the 64-bit address associated with World Wide Names. We explain this in the section that follows.

The 24-bit address scheme also 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.

2.4.3 24-bit port addresses

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 Simple Name Server (SNS).

The Simple 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 2-23.

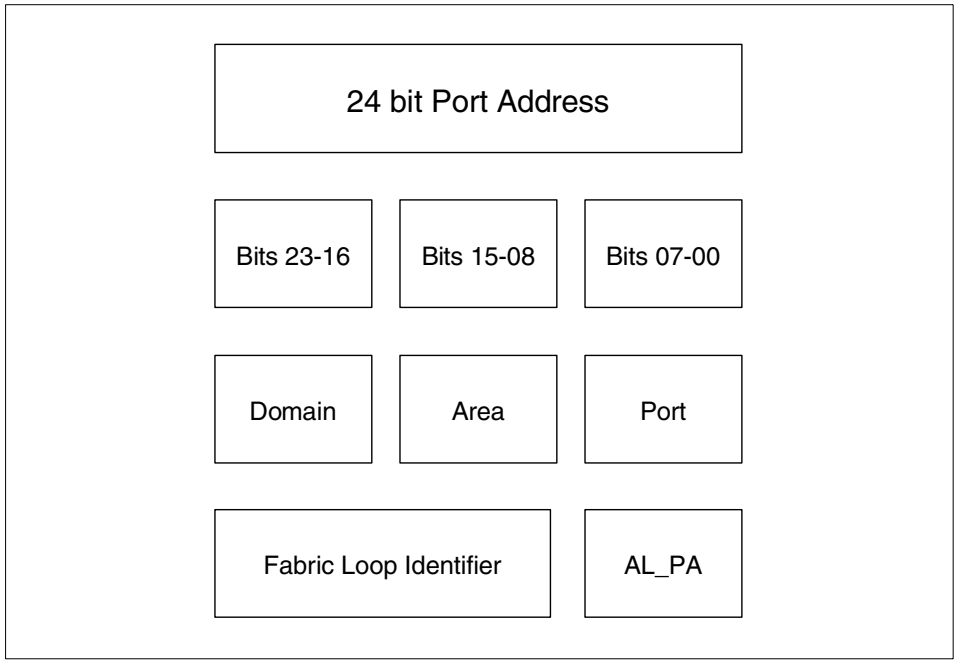


Figure 2-23 A Fibre Channel 24-bit address

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 “2.4.4 “Loop address” on page 74”.

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.

2.4.4 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.

2.5 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.

2.5.1 Management service

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.

2.5.2 Time service

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

2.5.3 Name services

Fabric switches implement a concept known as the Simple Name Server or SNS. 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.

2.5.4 Login service

In order to do a fabric login, a node communicates with the login server at address 0xFFFFFE. For more details see 2.6.1 “Fabric login” on page 76.

2.5.5 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.

2.6 Logins

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

- Fabric login
- Port login
- Process login

2.6.1 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.

2.6.2 Port 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.

2.6.3 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.

2.7 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.

2.7.1 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.

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 2-24, 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.

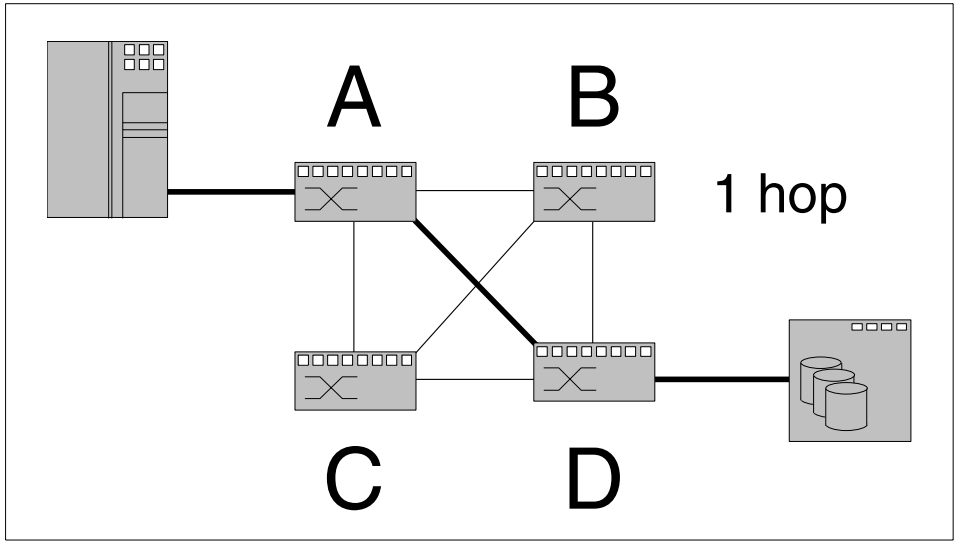


Figure 2-24 FSPF calculates the route taking the least hops

The other possible paths are shown in Figure 2-25.

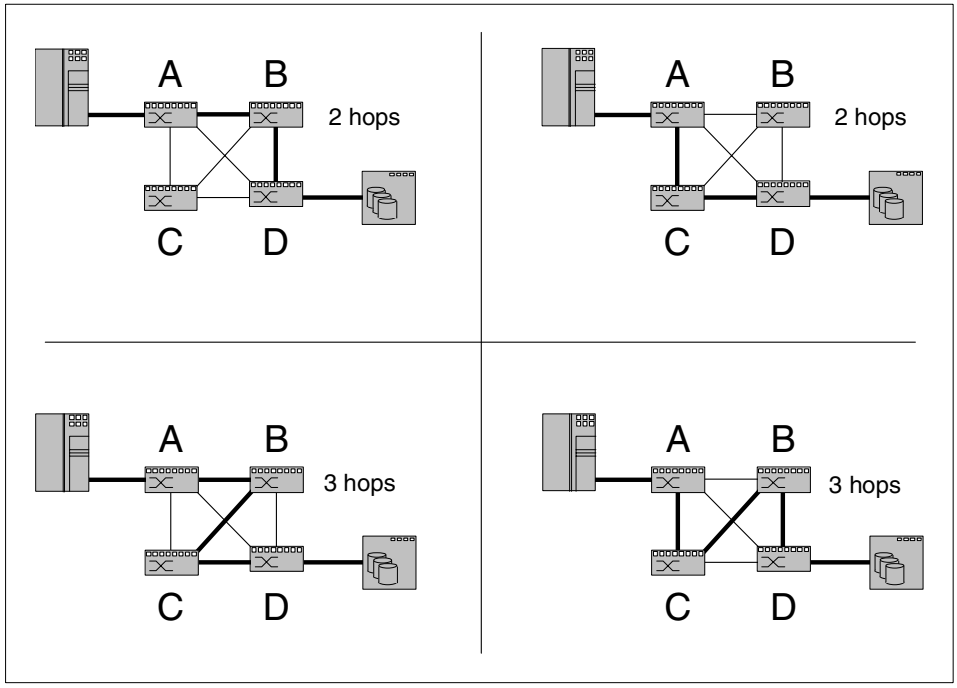


Figure 2-25 Other possible paths

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.

2.7.2 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 2.5.5 “Registered State Change Notification” on page 76) 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.

2.7.3 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.

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

If we look again at the example in Figure 2-24, 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 2-26.

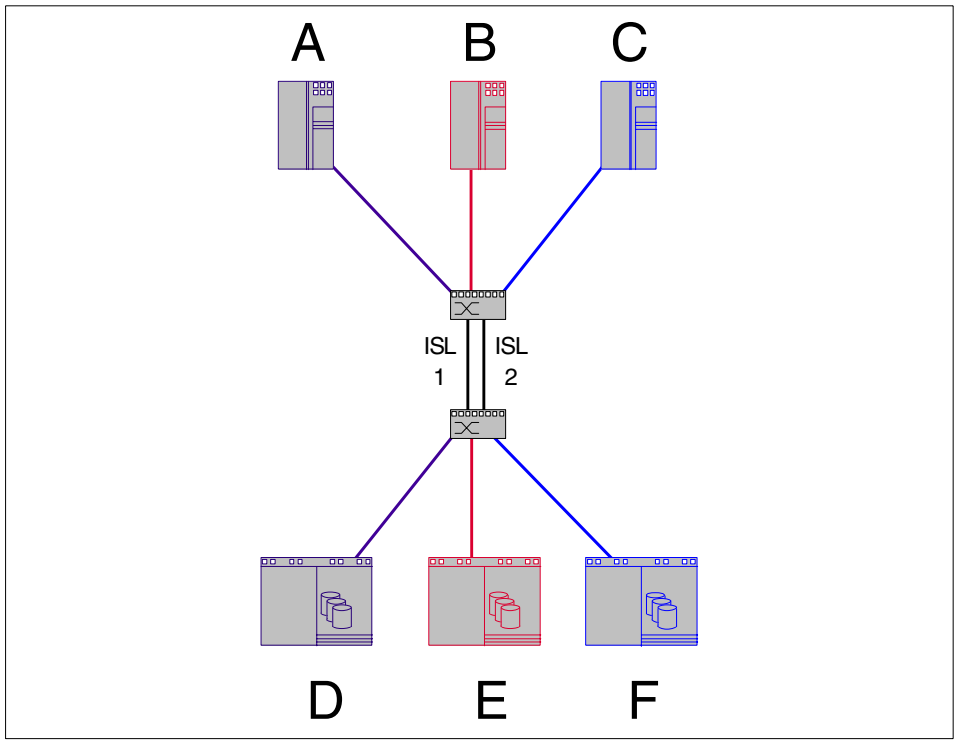


Figure 2-26 FSPF and round robin

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.

2.7.5 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 2-27.

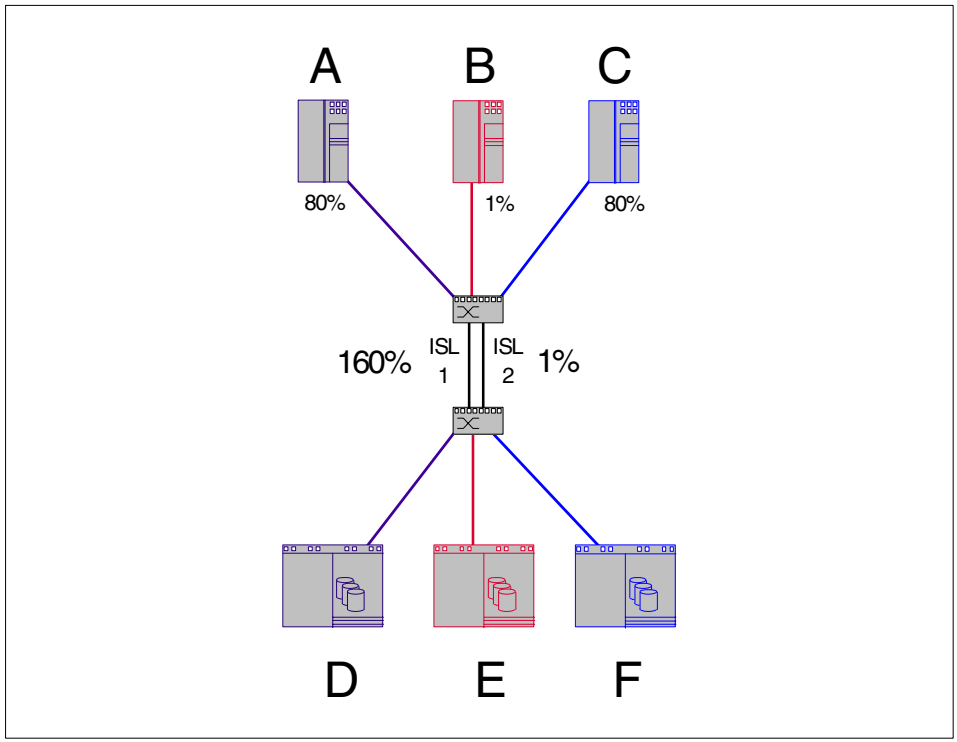


Figure 2-27 FSPF can lead to oversubscription and congestion

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 2.8 “Zoning” on page 93. Trunking is discussed in 2.9 “Trunking” on page 97.

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 2-28, however.

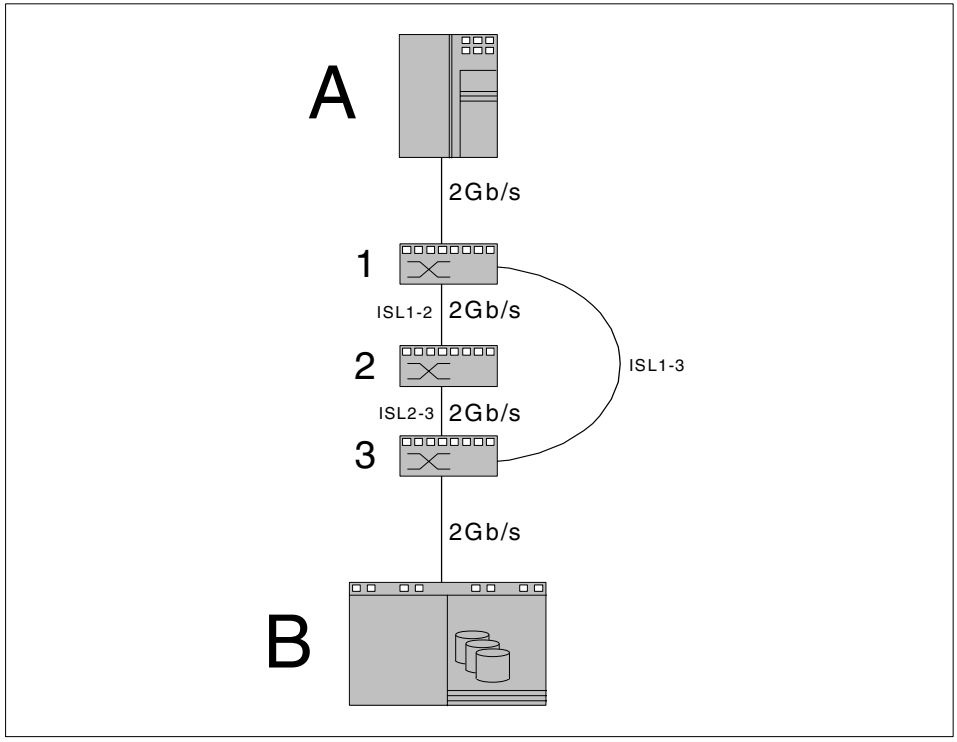


Figure 2-28 Hops and their cost, speed

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 2.9 “Trunking” on page 97, and 2.7.7 “1 Gb/s, 2 Gb/s and beyond” on page 86) 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

2.7.6 100 MB/s

Present day Fibre Channel devices generally operate at 1 Gb/s, that is, at one gigabit per second. What does this actually mean?

Data flows from the transducer (GBIC, GLM, transceiver) at a rate of 1 Gb/s. so, for example, in the case of a GBIC we actually tend to have a dataflow of 1,062,500,000 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 106,250,000 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
  - 1,062,500,000 bits per second
  - 106,250,000 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 2-2.

Table 2-2 Application bytes per second on 1 Gb/s link using 2048-byte payload

Acknowledged	Overhead Bytes	Bytes/sec	Megabytes/second
No	60	103,225,806	98.44
Yes	120	100,369,003	95.71

So, when the term 100 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 100 MB/s whether frames are acknowledged or not. If you prefer real computer (1024 x 1024) megabytes, then the maximum possible throughput is about 95.7 MB/s for an acknowledged class of service. This is still not a bad throughput, but, arguably, not 100 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 80 MB/s but this is by no means guaranteed, it might be higher or it might be lower.

2.7.7 1 Gb/s, 2 Gb/s and beyond

The FC-PH document describes and defines communication at 1 Gb/s, but 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 1 Gb/s is standard at the time of writing, 2 Gb/s is an emerging speed and manufacturers are beginning to implement it on their new products, or by making upgrades to their existing ones. The particular way that 2 Gb/s is implemented will vary from case to case. Manufacturers are beginning to market equipment at 4 Gb/s and some manufacturers mention that they are working towards 10 Gb/s.

The trend with 2 Gb/s is to use the SFF and LC form of connections, see “Small Form Factor Optical Transceivers” on page 48, and “LC connectors” on page 45.

Interoperability

As the transition is made from 1 Gb/s to 2 Gb/s a natural consideration is whether or not it is possible, and indeed sensible, 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, but as the technology surrounding 2 Gb/s is still emerging, it would be good advice to prove the concept of a particular solution or consult a suitable reference site before trying to incorporate it into a business critical SAN.

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 2-29.

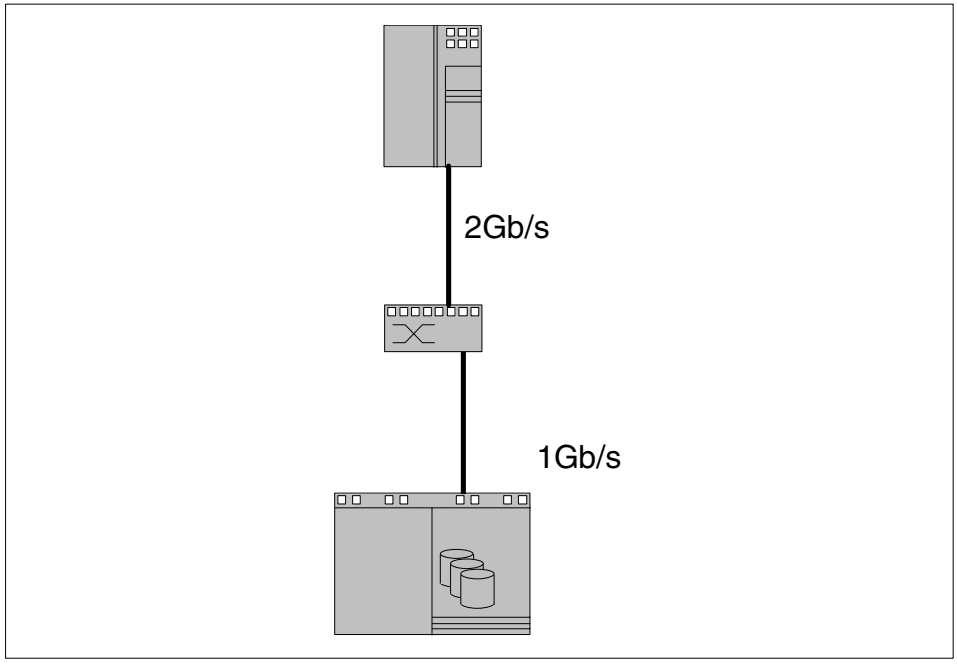


Figure 2-29 Mixing 2 Gb/s and 1 Gb/s

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 (see “Flow control” on page 114).

2.7.8 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 2.7.10 “Layers” on page 90). 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 2.2.1 “Classes of service” on page 56).
- The signalling layers (see also “Physical and signaling layers” on page 91)
- Speeds up to and including 100 MB/s
- Frames, Sequences, and Exchanges (see also 2.12 “Ordered Set, Frames, Sequences, and Exchanges” on page 108.)

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 bi-directional transfer.

Fractional bandwidth
This is Class 4 and is discussed in “Class 4” on page 57.

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 57.

2.7.9 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.

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.

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

2.7.10 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 parts:

- Physical and signaling layer
- Upper layer

The five layers are illustrated in Figure 2-30.

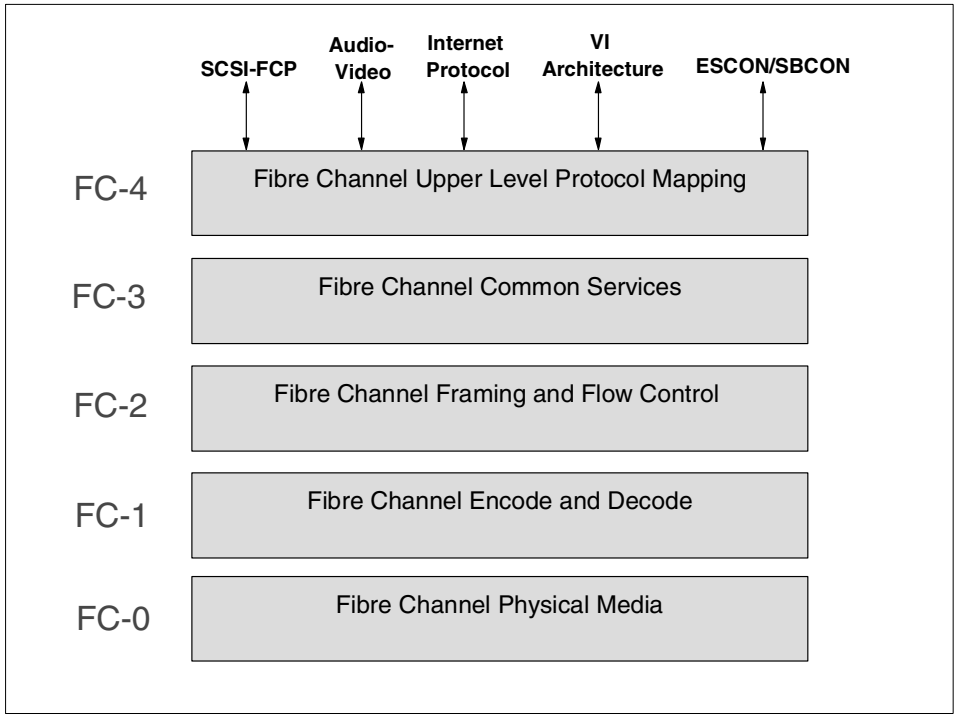


Figure 2-30 Fibre Channel layers

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 developing in popularity. The term is a contraction of FIbre CONnection and refers to running ESCON traffic over Fibre Channel

2.8 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.

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 2-31. 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.

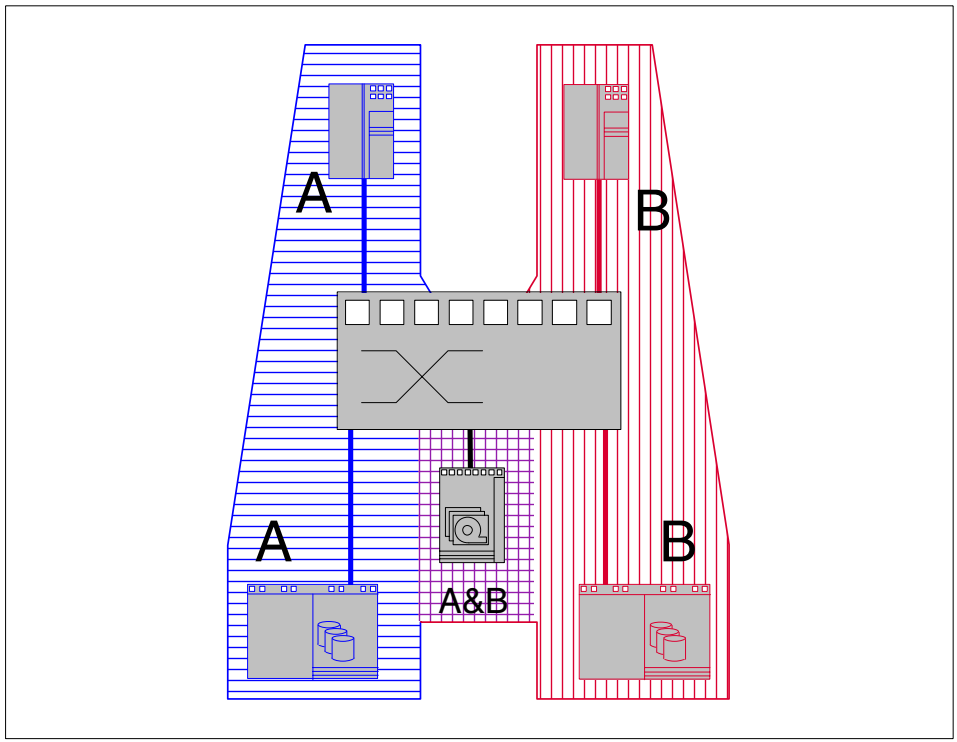


Figure 2-31 An example of zoning

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.

2.8.1 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

A single port can also belong to multiple zones. We show an example of hardware zoning in Figure 2-32.

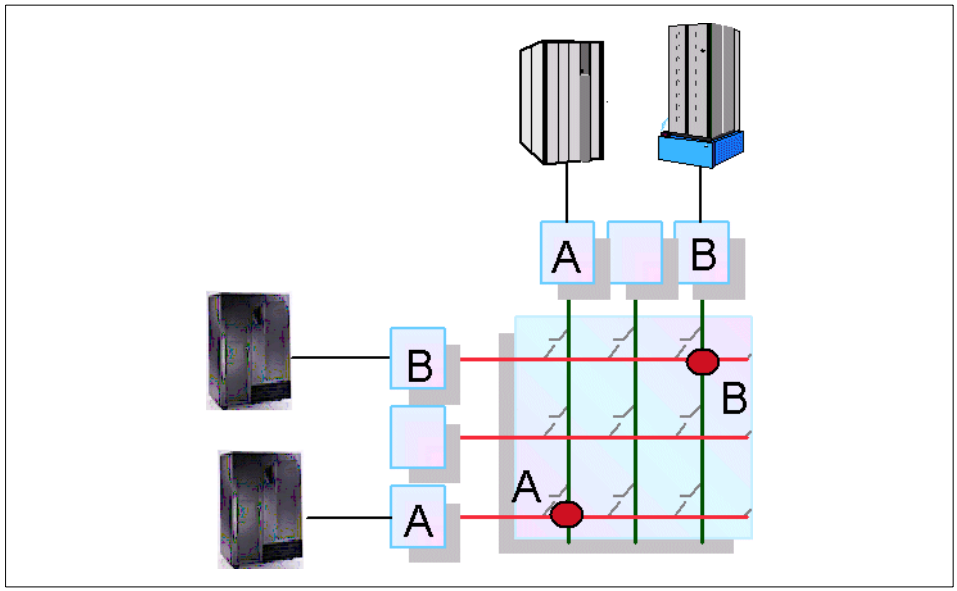


Figure 2-32 Hardware zoning

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.

2.8.2 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.

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. As the WWNs are used 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.

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.

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.

For maximum security, hardware zoning is recommended. But as the standards are evolving, and the industry is following them, it is likely that in the future, software zoning will probably be the preferred solution.

2.9 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 2-33.

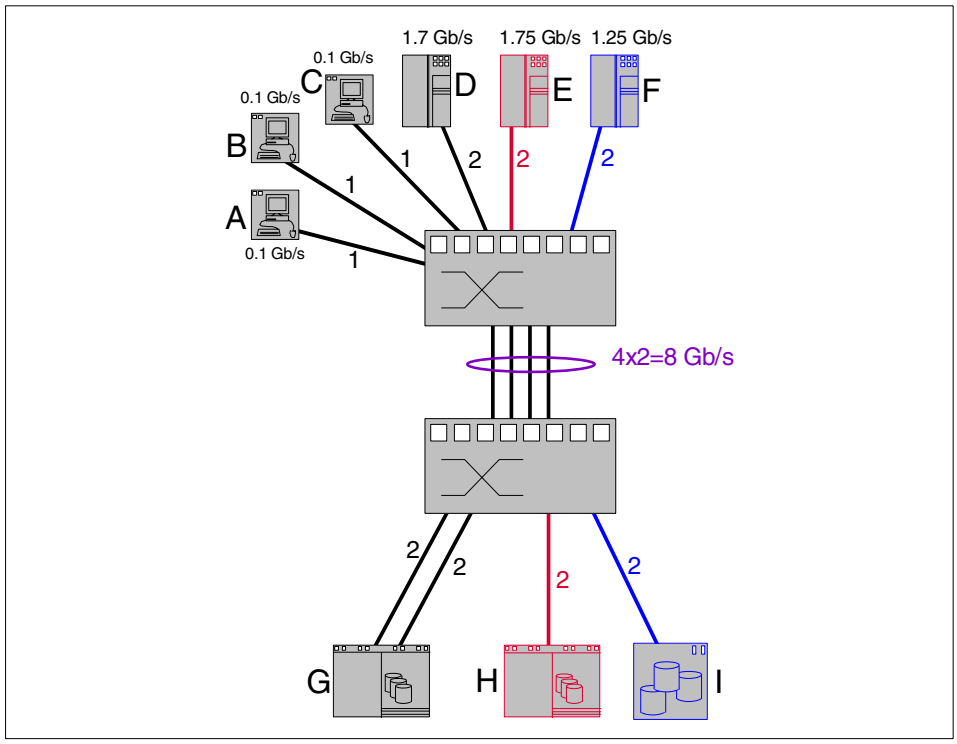


Figure 2-33 Trunking

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.

2.10 Logical unit number

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 possible for a server to see the same LUN more than once if there is more than one path from the server to the storage, see 2.11 “Multipathing” on page 99.

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. For this reason LUN masking becomes useful see “LUN masking” on page 129.

2.11 Multipathing

The idea of dual pathing, or multipathing in general, is to provide for a higher bandwidth so more data transfers can take place simultaneously and also to maintain data availability in case of path failures.

The way that a server deals with multipathing depends upon the particular operating system.

Multipathing is common in the S/390 environment and the channel subsystem takes care of it.

In some open systems, the operating system is able to handle multiple paths to a device by the implementation of a multipath aware fabric device driver in the kernel, an example is IBM’s DYNIX/ptx which runs on IBM e-server xSeries 430 and NUMA-Q equipment.

In most open system environments, we get an instance of each device on each path, so the appropriate software is required in addition to the Operating System kernel to handle multipath configurations.

Since a SAN typically provides more than one path between a server and a storage device, multipath software is, therefore, usually required.

Different vendors provide their own version of multipath software like Veritas Volume Manager. DMP, or HP PVLinks.

IBM has been offering Data Path Optimizer (DPO) AIX (5648-B58) and Windows NT (5639-F97). For all IBM TotalStorage Enterprise Storage Server (ESS) customers DPO has been superseded by the IBM Subsystem Device Driver (SDD).

2.11.1 IBM Subsystem Device Driver

The IBM Subsystem Device Driver (SDD) resides in the host server with the native disk device driver for the ESS. It uses redundant connections between the host server and disk storage in an ESS to provide enhanced performance and data availability.

The IBM Subsystem Device Driver provides the following functions:

- Enhanced data availability
- Automatic path failover and recovery to an alternate path
- Dynamic load balancing of multiple paths
- Path selection policies for the AIX operating system
- Concurrent download of licensed internal code

In most cases, host servers are configured with multiple host adapters with SCSI or Fibre Channel connections to an ESS that, in turn, provides internal component redundancy. With dual clusters and multiple host interface adapters, the ESS provides more flexibility in the number of I/O paths that are available.

When there is a failure, the IBM Subsystem Device Driver reroutes I/O operations from the failed path to the remaining paths. This function eliminates the following connections as single points-of-failure: a bus adapter on the host server, an external SCSI cable, a fiber-connection cable, or a host interface adapter on the ESS. This automatic switching in case of failures is called path failover.

In addition, multi-path load balancing of data flow attempts to prevent a single path from becoming overloaded with I/O operations, by using a round robin approach.

Concurrent Licensed Internal Microcode (LIC) installation

Concurrent download of licensed internal code is the capability to download and install licensed internal code onto an ESS while applications continue to run. During the time when new licensed internal code is being installed in an ESS, the upper-interface adapters inside the ESS may not respond to host I/O requests for approximately 30 seconds. The IBM Subsystem Device Drivers makes this transparent to the host through its path selection and retry algorithms.

Path algorithms

The path algorithms basically work the same for all the platforms the Subsystem Device Driver runs on. There are two modes of operation:

Single-path mode

The host server has only one path configured to an ESS logical unit number (LUN). The Subsystem Device Driver in single-path mode has the following characteristics:

- When an I/O error occurs, the I/O is retried a sufficient number of times to bypass the interval when the ESS upper-interface adapters are not available.
- This path is never put into the *dead* path mode.

Multiple-path mode

The host server has multiple paths configured to an ESS LUN. The Subsystem Device Driver in multiple-path mode has the following characteristics:

- If an I/O error occurs on a path, the Subsystem Device Driver does not attempt to use the path again, until 2000 successful I/Os have been performed on an operational path. This process is known as bypassing a path. The Subsystem Device Driver bypasses a failing path twice (until the I/O error count reaches three) and then the path is put into the *dead* state. After the path is put into the dead state, the Subsystem Device Driver uses this same bypass algorithm an additional two times.
- After the Subsystem Device Driver puts a path into the dead state, it puts the path into the *open* state after a certain number of successful I/Os have completed on an operational path. This number is operating system specific. If the first I/O operation fails after the path is put back into the open state, the Subsystem Device Driver puts the path into the dead state immediately and permanently. When fixed you must manually bring the path online by using the **datapath** command.
- If an I/O error occurs on the last operational path to a device, the Subsystem Device Driver attempts to reuse (or fail back to) a previously failed path. Only after a fixed number of fail back attempts is the operational path placed permanently offline.

- Host servers with only one operational path are in single-path mode.
- If an I/O error occurs on all the paths to a LUN, the Subsystem Device Driver returns an I/O error back to the application.

Currently IBM SDD supports AIX, Windows NT and 2000, Solaris, and HP-UX. For the specific version that are supported, and additional information about SDD refer to the following Web site:

http://www.ibm.com/storage/support/techsup/swtechsup.nsf/support/sddupdates

Linux

Due to the changing nature of Linux, it is difficult to list what is supported. One might almost say that more devices will be supported at the end of this sentence than were supported at the beginning of it.

What we can say, though, is that IBM is working to port SDD to Linux and this will be available in the near future.

2.11.2 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.

2.11.3 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.

2.11.4 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.

2.11.5 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 92. The Fibre Channel sequence is described in 2.12 "Ordered Set, Frames, Sequences, and Exchanges" on page 108.

2.11.6 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 2.11.7 “Data encoding” on page 104, 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.

2.11.7 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 2-34.

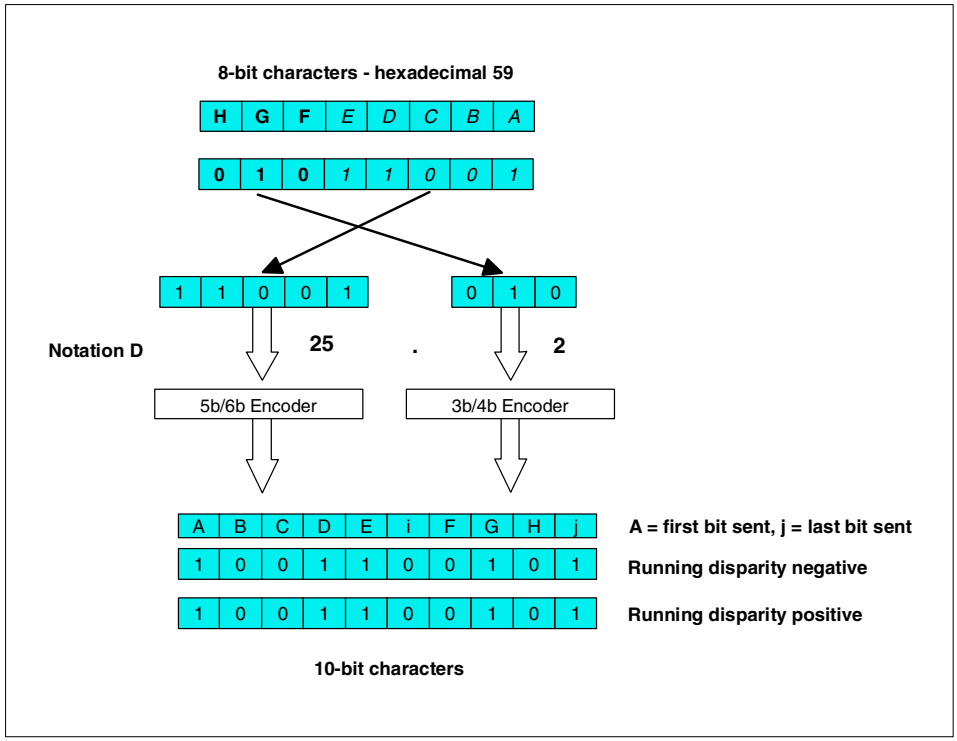


Figure 2-34 8b/10b encoding logic

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 2-3 below.

Table 2-3 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.

2.12 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.

2.12.1 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.

2.12.2 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 2-4 “Transmission words in a frame” on page 110.) 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.

2.12.3 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.

2.12.4 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.

2.12.5 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 2-4.

Table 2-4 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 2-5. The abbreviations are explained below the table.

Table 2-5 The frame header

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

2.12.6 “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.

2.12.7 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

2.12.8 Time-outs

It is necessary to consider various time-outs when designing a SAN. There are:

- Upper Layer Protocol time-outs, such as SCSI
- Application time-outs - such as DB2
- Fibre Channel time-outs

Whilst the first two are, of course, important to consider, the discussion of them falls outside the scope of this Redbook.

The FC-PH standard defines three time out values used for error detection and recovery:

R_T_TOV

This is the Receiver Transmitter time out value. It is used by the receiver logic to detect Loss of Synchronization with the transmitter. It has a fixed value of 100 ms.

E_D_TOV

This is the Error Detect time-out value. It represents the period in which a response should come back for a timed event. For example, during data transmission it represents a time-out value for a data frame to be delivered, the receiving port to transmit a response and the response be received by the initiator. E_D_TOV can normally be configured. The selected value should consider configuration and switch characteristics.

E_D_TOV is used in class of services 1 and 2, since, Class 3 does not check for acknowledgment.

A typical value for E_D_TOV is 2000 ms.

R_A_TOV

This is the Resource Allocation time-out value. It is used as a time out value during the recovery process. It should be set to E_D_TOV plus twice the maximum time a frame may be delayed within a fabric and still be delivered.

A typical value for R_A_TOV might be 5000 or 10000 ms.

Time out value settings

Without entering into the details of error detection and recovery, it is important to know the consequences of a wrong time-out value setting. Small E_D_TOV values may affect performance due to sequences being timed out and retried when they can still be correctly finished; too small R_A_TOV values may cause duplicated frames during recovery. On the other hand if the values are too long, error detection and recovery may be delayed when it is really needed.

Switch manufacturers provide default values that should work fine for normal distances (up to 10 Km). Delay considerations should be taken into account for extended distances. Each kilometer of fiber adds approximately 5 microseconds delay. Also the delay introduced by repeaters or extenders should be considered.

2.12.9 Buffers and credits

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 fabric login, N_Ports and F_Ports at both ends of a link establish its Buffer to Buffer Credit (BB_Credit). Each port states the maximum BB_Credit that they can offer and the lower of the two is used.

EE_Credit

In the same way during port login all N_Ports establish End to End Credit (EE_Credit) with each other.

Flow control

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, both are initialized to 0 during login.

Each time a port sends a frame it increments BB_Credit_CNT and EE_Credit_CNT by one. When it receives R_RDY from the adjacent port it decrements BB_Credit_CNT by one, when it receives ACK from the destination port it decrements EE_Credit_CNT by one. 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.

Performance

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 between about 75 m and 4 Km of fiber depending on the size of the data payload. In a 100 Km link we could send many frames before the first one reaches destination. We need an ACK back to start replenishing EE_Credit or an R_RDY to replenish BB_Credit.

For a moment, let us consider frames with 2 KB of data. These occupy approximately 4 Km of fiber. We will be able to send about 25 frames before the first arrives at the far end of our 100 Km link. We will be able to send another 25 before the first R_RDY or ACK is received, so we would need at least 50 buffers to allow for non stop transmission at 100 Km distance with frames of this size. If the frame size is reduced, more buffers would be required to allow non-stop transmission.

2.12.10 Ports

In the various discussions in this redbook, we will mention different kinds of Fibre Channel ports. It is, therefore, important to understand what is meant by these different types of ports.

E_Port

An E_Port is an expansion port. A port is designated an E_Port when it is used as an interswitch expansion port to connect to the E_Port of another switch, to build a larger switched fabric. These ports are found in Fibre Channel switched fabrics and are used to interconnect the individual switch or routing elements. They are not the source or destination of IUs, but instead function like the F_Ports and FL_Ports to relay the IUs from one switch or routing elements to another. E_Ports can only attach to other E_Ports.

An Isolated E_Port is a port that is online but not operational between switches due to overlapping domain ID or nonidentical parameters such as E_D_TOVs see 2.12.8 “Time-outs” on page 113.

F_Port

An F_Port is a fabric port that is not loop capable. Used to connect an N_Port to a switch. These ports are found in Fibre Channel switched fabrics. They are not the source or destination of IUs, but instead function only as a “middle-man” to relay the IUs from the sender to the receiver. F_Ports can only be attached to N_Ports.

FL_Port

An FL_Port is a fabric port that is loop capable. Used to connect NL_Ports to the switch in a loop configuration. These ports are just like the F_Ports described above, except that they connect to an FC-AL topology. FL_Ports can only attach to NL_Ports.

G_Port

A G_Port is a generic port that can operate as either an E_Port or an F_Port. A port is defined as a G_Port when it is not yet connected or has not yet assumed a specific function in the fabric.

L_Port

An L_Port is a loop capable fabric port or node. This is a basic port in a Fibre Channel Arbitrated Loop (FC-AL) topology. If an N_Port is operating on a loop it is referred to as an NL_Port. If a fabric port is on a loop it is known as an FL_Port. To draw the distinction, throughout this book we will always qualify L_Ports as either NL_Ports or FL_Ports.

N_Port

N_Port is a node port that is not loop capable. Used to connect an equipment port to the fabric. These ports are found in Fibre Channel nodes, which are defined to be the source or destination of information units (IU). I/O devices and host systems interconnected in point-to-point or switched topologies use N_Ports for their connection. N_Ports can only attach to other N_Ports or to F_Ports.

NL_Port

An NL_Port is a node port that is loop capable. Used to connect an equipment port to the fabric in a loop configuration through an FL_Port. These ports are just like the N_Port described above, except that they connect to a Fibre Channel Arbitrated Loop (FC-AL) topology. NL_Ports can only attach to other NL_Ports or to FL_Ports.

U_Port

U_Port is a universal port. A generic switch port that 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.

In addition to these Fibre Channel port types, the following port types are only used in the INRANGE products.

T_Port (INRANGE specific)

A T_Port is an ISL port more commonly known as an E_Port.

TL_Port (INRANGE specific)

A TL_Port is a private to public bridging of switches or directors.

In addition to these Fibre Channel port types, the following port type is used only in the McDATA products:

B_Port (McDATA specific)

A B_Port is a bridge port that provides fabric connectivity by attaching to the E_Port of a director. This B_Port connection forms an ISL through which a fabric device can communicate with a public loop device.

Figure 2-35 shows an example SAN with representations of some of the different Fibre Channel port types.

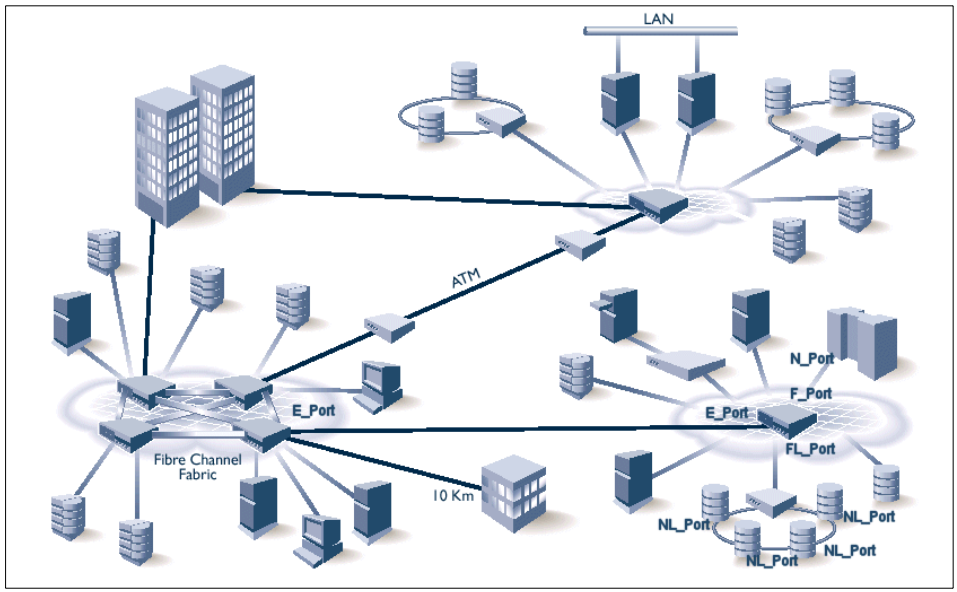


Figure 2-35 An example SAN showing FC port types

2.12.11 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.

2.12.12 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.

2.13 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.

An introduction to FC-AL is given in “Arbitrated Loop” on page 59.

2.13.1 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 2-36 we show a private loop.

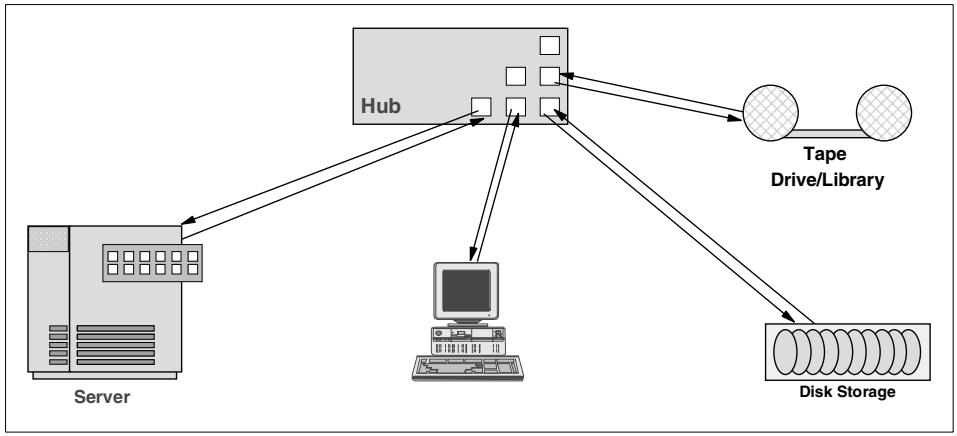


Figure 2-36 Private loop implementation

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 2-37 shows a public loop.

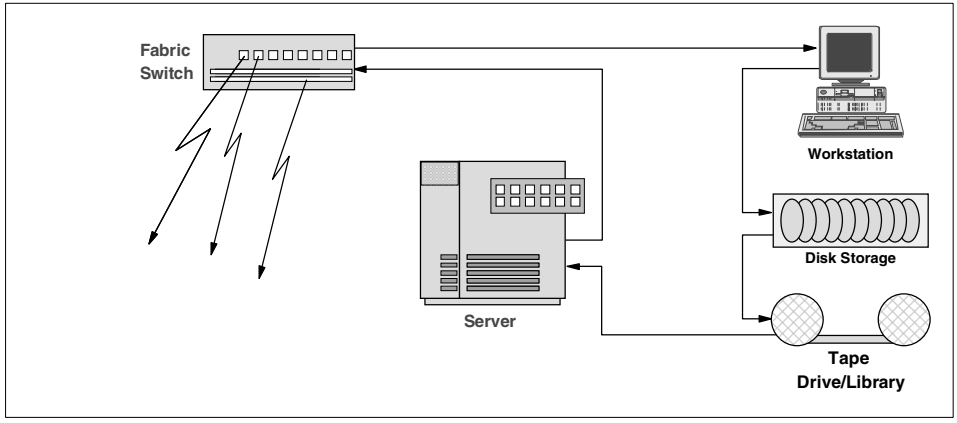


Figure 2-37 Public loop implementation

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.

2.13.2 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.

2.13.3 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.

2.13.4 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 2-38.

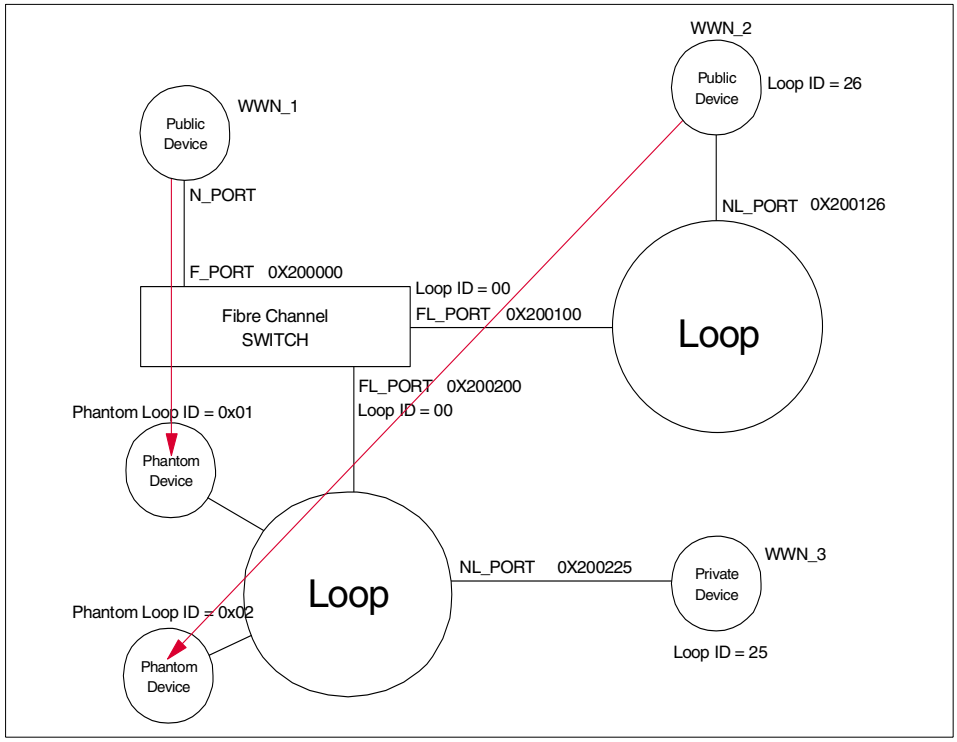


Figure 2-38 Arbitrated loop address translation

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 2-6.

Table 2-6 Simple 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 2-7, is created in the switch.

Table 2-7 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 2-7.

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 2.13.4 “Private devices on NL_Ports” on page 124.

2.14 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.

2.14.1 Limits

There are various limits encountered in Fibre Channel.

Distances

In Table 2-8 we show the copper and Fibre Channel limits.

Table 2-8 Copper and Fibre Channel limits

Type	Fiber type	Distance	Speed**
Extended LW Laser	Single Mode 9 µm	80-100Km*	100MB/s 200MB/s
LW Laser	Single Mode 9 µm	10Km	100MB/s 200MB/s
SW Laser	Multi Mode 50 µm	500m 300m	100MB/s 200MB/s
SW Laser	Multi Mode 62.5 µm	300m 150m	100MB/s 200MB/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 two speeds are specified, the actual speed will depend upon the GBIC or other transducer selected. The speeds are the nominal throughput of application data and the considerations discussed in 2.7.6 “100 MB/s” on page 84.

2.14.2 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 2.8.1 “Hardware zoning” on page 94).

LUN masking

One 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.

2.14.3 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

2.14.4 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 2-39.

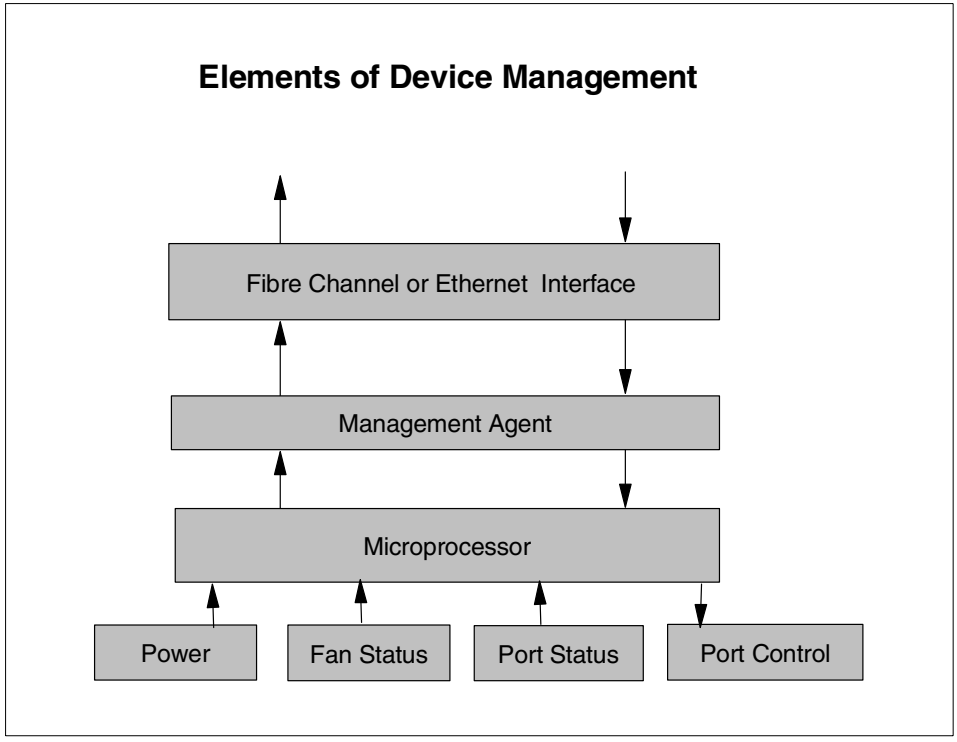


Figure 2-39 Device management elements

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.

2.14.5 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.

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 131, and inband in “Inband management” on page 131.

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 2-9 summarizes the management access methods available.

Table 2-9 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 2-40 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 2-40 illustrates the tree structure, with special attention to the internet MIB tree and the Fibre Channel MIB tree.

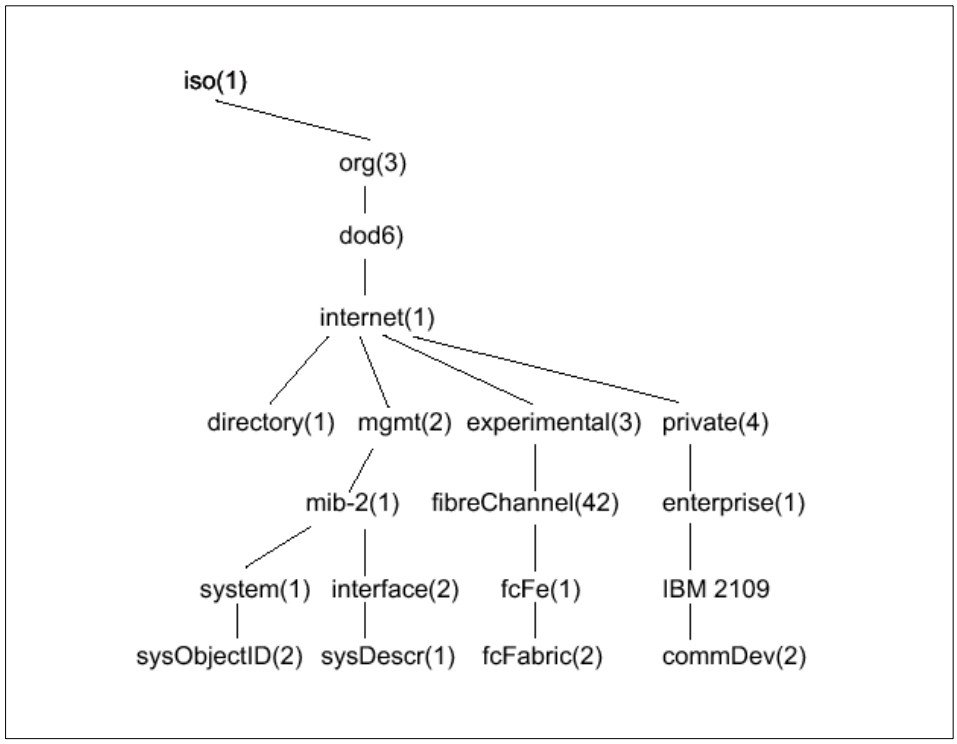


Figure 2-40 MIB tree

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 139.

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 139. 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.

2.14.6 Long distance links

The first thing to consider regarding a long distance link is that the technology selected will cope with the distance of the link, see 2.14.1 “Limits” on page 127.

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.

2.14.7 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.

2.14.8 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

SAN fabric products

In this chapter we will overview the products which represent the constituent pieces of the IBM SAN fabric jigsaw.

The products we will overview are:

- IBM SAN Data Gateway SCSI Tape Router
- IBM SAN Data Gateway
- IBM TotalStorage SAN Controller 160
- IBM Fibre Channel Storage Hub
- IBM TotalStorage SAN Managed Hub
- IBM TotalStorage SAN Switch F08
- IBM TotalStorage SAN Switch S08
- IBM TotalStorage SAN Switch S16
- IBM TotalStorage SAN Switch F16
- IBM TotalStorage SAN Switch M12
- INRANGE FC/9000 Fibre Channel Director
- McDATA ES-1000 Loop Switch
- McDATA ES-3016 Fabric Switch
- McDATA ES-3032 Fabric Switch
- McDATA ED-6064 Enterprise Fibre Channel Director

3.1 IBM SAN Data Gateway SCSI Tape Router

The IBM SAN Data Gateway Router is a SCSI to Fibre Channel protocol converter for tape libraries, with one Fibre Channel adapter, and up to two SCSI ports. It is a low-cost solution, compared to the IBM SAN Data Gateway product, which offers up to three FC x four SCSI ports configurations.

The IBM SAN Data Gateway Router (2108-R03) can accommodate either Ultra SCSI single-ended ports or Ultra SCSI differential ports.

The Router supports full mapping of SCSI IDs and LUNs between the Fibre Channel attached host and the SCSI tape library. The IBM SAN Data Gateway Router can be attached to an IBM Fibre Channel Switch for connectivity.

Fibre Channel attachment

Industry-standard Fibre Channel technology is rapidly replacing SCSI channel attachment between open system servers and tape storage systems. However, many tape storage systems do not provide Fibre Channel attachment. To bridge the gap between Fibre Channel server adapters and SCSI-attached tape storage, IBM has developed the Storage Area Network (SAN) Data Gateway Router.

The SAN Data Gateway Router is a hardware solution that enables the attachment of SCSI storage systems to Fibre Channel adapters on specific Intel-based servers running Windows NT and UNIX-based servers from IBM and Sun Microsystems.

For the most current list of supported products, visit:

www.ibm.com/storage/sangateway

The SAN Data Gateway Router with short-wave ports can provide Fibre Channel distance extension up to 500 meters between an open system server and a storage system. This is ideal for server and storage consolidation.

With IBM Fibre Channel Storage Hubs and Managed Hubs, and Fibre Channel Switches, connectivity options enable distances up to 10 kilometers and many server and storage connections. This any-to-any switched fabric capability supports large and rapidly growing storage consolidation and data sharing requirements.

IBM StorWatch SAN Data Gateway Specialist

The SAN Data Gateway Router provides access between its Fibre Channel ports and SCSI ports. Channel zoning controls access between ports. The IBM StorWatch SAN Data Gateway Specialist — an easy-to-use graphical user interface — includes the tools to define SAN Data Gateway Router channel zoning and to control access to specific storage devices.

Multiple configuration options

The SAN Data Gateway Router utilizes Fibre Channel and Ultra SCSI channel bandwidth for high-performance attachment of the following devices:

- Magstar 3590 Tape Subsystem in stand-alone, Magstar 3494 Tape Library, and Magstar 3590 Silo Compatible Tape Subsystem environments
- Magstar MP 3570 Tape Subsystem or Magstar MP 3575 Tape Library Dataserver
- IBM 3580 Ultrium** Tape Drive, 3584 UltraScalable Tape Library, 3583 Ultrium Scalable Tape Library, and 3581 Ultrium Tape Autoloader
- IBM 3502 DLT Tape Library

The SAN Data Gateway Router is available as a rack-mounted unit or as a stand-alone tabletop unit. The low-cost Router provides one short-wave Fibre Channel port and two Ultra SCSI Differential or Ultra SCSI Single-End ports for tape storage attachment.

3.2 IBM SAN Data Gateway

The IBM SAN Data Gateway (2108-G07) was one of the first components of the IBM SAN solution that allows an easy migration to the SAN environment using Fibre Channel technology. The SAN Data Gateway connects SCSI and Ultra SCSI storage devices to Fibre Channel environments. It attaches new or existing SCSI storage products to the SAN using an industry standard Fibre Channel arbitrated loop (FC-AL) interface. The SAN Data Gateway solves three immediate problems:

- The 25m cable length restriction for SCSI: the cable can extend up to 500m
- The increased bandwidth demand that Ultra SCSI storage products can place on the SCSI bus
- The address limitations of SCSI

The use of hubs in SAN configurations increases the device connectivity, but hubs have some impact with respect to multiple hosts on the FC-AL loop. These include loop initialization process and arbitration. If a system is turned off and then on, or rebooted, it might impact the operation of other systems in the FC-AL loop. Many integrators will not support multi-host loop at all.

The use of switches or directors increases the host fan-out which is another way of saying the number of host connections of SAN configurations.

The SAN Data Gateway utilizes Fibre Channel and Ultra SCSI channel bandwidth for high-performance attachment of the following devices:

- IBM Enterprise Storage Server
- IBM Magstar 3590 Tape Subsystem in stand-alone, Magstar 3494 Tape Library, and Magstar 3590 Silo Compatible Tape Subsystem environments
- IBM Magstar MP 3570 Tape Subsystem or Magstar MP 3575 Tape Library Dataserver
- IBM 3502 DLT Tape Library
- IBM Ultrium 358X Tape Subsystems with LTO Tape Drives

For the latest and most up-to-date list of supported servers, adapters, disk and tape subsystems on the SAN Data Gateway, visit:

http://www.storage.ibm.com/hardsoft/products/sangateway/supserver.htm

Sharing the Gateway between disk and tape products is currently not supported or practical, because:

- The Enterprise Storage Server needs all the SCSI attachments.
- The levels of the HBA driver required for disk and for tape are different, which makes it impossible to use Gateway-attached disks and tapes on the same host. This will eventually be fixed, but is a nice illustration of an interoperability problem.

The Gateway can either be used as a stand-alone table top unit or mounted in a standard 19" rack. The rack can be either the IBM 2101 Seascape Solutions rack or an industry standard rack.

The SAN Data Gateway is equipped with:

- Four Ultra SCSI differential ports
- One to six FC-AL short-wave and long-wave ports and Fibre Channel optic cables
- StorWatch SAN Data Gateway Specialist (included on CD)

Features and functions of the SAN Data Gateway:

- **SAN connectivity:** Creates reliable SAN solutions without needing hubs, switches and bridges. The SAN Data Gateway provides a distance or connectivity solution for SCSI attached storage devices.
- **Heterogeneous systems and storage:** Provides seamless support for different host platforms and multiple device types.
- **SAN resource sharing:** Zoning or partitioning enables a simple and effective resource sharing solution. Zones are created by controlling the access between different channels or ports and are implemented with the StorWatch SAN Data Gateway Specialist access control function.
- **SAN value added functions:**
  - Supports up to 256 LUNs across multiple interfaces
  - Persistent address maps are preserved in non-volatile memory
  - Full awareness of SCSI 3 protocol for disk and tape
  - SCSI over TCP for remote transfer, management and control. SCSI commands and data are encapsulated in TCP packets
  - Support for SNIA Extended Copy Command specification. This is the basis for server-free backup solutions in the future
- **Transparent SAN performance:** The total bandwidth of the SAN Data Gateway is 120 MB/s; The overall performance is driven by the maximum available device performance.
- **SAN Management:** The SAN Data Gateway is remotely managed and controlled by the StorWatch SAN Data Gateway Specialist.
- **SAN Scalability:** The SAN Data Gateway is offered with up to *six FC ports* to provide 6 x 4 configurations.

Zoning or access control

The SAN Data Gateway has the ability to connect to more than one host. In the default configuration, there is no restriction between the channels for access to the target devices. Without additional controls, host operating systems do not handle multiple systems using the same target devices simultaneously. The result is corrupted file systems when two hosts try to use the same disk drives or LUN. Or, tape backup and restore operations might be interrupted. The IBM StorWatch SAN Data Gateway Specialist Channel Access options can be used to disable access between the SAN Connections and individual SCSI channels.

IBM StorWatch SAN Data Gateway Specialist

The SAN Data Gateway provides access between its Fibre Channel ports and its SCSI ports. Channel zoning provides access control between ports. While channel zoning provides control of paths between host adapters and SCSI storage ports, it does not limit access to specific devices (LUNs) within the storage system. Virtual Private SAN (VP SAN) provides LUN masking to limit access between host adapters and LUNs attached to SAN Data Gateway SCSI ports. The IBM StorWatch SAN Gateway Specialist, an easy to use graphical user interface, provides the tools to define SAN Data Gateway channel zoning, the VP SAN LUN-masking, and control which host systems have access to specific storage devices.

This Access Control function, also called zoning, partitions the SAN configuration by either allowing or denying access between the FC and SCSI ports of the Gateway.

Advantages of SAN Data Gateway versus a hub

- Concurrency: Aggregate throughput is not limited to one loop.
- Zoning: Access control available based on FC and SCSI ports.
- Hosts are each point-to-point and they can be heterogeneous.
- Smaller configurations with fewer devices lower administration cost for customer and lower service/support cost for IBM (easier to isolate problems).
- Avoids the inherent multi-host issues of the FC-AL loop, such as loop initialization process (LIP) and arbitration. If a system is turned OFF/ON or rebooted it might impact the operation of other systems in the FC-AL loop. Many integrators will not support multi-host loop at all.

Advantages of SAN Data Gateway versus a switch

- Defers or completely avoids the high entry cost of a switch.
- Smaller configurations with fewer devices lower administration cost for customer and lower service/support cost for IBM (easier to isolate problems).
- Interoperability issues with switches: fabric support is limited, resource sharing requires middleware.

The IBM SAN Data Gateway is shown in Figure 3-1.

Figure 3-1 IBM SAN Data Gateway

3.3 IBM TotalStorage SAN Controller 160

The IBM TotalStorage SAN Controller 160 (7140-160) enables all IBM 7133, 7131, and 3527 Serial Disk Systems to attach to host systems using Fibre Channel host adapters and drivers.

The IBM TotalStorage SAN Controller 160 should be considered for configurations where native SSA connectivity is not possible, and where the RAID-1 data replication capability can provide added data protection. The SAN Controller 160 is designed to bring the performance, availability, and scalability advantages of the Serial Storage Architecture (SSA) to customers with new or existing fibre channel based host servers.

The SAN Controller replicates data across or within serial disk systems; simultaneously mirroring two or three copies of data without host involvement. With global hot disk sparing, data is automatically rebuilt if a mirrored disk fails. In this way, the SAN Controller improves performance and data availability while simplifying storage operations.

The Instant Copy function can create a separately addressable copy of mirrored data that can be used for tape backup. After the backup has completed, data is re-synchronized with the primary copy. To support remote storage operations, mirrored 7133 Advanced Models D40 or T40 can be separated by up to 10 km with serial storage fiber optic extenders.

With 36.4 GB disks, logical volume groups or partitions as large as 580 GB can be created for Windows NT servers, which have limited volume addressing. The SAN Controller can also create composite drives by concatenating up to 16 physical disks. These capabilities provide excellent configuration flexibility for growing storage environments.

Simplified management
The SAN Controller 160 Manager is a Windows NT based management tool that provides configuration and service functions, including mirror group definition, the ability to create composite drives, and Instant Copy disk management. The Manager can manage multiple SAN Controllers across the enterprise.

A highly scalable solution
The SAN Controller supports up to 64 serial disk drives in a single loop and enables non-disruptive growth in disk capacity from 18.2 GB to 2.3 TB. Up to eight UNIX and Windows NT host systems can be attached to a single loop. Performance scales up as more SAN Controllers are added to the serial loop.

A standalone, tabletop SAN Controller unit provides one fibre channel port and two SSA ports. Short and long-wave laser optical interfaces are supported. An optional rack-mounted enclosure can hold up to four SAN Controllers in a compact 2U-high space in an industry-standard 19-inch rack.

Fully utilized bandwidth potential
The IBM TotalStorage SAN Controller 160 enables fibre channel servers to benefit from high-performance, non-arbitrated serial disk technology. The fibre channel host-based adapter views the IBM TotalStorage SAN Controller 160 as a single FC-AL target, which minimizes loop arbitration overhead. It has measured up to 90 MB/s sustained throughput, or up to 25,000 I/Os per second per logical channel in full-duplex, simultaneous read/write mode.

3.4 IBM Fibre Channel Storage Hub

The IBM Fibre Channel Storage Hub (2103-H07) is an entry level component for SAN fabric installations designed for connecting one or more storage devices, to one or more servers. It is not a very scalable solution and should not be chosen if many devices are to be connected later. A hub can also be used to connect to a remote location to extend the distance.

The Fibre Channel Storage Hub is designed to provide a centralized point of connectivity, to provide loop fault tolerance, and to simplify configuration management.

Fibre Channel products that are commonly interconnected to the Fibre Channel Hub are Fibre Channel host bus adapters, FC-AL storage devices, and FC-AL storage arrays.

In terms of scalability of bandwidth, one FC-AL loop by itself is not scalable. All devices share the bandwidth of 100 MB/s, rather than that offered by the Managed Hub.

In Figure 3-2 we show the hub.

Figure 3-2 IBM Fibre Channel Storage Hub

3.4.1 Hub configuration

The IBM Fibre Channel Storage Hub interconnects multiple servers and storage systems, over fiber optic media, and transfers data at speeds up to 100 MB/s.

Each port requires a gigabit interface converter to connect it to each attached node. The Fibre Channel Storage Hub supports any combination of shortwave or longwave optical GBICs. We show a GBIC in Figure 3-3.

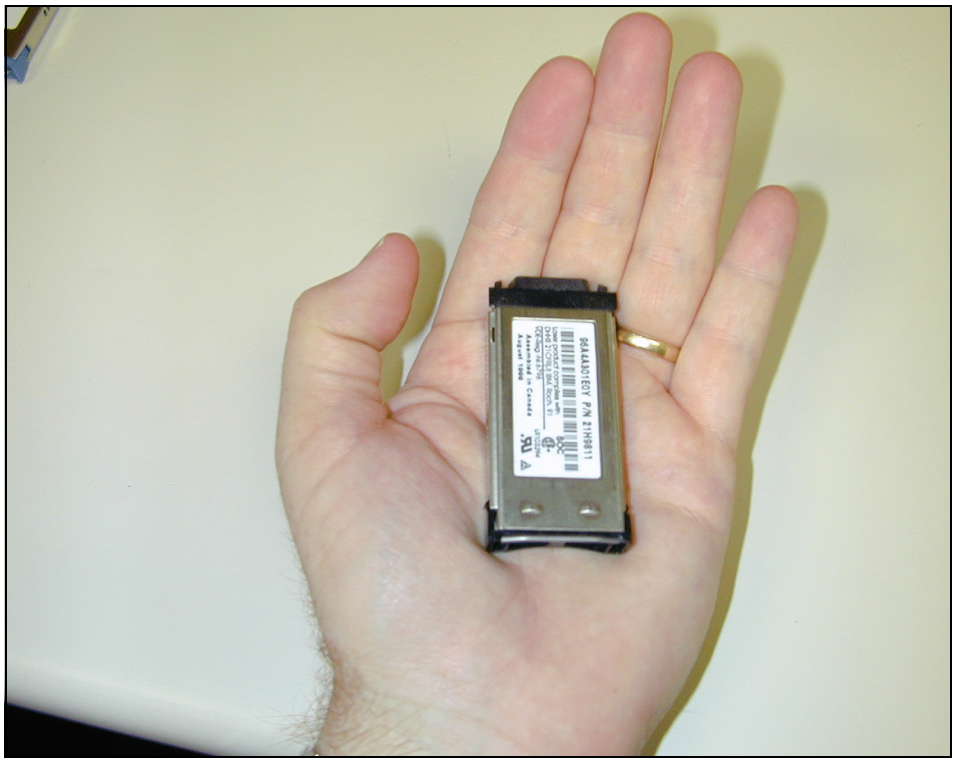


Figure 3-3 Gigabit interface converter

The GBICs are hot-pluggable into the IBM Fibre Channel Storage Hub, which means you can add host computers, servers, and storage modules to the arbitrated loop dynamically, without powering off the Fibre Channel Storage Hub or any connected devices.

If you remove a GBIC from a Fibre Channel Storage Hub port, that port is automatically bypassed. The remaining hub ports continue to operate normally with no degradation of system performance. Conversely, if you plug a GBIC into the Fibre Channel Storage Hub, it will automatically be inserted and become a node on the loop, if valid Fibre Channel data is received from the device.

3.5 IBM TotalStorage SAN Managed Hub

The IBM TotalStorage SAN Managed Hub (3534-1RU), is an 8-port Fibre Channel hub that consists of a system board with connectors for supporting up to eight ports. This includes seven fixed, short wavelength ports, one pluggable GBIC port, and an operating system for building and managing a switched loop architecture.

The Managed Hub is a non-blocking architecture that provides 8 x 100MB/sec throughput.

The hub is supported on IBM PC, Netfinity servers, and other Intel-based servers. Shown in Figure 3-4 is a picture of the Managed Hub with a shortwave GBIC and a longwave GBIC.

Figure 3-4 IBM TotalStorage SAN Managed Hub

The latest support matrix, including adapters and operating system requirements, can be found at the following Web page:

http://www.storage.ibm.com/hardsoft/products/fchub/msupserver.htm

In Figure 3-5, we show the faceplate of the IBM TotalStorage SAN Managed Hub. The ports are numbered sequentially, starting with zero for the left-most port.

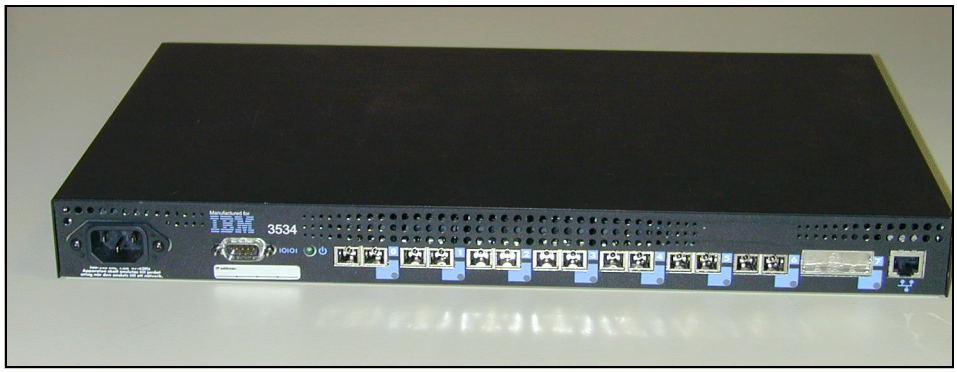


Figure 3-5 IBM TotalStorage SAN Hub faceplate

The system board is enclosed in an air-cooled chassis, which may be either mounted in a standard rack or used as a stand-alone unit.

The chassis includes a power supply, an RJ-45 Ethernet connection for set up and management, and a serial port. If the default address is not known, the serial port is used for recovering the factory settings and initial configuration of the IP address.

The Managed Hub can accommodate one GBIC module and can be connected to one other Managed Hub, to expand the loop capabilities to 14 ports. It can also be connected with a single port into a SAN fabric as a loop extension.

The Managed Hub may be configured using the serial port or the 10/100BaseT Ethernet port. Management interfaces include Telnet or Web-based management using the IBM StorWatch SAN Managed Hub Specialist. This is similar to the IBM TotalStorage Specialist.

The Managed Hub supports either the 50/125 or 62.5/125 Short Wave Length (SWL) cable when attached to a shortwave GBIC or through 9/125 Long Wave Length (LWL) cable when attaching to the longwave GBIC.

Fibre Channel Storage Hubs are designed to provide a centralized point of connectivity, to provide loop fault tolerance, and to simplify configuration management. Specifically designed for entry-level workgroup FC-AL applications, the hubs provide considerable flexibility in configuring loops and segmenting them for performance or high-profile availability applications.

Performance

The Managed Hub supports a minimum aggregate routing capacity of 4,000,000 frames per second for Class 2, Class 3, and Class F frames. Non-blocking throughput of up to 8 x 100 MB/s is provided.

Upgrading the Managed Hub to a switch

The Managed Hub can be upgraded to switched fabric capabilities with the Entry Switch Activation Feature.

The upgrade provides a cost effective, and scalable approach to developing fabric based SANs. The Entry Switch Activation feature provides the license key necessary to convert the FC-AL based Managed Hub to fabric capability. This provides up to eight F_Ports, one of which can be an interswitch link capable port (E_Port), for attachment to the IBM TotalStorage SAN Fibre Channel Switch or other supported switches.

The IBM TotalStorage SAN Managed Hub's Entry Switch Activation feature offers:

- Upgrade to fabric switch capability on all eight ports, F_Ports
- Support for one interswitch link, or E_Port
- Fabric services such as the Simple Name Server and fabric-wide management
- High-speed performance utilizing non-blocking switch-based technology
- Investment protection for small SAN environments moving from Arbitrated Loop to switched fabrics

The Entry Switch Activation feature, P/N 19P3127, does not alter any of the current operating environment specifications for the IBM TotalStorage SAN Managed Hub.

3.6 IBM TotalStorage SAN Switch F08

IBM introduces a new entry-level model of the IBM TotalStorage SAN Switch F08 (3534-F08), with 2 Gigabit Fibre Channel (FC) performance and additional functions to support the storage networking demands for throughput and management controls.

Based on the Brocade SilkWorm 3200, the IBM TotalStorage SAN Switch Model F08 is an 8-port Fibre Channel switch based on a new generation of switch technology. It is designed to provide 2 Gb/s port-to-port throughput with auto-sensing capability for connecting to existing 1 Gb/s host servers, storage, and switches, in a smaller 1U form factor, requiring half the space of the previous 2109-S08 8-port SAN Switch. The new model is fully interoperable with the current IBM 2109 SAN Switches (Models S08, S16, and F16).

The F08 extends the broad range of scalable SAN connectivity solutions available from IBM for a wide variety of host and storage types.

We show a picture of the F08 in Figure 3-6.

Figure 3-6 IBM TotalStorage SAN Switch F08

The F08 is designed to provide:

- Eight non blocking ports, each with full-duplex throughput at either 2 gigabits per second or 1 gigabit per second.
- Auto-sensing ports that self-negotiate to the highest speed supported by the attached server, storage, or switch.
- Hardware zoning controlled at the port level, and software zoning controlled at the worldwide name level.
- Support for high-speed data traffic with the Performance Bundle feature, which provides Inter-Switch Link (ISL) Trunking and Performance Monitoring. Up to four ISLs can be combined for throughput capability of up to 8 gigabits per second.
- Cascading support for flexibility in creating scalable fabric topologies.
- The IBM TotalStorage SAN Switch Specialist, which provides a comprehensive set of management tools that support a Web browser interface for flexible, easy-to-use operations.

The F08 supports Fibre Channel connectivity for the following servers:

- IBM e(logo) server pSeries and selected RS/6000 servers
- IBM e(logo) server xSeries and selected Netfinity servers
- IBM TotalStorage Network Attached Storage 300G
- Other Intel-based servers with Microsoft Windows NT and Windows 2000
- Selected Sun and HP servers

The F08 supports Fibre Channel connectivity for the following servers storage systems:

- IBM Enterprise Storage Server (ESS)
- IBM FAStT Family of Storage Servers
- IBM Magstar 3590 Subsystems and 3494 Tape Libraries
- IBM Ultrium and UltraScalable Tape Libraries
- Other selected storage systems

Additionally, the F08 offers:

- Two gigabit per second industry-standard Fibre Channel switch throughput
- Scalability from small to very large enterprise SAN fabric environments
- A high availability design with hot-pluggable components, and automatic path rerouting
- Modularity for flexible system configurations, including diagnostics to isolate problems quickly
- An Inter-Switch Link Trunking option to combine up to four physical links into one logical high-speed trunk with up to 8 Gb/s full-duplex throughput
- A Performance Monitoring feature for end-to-end measurement of Fibre Channel traffic, including cyclic redundancy checking (CRC) error counts
- Support for Public Fibre Channel Arbitrated Loop devices
- TotalStorage SAN Fibre Channel Specialist for fabric management

The F08 Switch is ideally suited for disaster tolerance solutions such as remote tape vaulting and remote disk mirroring. F08 Switches can provide up to twice the throughput of SAN Managed Hubs. This performance capability can be used to either reduce the number of expensive extended distance ISL connections or to improve the performance with the same number of connections.

F08 Switches are ready to exploit the performance potential of newer servers and storage devices with 2 Gb/s capabilities. F08 Switches can be used as edge switches to expand an existing core-to-edge SAN fabric infrastructure. As F16 Switches and larger, high availability core switches are added to the core, installed F08 Switches can be migrated to the edge. This approach supports scalable network growth in a modular, cost-effective and non-disruptive manner with investment protection for installed switches.

3.7 IBM TotalStorage SAN Switches, S08, and S16

The IBM TotalStorage SAN Switch (2109-S08 and 2109-S16) interconnects multiple host servers with storage servers and devices, creating a Storage Area Network or SAN. An IBM TotalStorage SAN Switch can be used either as a standalone device to build a simple SAN Fabric, or 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 hundreds of Fibre Channel ports. The SAN Fabric provides the 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.

3.7.1 Product overview

The IBM TotalStorage SAN Switch operates at up to 100 MB/s per port with full-duplex data transfer with the 2109-S08 and 2109-S16 models, and up to 200 MB/s with the 2109-F16 model. Unlike hub-based Fibre Channel Arbitrated Loop (FC-AL) solutions which reduce performance as devices are added, the SAN Fabric performance increases as additional switches are interconnected.

IBM offers three different types of IBM TotalStorage SAN Switches which are OEM products from the Brocade SilkWorm family:

- IBM TotalStorage SAN Switch Model 2109-S08 is an 8-port model:
  - Equivalent to a SilkWorm 2400
- IBM TotalStorage SAN Switch Model 2109-S16 is a 16-port model:
  - Equivalent to a SilkWorm 2800
- IBM TotalStorage SAN Switch Model 2109-F16 is a 16 port model:
  - Equivalent to a SilkWorm 3800

> **Note:** As there are significant differences between the 2109-S models and the 2109-F16, we describe the functions and features of the 2109-F16 in 3.8 “IBM TotalStorage SAN Switch F16” on page 167.

Figure 3-7 shows the 8-port model and Figure 3-8 shows the 16-port model.

Figure 3-7 The IBM TotalStorage SAN Switch (2108-S08)

Figure 3-8 IBM TotalStorage SAN Switch (2109-S16)

3.7.2 IBM TotalStorage SAN Switch hardware components

The IBM TotalStorage SAN Switch is available as two models:

- **2109-S08**

  This is an 8-port Fibre Channel gigabit switch that consists of a motherboard with connectors for supporting up to 8 ports.

- **2109-S16**

  This is a 16-port Fibre Channel gigabit switch that consists of a motherboard with connectors for supporting up to 16 ports.

The motherboard is enclosed in an air-cooled chassis which may be a standard rack or used as a standalone unit. The chassis includes one or two power supplies, a fan tray, and an RJ-45 Ethernet connection for switch set up and management. The S08 also has a serial port.

Serial port connection

The serial port is used for recovering factory settings only and for the initial configuration of the IP address for the switch, if the default address is not known. It is not used during normal operation.

The IBM SAN Fibre Channel Switch, 2109-S16, does not have a serial port, but rather uses the LCD display.

Ethernet connection

It is possible to connect an existing Ethernet 10/100BaseT LAN to the switch using the front panel RJ-45 Ethernet connector. This allows access to the switch's internal SNMP agent, and also allows remote Telnet and Web access for remote monitoring and testing. The IP address may be changed using the Ethernet port.

The front panel of the 2109-S16, as shown in Figure 3-9, has a display panel that is used to configure the switch. Generally, this mainly used as part of the initial configuration when setting up the IP address or changing factory settings.

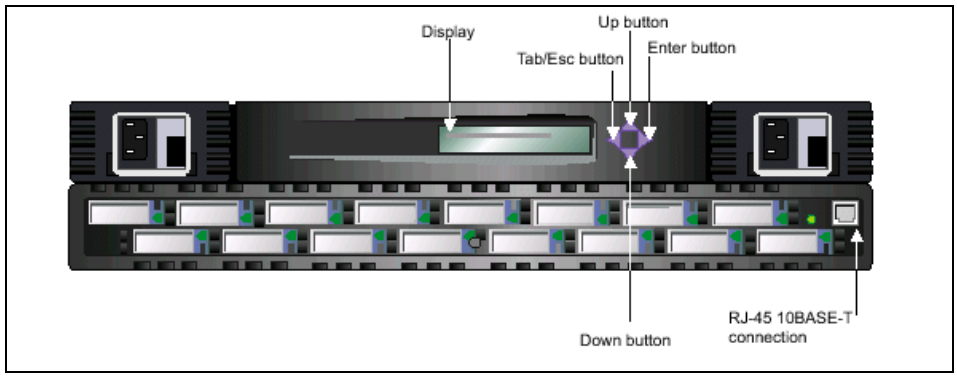


Figure 3-9 2109-S16 display panel and controls

GBICs

The IBM TotalStorage SAN Switch uses Gigabit Interface Convertors which are laser-based, hot-pluggable transceivers that use high data rates (typically 1063 or 1250 Mb/s). The GBICs for the IBM TotalStorage SAN Switch are available in both short wave (SWL) and long wave (LWL) options and provide flexibility for configuring a Storage Area Network. The short wave GBIC supports a distance of up to 550 metres, whereas the long wave GBIC can support up to 10 km.

The IBM TotalStorage SAN Switch comes standard with four short wave GBICs, and supports a mix of additional long or short wave GBICs. The GBIC supports fiber optic cables of 9 microns for long wave ports and 50 or 62.5 microns for the short wave ports.

The GBICs are hot pluggable and are easy to configure and replace. The unused port positions are protected by a metal, spring-loaded door that protects the switch.

The GBICs have status lights which are visible on the front panel, giving a quick, visual check of the GBICs status and activity.

Fibre Channel connections

The IBM TotalStorage SAN Switch supports the following types of Fibre Channel connections:

- Fabric (F_Port)
- Arbitrated Loop that is public and private (FL_Port)
- Interswitch connection (E_Port)

Supported port types

The IBM 2109-S08 and 2109-S16 support the following port types:

- E_Port is an expansion port.
- A port is designated an E_Port when it is used as an interswitch expansion port to connect to the E_Port of another switch, to build a larger switch fabric.
- F_Port is a fabric port that is not loop capable.
- Used to connect an N_Port to a switch.
- FL_Port is a fabric port that is loop capable.
- Used to connect NL_Ports to the switch in a loop configuration.
- G_Port is a generic port.
- It can operate as either an E_Port or an F_Port. A port is defined as a G_Port when it is not yet connected or has not yet assumed a specific function in the fabric.
- Isolated E_Port
- This is a port that is online but not operational between switches due to overlapping domain ID or nonidentical parameters such as E_D_TOVs.
- L_Port is a loop capable fabric port or node.
- N_Port is a node port that is not loop capable.
- Used to connect an equipment port to the fabric.
- NL_Port is a node port that is loop capable.
- Used to connect an equipment port to the fabric in a loop configuration through an FL_Port.
- U_Port is a universal 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.

Currently, the IBM TotalStorage SAN Switch only supports the same vendor switch interconnection through the use of an E_Port. For example, the 2109-S08 can be connected to other 2109-S08, or 2109-S16 or even Brocade Silkworm switches.

Switch electronics

The system board incorporates the Fibre Channel port interfaces, two ASICs in the IBM TotalStorage SAN Switch Model 2109-S08 and four in the Model 2109-S16, switching mechanism, the embedded switch control processor (i960RP), and support logic for the embedded processor logic.

ASIC

The ASIC provides four Fibre Channel ports that may be used to connect to external N_ports (as an F_port), external loop devices (as an FL_port), or to other IBM TotalStorage SAN Switch (as an E_port). Each port operates at 1.0625 Gb/s. The ASIC contains the Fibre Channel interface logic, message/buffer queuing logic, receive buffer memory for the four on-chip ports, and other support logic.

Central Memory Module

The IBM TotalStorage SAN Switch is based on a central memory architecture and has a central memory module (CMM). In this scheme, a set of buffers in the central memory is assigned to each port, to be used for receipt of frames. As an ASIC port receives and validates a frame, it stores the frame in one of its receive buffers in the central memory and forwards a routing request (`Put` message) to the appropriate destination ports.

When a destination port is capable of transmitting the frame, it reads the frame contents from central memory and forwards the frame to its transmit interface. It does not wait for the frame to be written in memory, unless the port is busy. Once it has removed an entry for a frame from its internal transmit queue in preparation for frame transmission, the destination port sends a `Finish` message to indicate “transmission complete” to the port that received the frame, allowing the receiving port to reuse the buffer for subsequent frames received.

The IBM TotalStorage SAN Switch central memory is incorporated into the ASICs. Frames received on the four ports in an ASIC are written into the portion of central memory in the receiving chip; received frames may not be written into the sections of central memory located in other ASICs. All transmitters in a switch may read from the memories in any of the ASICs, through inter-chip connections clocked at 106.25 MHz.

Inside each ASIC, there are a total of 6272-by-34-bit static RAM devices plus data path crossbar logic used to implement the central memory. This provides 112 receive buffers that accommodate full 2112-byte payload frames for 4 ports (or 128 2048-byte typical frames). Each memory block is accessed in a time-sliced fashion. The buffer design is efficient in that if frames are smaller than 2112 bytes, the buffer pool will expand proportionately, effectively providing greater than 128 receive buffers. A single 4-port ASIC can buffer a total of 448 small frames (36-576 bytes), enabled using mini-buffers of 308 bytes in size.

Control Message Interface

The IBM TotalStorage SAN Switch control message interface (CMI) consists of a set of control signals used to pass hardware-level messages between ports. These control signals are used by recipient ports to inform transmitting ports when a new frame is to be added to the transmitter’s output queue. Transmitting

ports also use the CMI to inform recipient ports that a frame transmission has been completed. A recipient port is free to reuse a receive buffer when it receives notification that the frame has been transmitted. Multiple notifications are required, in the case of multicast, to determine when a receive buffer is freed.

The CMI interfaces for the ASICs are connected inside each ASIC through a message crossbar, implementing a barrel shift message scheme. Each chip time slices its output port to each possible destination chip in the switch. If it has a message to send to a particular destination during the corresponding time slot, the chip will use the time slot to send the message; otherwise, the output port lines will be driven to indicate no message is present.

The time slicing of the output CMI control signals of the ASICs are arranged out of phase from each other, such that, in any given clock cycle, each chip's output port is time sliced to a different destination chip. Thus, messages appearing at the input control signal interface of a given ASIC are also time sliced through each possible source chip in the switch.

PCI bus operation

In the IBM TotalStorage SAN Switch, the interface between the embedded i960RP processor and the ASICs is implemented using a 33 MHz PCI bus. ASICs are connected directly to one of the PCI bus interfaces (Primary PCI@3.3 V) of the processor. A slave-only PCI interface is provided by each ASIC, to allow the processor to program various registers, routing tables, and so on, within the chip.

The PCI bus interface to the ASICs operates in 32-bit mode, and has a word-wide even parity bit. The second PCI bus (Secondary, PCI@5V) connects to an Ethernet MAC (media access controller) IC. This provides 10/100 BaseT Ethernet connectivity. Either the i960RP or MAC may become bus master.

Embedded processor

The embedded processor is an Intel i960RP processor, with a clock speed of 33 MHz. It contains an integrated memory controller, a bridged dual PCI bus, and an I^2C controller. The I^2C bus provides peripheral I/O control for the LCD module, thermometers, general I/O functions, and others. In addition, the design includes an RS232 serial port, 10/100 BaseT Ethernet port, SDRAM, and FLASH EEPROM for firmware text, initialized data, and switch configuration information. Logic is also provided for the 16-port switch to allow the processor to display characters on the Model 2109-S16 switch's front panel and to read the state of the front panel buttons.

Host attachment

The IBM TotalStorage SAN Switch supports attachments to multiple host systems, including these:

- IBM Netfinity and Intel-based servers running Microsoft's Windows NT, Windows 2000, or Novell Netware
- IBM RS/6000 servers running AIX
- IBM NUMA-Q servers
- SUN servers running Solaris
- Hewlett Packard servers running HP-U

For a complete list of the supported platforms, please check the IBM TotalStorage SAN Switch Supported Servers site:

http://www.storage.ibm.com/hardsoft/products/fcswitch/supserver.htm

Device attachment

The SAN connectivity products and storage systems that can be attached to the IBM TotalStorage SAN Switch include:

- IBM Enterprise Storage Sever
- IBM Modular Storage Server
- IBM FAStT 200 RAID / Storage unit
- IBM FAStT500 RAID controller
- IBM Fibre Channel RAID Storage Server; and the Netfinity Fibre Channel RAID Controller Unit
- IBM 3494 Automated Tape Library and IBM Magstar 3590 Tape Drives with Native Fibre Channel attachment feature
- IBM Fibre Channel Managed Hub, IBM Fibre Channel Hub (including the Netfinity Fibre Channel Hub)
- IBM SAN Data Gateway with:
  - IBM Magstar and Magstar MP libraries
  - IBM 3502 DLT Tape Library
  - IBM Enterprise Storage Servers (SCSI attachment)
  - IBM LTO 3584 Tape Library

3.7.3 IBM TotalStorage SAN Switch software features

The IBM TotalStorage SAN Switch can be managed using three different methods:

- By using the IBM TotalStorage Specialist, which provides a user-friendly, Web browser interface
- By using Telnet commands
- With SNMP

For our purposes we mainly used the Fibre Channel Switch Specialist, which provides advanced management capabilities for the following:

- Automatic discovery and registration of host and storage devices
- Intelligent rerouting of connection paths, should a port problem occur
- Cascading of switches, for scaling to larger configurations and to provide resiliency for high data availability
- Switch zoning for fabric segmentation
- Configuration with hot pluggable port GBICs for shortwave or longwave optical connections of up to 10 kilometers
- Fabric Watch, which provides monitoring and alert management for the switch components.

3.8 IBM TotalStorage SAN Switch F16

The IBM TotalStorage SAN Switch F16 (2109-F16) provides 2 Gigabit Fibre Channel (FC) performance and additional functions to support the storage networking demands for higher security, throughput, and management controls.

As there is a lot of similarity between the S08/S16 and F16 models, we restrict this particular topic to those new features that this brings to the market. Users that are familiar with the S08/S16 will have no problem upgrading their awareness to successfully encompass the F16.

We show a picture of the 2109-F16 in Figure 3-10.

Figure 3-10 2109-F16 switch

3.8.1 Product overview

The F16 is a 16-port Fibre Channel switch based on a new generation of switch technology. It provides 2 gigabit per second port-to-port throughput with auto-sensing capability for connecting to existing 1 gigabit per second host servers, storage, and switches, in a smaller 1U form factor, requiring half the space of the previous 16-port SAN Switch.

The new model is fully interoperable with the current IBM TotalStorage SAN Switches (Models S08 and S16), and can be added to existing fabrics with minimal disruption, enabling an easy transition for existing Fibre Channel storage networks to the faster technology.

The IBM TotalStorage SAN Switch Model F16 extends the broad range of scalable SAN connectivity solutions available from IBM for a wide variety of host and storage types. IBM TotalStorage SAN Switches enable storage resources to be shared efficiently and to scale rapidly to meet the demands by users for highly available, heterogeneous access to expanding storage pools.

The new Model F16 provides:

- Sixteen non blocking ports, each with full-duplex throughput at either 2 gigabits per second or 1 gigabit per second
- Auto-sensing ports that self-negotiate to the highest speed supported by the attached server, storage, or switch
- Universal ports that 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 smaller 1U package that can be either rack-mounted or used in a table-top configuration, with the option of a redundant power supply, providing a highly available switch.
- Hardware zoning controlled at the port level, and software zoning controlled at the worldwide name level.
- Support for high-speed data traffic with the Performance Bundle feature, which provides Inter-Switch Link (ISL) Trunking and Performance Monitoring. Up to four ISLs can be combined for throughput capability of up to 8 gigabits per second.
- Cascading support for flexibility in creating scalable fabric topologies.
- Distributed fabric services such as name serving, zoning, routing, and microcode upgrade.

- The IBM TotalStorage SAN Switch Specialist, which provides a comprehensive set of management tools that support a Web browser interface for flexible, easy-to-use operations.

3.8.2 Hardware components

The 2109 Model F16 system board is a single-board design with a highly integrated CPU. The Intel 80960VH CPU is a RISC core processor and is the top choice for this platform. It provides over 70% of the functionality for the digital section of the system board. The system uses three types of memory devices: DRAM, Flash File, and Boot Flash. On the Fibre Channel section of the system board, the Bloom ASICs, the Serializer/Deserializer (SERDES), and the SFP media are the key components that provide high-speed data transfer. SFP media interfaces support SWL and LWL.

The system chassis is a 1U height enclosure with space for two power supply units and one system board. The system board is placed in an Electromagnetic Interference (EMI) enclosure tray as an EMI-proof system unit. Two 126-watt removable, redundant power supplies provide hot-swappable capability. Cooling fans are mounted in the rear to provide airflow for system cooling.

CPU subsystem

An Intel 80960VH CPU is used for switch initialization and management functions. The CPU runs the fabric OS and is responsible for switch initialization, configuration, and management. IBM-designed ASICs provide the switching functionality.

The following peripherals are supported as well:

- An Ethernet port
- A serial port
- Three digital thermometers
- A real-time clock
- Two power supply controls
- General I/O

The CPU subsystem is a mixed voltage system using 1.8 V, 2.5 V, 3.3 V, and 5 V depending on the device. The maximum board power consumption is 78 W.

Features

The 2109-F16 CPU subsystem includes the following features:

- A 80960VH-100 MHz CPU
- A SDRAM controller with parity check at 33 MHz
- A peripheral control interconnect (PCI) bus arbiter

- An on-board SDRAM with data parity to support a 16 MB configuration
- One PLCC32 Boot Flash socket to support up to 512 KB of Flash memory
- 8 MB (2 x 4 MB) Flash memory for software storage
- 10BASE-T or 100BASE-T port for management connection with RJ45 connector
- One RS232 port with DB9 connector
- 16 LEDs to indicate the status for each port
- 16 LEDs to indicate the link speed for each port
- One LED (green) to indicate the system power-on status
- Three digital thermometers for temperature sensing
- Two analog switches to control the power supply inter-integrated circuit (I2C) bus access
- One 3.3 V to 1.8 V dc/dc converter for Bloom ASIC core supply
- Two Bloom ASICs supporting up to 16 non-blocking ports
- 16 SERDES
- One real-time clock with a battery and 56 bytes of nonvolatile RAM (NVRAM)

Embedded processor

The embedded processor is an Intel 80960VH processor with a clock speed of 100 MHz. It contains the following:

- A high-performance RISC processor core (compatible with the 2109 series of switches and the 3534 switch)
- An integrated EDO memory controller (for DRAM, SRAM, ROM, and Flash memory)
- A PCI bus interface
- A complex programmable logic device (CPLD) for SDRAM control
- Two direct memory access (DMA) channels
- An I2C interface
- General purpose I/O

You can access system memory through the local bus. The external CPLD SDRAM device provides SDRAM controller functionality at 33 MHz. It supports parity checking to enhance the data integrity of the system. The CPU communicates with the ASIC and the 10BASE-T or 100BASE-T Ethernet media access controller (MAC) through the PCI interface. An external PCI bus arbiter enables the Ethernet device to be a bus master.

You can also access the RS232 Universal Asynchronous Receiver Transmitter (UART) serial port through the local bus. Other I/O peripherals, such as the real-time clock, the two power supply controls, the LEDs, the three digital thermometers, and miscellaneous I/O are handled by the I2C bus of the CPU. The CPU is the only I2C bus master in the system. The RS232 port and drivers, Ethernet MAC/PHY, and LEDs are external components to the CPU. An RJ45 connector provides Ethernet connection to external systems. The DB9 RS232 is a ribbon-cable connection through the on-board 10-pin header.

Bus operations

The interface between the embedded processor, the ASICs, and the 10BASE-T or 100BASE-T Ethernet MAC is implemented using a PCI bus. All PCI devices on the bus are PCI Revision 2.2 compliant. The PCI bus interface operates at 32-bit, up to 33 MHz and has a worldwide even parity bit. A slave-only PCI interface is provided by each ASIC to allow the processor to program various registers, routing tables, and so on within the chip. An external PCI bus arbiter enables the Ethernet device to be a bus master.

The local bus, a 32-bit multiplexed burst bus, provides the interface between the system memory and the I/O. Because the integrated EDO memory controller on the CPU allows only direct control for DRAM, SRAM, ROM, and Flash memory, the external CPLD controller is included to provide SDRAM controller functionality.

The I2C bus provides peripheral I/O control for the LEDs, the thermometers, and general I/O functions. The 80960VH CPU serves as the master on the I2C bus.

Each Bloom ASIC is an eight-port Fibre Channel switch controller. There are two Bloom ASICs to support up to 16 ports. The communication between ASICs is over a proprietary 10-bit wide SSTL2 bus running at 106.25 MHz. An SSTL2 bus is also used between the Bloom ASICs and the SERDES.

Memory

The system design uses the following three types of memory devices:

- DRAM
- Flash File
- Boot Flash

Two on-board SDRAM chips provide up to 16 MB for system memory. Two additional SDRAM chips provide data parity. The printed circuit board (PCB) SDRAM footprint is designed to be compatible with 64 MB, 128 MB, and 256 MB devices. An external CPLD device added to the local bus provides control functions for the 80960VH processor.

The system provides 4 MB of on-board redundant Flash File memory for software and data storage. The Boot Flash is an 8-bit Flash device socket that is used only for system start. The Boot Flash device contains a block area for startup code protection. The PLCC32 socket supports 3.3 V Boot Flash memory up to 512 KB.

Central memory

As with the 2109 series of switches and the 3534 switch, the 2109 Model F16 is based on a central memory architecture. In this scheme, a set of buffers in the central memory is assigned to each port, to be used for receipt of frames. As an ASIC port receives and validates a frame, it stores the frame in one of its receive buffers in the central memory and forwards a routing request (a `Put` message) to the appropriate destination ports.

When a destination port is capable of transmitting the frame, it reads the frame contents from central memory and forwards the frame to its transmit interface. It does not wait for the frame to be written in memory, unless the port is busy. After it has removed an entry for a frame from its internal transmit queue in preparation for transmitting a frame, the destination port sends a transmission complete message (a Finish message) to the port that received the frame. This allows the receiving port to reuse the buffer for subsequent frames received.

The central memory is also incorporated into the ASICs. Frames received on the ports in an ASIC are written into the portion of central memory in the receiving chip; received frames cannot be written into the sections of central memory located in other ASICs. All transmitters in a 2109 Model F16 switch can read from the memories in any of the ASICs, through inter-chip connections clocked at 106.25 MHz.

Each ASIC contains RAM devices plus data path crossbar logic that is used to implement the central memory. Memory blocks are accessed in a time-sliced fashion. The buffer pool can be split into 2112-byte buffers or into 312-byte mini-buffers. If frames that need to be buffered are smaller than the maximum 2112 bytes, using mini-buffers effectively expands the buffer pool and increases the efficiency of memory usage by providing more (but smaller) receive buffers.

Additionally, the Bloom ASIC provides a special memory interface (SMI). The SMI provides the firmware with a mechanism to read and write frame contents to and from the ASIC. It also supports higher throughput transfers. The SMI includes a set of two buffers that are large enough for an entire maximum-sized frame to be transferred in a single operation. Additionally, because there are two buffers available, the firmware can perform a read or write on a frame in one of the buffers while the ASIC streams another frame into the other buffer.

ASICs

Two ASICs within the system provide the switching functionality. Each ASIC provides eight Fibre Channel ports that can be used to connect to external N_ports (as an F_port), external loop devices (as an FL_port), or to other 3534 or 2109 series boxes (as an E_port).

Each port can operate at either 1.0625 Gb/s or 2.125 Gb/s link speeds. The ASIC contains the Fibre Channel interface logic, message and buffer queuing logic, receive buffer memory for the eight on-chip ports, and other support logic.

The Bloom ASICs are PCI slaves to the CPU. The two ASICs interface through an inter-chip 10-bit SSLT2 bus connection clocked at 106.25 MHz. A 16-channel SERDES is used to support 16 ports. The interface between ASIC and SERDES is also a 10-bit SSTL2 bus running at 106.25 MHz. The SERDES converts the 10-bit wide parallel data from the SSTL2 bus into high-speed serial data for the SFP media and vice versa. The SERDES supports single data rate (SDR) or double data rate (DDR) transfer between the SERDES and the SFP media. The DDR operation supports 2.125 Gb/s data transfer rate between ASICs. Implementing the SERDES external to the ASIC reduces the risk of silicon packaging as well as the risk of running 2.125 Gb/s signals on a board with a long trace length.

The SFP media interfaces to external devices and enables support for short-wave laser and long-wave laser. Two LEDs for each port provide port status and link speed information.

Control Message Interface (CMI)

The 2109 Model F16 Control Message Interface (CMI) consists of a set of control signals that are used to pass hardware-level messages between ports. Recipient ports use these control signals to inform transmitting ports when a new frame needs to be added to the output queue of the transmitter. Transmitting ports also use the CMI to inform recipient ports that a frame transmission has been completed. A recipient port is free to reuse a receive buffer when it receives notification that the frame has been transmitted. In the case of multicast, multiple notifications are required to determine when a receive buffer is freed.

The CMI interfaces for the ASICs are connected inside each ASIC through a message crossbar, implementing a barrel shift message scheme. Each chip time slices its output port to each possible destination chip in the switch. If it has a message to send to a particular destination during the corresponding time slot, the chip uses the time slot to send the message. Otherwise, the output port lines are driven to indicate that no message is present.

The time slicing of the output CMI control signals of the ASICs are arranged out of phase from each other so that each chip's output port is time sliced to a different destination chip in any given clock cycle. Messages that are displayed at the input control signal interface of a given ASIC are also time sliced through each possible source chip in the switch.

Ports

The 2109 Model F16 supports the following port types:

- Optical ports
- Ethernet port
- Serial port

Each ASIC in the 2109 Model F16 switch connects up to eight SFP media. SFP devices are encased in metal to ensure low emissions and high thermal management. They are hot-swappable and use industry-standard local channel connectors. Each port provides ISL, loop, and fabric (E, F, and FL respectively) type connectivity that the 2109 Model F16 senses automatically; it requires no administration to identify the port type.

Fibre Channel connections

The IBM TotalStorage SAN Switch supports the following types of Fibre Channel connections:

- Fabric (F_Port)
- Arbitrated loop — public and private (FL_Port)
- Interswitch connection (E_Port)

Supported port types

The IBM 2109-S08 and 2109-S16 support the following port types:

- E_Port is an expansion port.
- A port is designated an E_Port when it is used as an interswitch expansion port to connect to the E_Port of another switch, to build a larger switch fabric.
- F_Port is a fabric port that is not loop capable.
- Used to connect an N_Port to a switch.
- FL_Port is a fabric port that is loop capable.
- Used to connect NL_Ports to the switch in a loop configuration.
- G_Port is a generic port
- It can operate as either an E_Port or an F_Port. A port is defined as a G_Port when it is not yet connected or has not yet assumed a specific function in the fabric.
- Isolated E_Port

- This is a port that is online but not operational between switches due to overlapping domain ID or nonidentical parameters such as E_D_TOVs.
- L_Port is a loop capable fabric port or node.
- N_Port is a node port that is not loop capable.
- Used to connect an equipment port to the fabric.
- NL_Port is a node port that is loop capable.
- Used to connect an equipment port to the fabric in a loop configuration through an FL_Port.
- U_Port is a universal 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.

Optical ports

For optical ports, the 2109 Model F16 uses SFP fiber-optic transceivers that convert electrical signals to optical signals (and optical signals to electrical signals). Capable of transmitting at both 1 and 2 Gb/s speeds, each SFP fiber-optic transceiver supports 850 nm SWL on multimode fiber-optic cable, 1310 nm LWL on single-mode fiber-optic cable, and 1550 nm ELWL 5 on single-mode fiber-optic cable. These miniature optical transceivers provide high port density and deliver twice the port density of standard removable GBIC transceivers.

Ethernet port

The 2109 Model F16 provides a fully IEEE-compliant 10BASE-T or 100BASE-T Ethernet port for switch management console interface. When a device is connected to the port, both ends negotiate to determine the optimal speed. The Ethernet port uses an RJ45 connector. There are two LEDs for the port. One LED indicates transmit and receive activity and one LED indicates speed (10 Mb/s or 100 Mb/s). The TCP/IP address for the port can be configured from the serial port.

Serial port

An RS232 serial port is provided on the 2109 Model F16. The serial port uses a DB9 connector. The connector is a header pin block on the system board. The parameters of the serial port are fixed at 9600 baud, 8 data bits, no parity, no hardware flow control, 1 start and 1 stop bit.

You use this connector to configure the IP address and to recover the factory default settings of the switch should Flash memory contents be lost. The serial port connection should not be used to perform normal administration or maintenance functions. Accessible functions are limited to connecting a terminal to the port to re-initialize the switch defaults, which restores the switch to its factory configuration. This is required to restore the switch passwords to a known state and to allow customers to set a specific switch IP address.

Enclosure

The 2109 Model F16 enclosure is designed to be mounted in a 19-inch rack, with a height of 1 RETMA unit (1µ), but it can also be used in a tabletop configuration. The enclosure houses dual-redundant power supplies, dual-redundant fan assemblies, and a system board that supports the two ASICs and the CPU.

The 2109 Model F16 enclosure has forced-air cooling. The fans push the air from the rear chassis intake through the enclosure and exhaust the air through venting holes in the front panel. The SFP media, the cooling fan, and the power supplies are hot-swappable so that they can be removed and replaced without interrupting the system power.

The top panel of the 2109 Model F16 enclosure can be removed without tools, allowing access to the system board. The enclosure design provides for simple assembly of the system board into the enclosure, allowing for ease of manufacture and maintenance. All pieces of the product are modular, and all maintenance can be performed without special tools.

On the front of the unit, there are two port connections (an RS232 connection and an RJ45 connection). The RJ45 connection provides a 10BASE-T or 100BASE-T Ethernet port for a full system management console interface. The RS232 connection provides a serial port interface for setting the IP address of the switch and for resetting the switch to factory defaults

The fibre-optic cables, Ethernet cables, and serial port cables are located on the front of the switch. AC power input cables, power supplies, and cooling modules are inserted and removed from the rear of the switch.

Power supply

The 2109 Model F16 power supply is a hot-swappable switching unit, allowing 1+1 redundant configurations. The unit is a universal power supply capable of functioning worldwide without voltage jumpers or switches. The fully enclosed, self-contained unit has its own internal fans to provide cooling. It is auto-ranging in terms of accommodating input voltages.

The power supply has three DC outputs (3.3 V, 5 V, and 12 V) that provide a total output power of 126 maximum usable watts. The power supplies plug directly into the enclosure from the rear of the unit, mating to internal blind connectors that connect both the DC outputs and the interface signals to the system backplane. An integral on/off switch, input filter, and power indicator are provided in the power supply.

LEDs

The 2109 Model F16 provides several LEDs to indicate status on the switch. Each of the 16 ports has two status indicators. The first LED for the port is a two-color (green and yellow) LED, and indicates the status for the port. Green indicates normal status, and yellow indicates an error. The second LED is a single-color (green) LED and indicates the link speed for the port. Green indicates 2 Gb/s; if the LED is not lit (dark), it indicates 1 Gb/s.

A single-color (green) LED is located at the front of the unit and indicates system power-on status. On the back of the unit, there is a two-color (green and yellow) LED driven by an I2C I/O expander that indicates the mode of the unit (Green indicates normal mode and yellow indicates diagnostic mode). All LEDs are surface mount components with on-board light pipe and are visible externally with full chassis enclosure. There are two LEDs for the Ethernet port. One LED indicates the transmit and receive activity and one LED indicates speed (10 Mb/s or 100 Mb/s).

3.8.3 Software specifications

The 2109 Model F16 switch is supported by the Fabric OS Version 3.0. The Fabric OS is implemented in firmware and manages the operation of the 2109 Model F16 switch. The switch firmware is designed to make a 2109 Model F16 easy to install and use while retaining the flexibility needed to accommodate user requirements. A fabric constructed with cascaded 2109 Model F16 switches automatically assigns individual switch addresses, establishes frame routes, configures the internal name server, and so on.

Users can access internal management functions using standard host-based Simple Network Management Protocol (SNMP) software or Web browsers. They can access these functions using network connectivity through the Ethernet port or using Internet Protocol (IP) over the Fibre Channel ports. SCSI Enclosure Services (SES) is also supported as a management method. The management functions of the switch allow a user to monitor frame throughput, error statistics, fabric topology, fans, cooling, media type, port status, IDs, and other information to aid in system debugging and performance analysis.

The Fabric OS includes all basic switch and fabric support software as well as optionally licensed software that is enabled using license keys. The fabric license is pre-installed on the 2109 Model F16 switch to ensure fabric operation. The Fabric OS is composed of two major software components:

- Firmware that initializes and manages the switch hardware.
- Diagnostics that perform component self-testing algorithms for fault isolation during the manufacturing process and in customer installations.

The internal firmware can be viewed as a set of embedded applications running on top of a proprietary real-time operating system.

Additionally, host-based software includes the drivers, utilities, and applications that use the switch. You can obtain these components from your system vendor or Fibre Channel component supplier.

2109 Model F16 software

The 2109 Model F16 software consists of a set of embedded applications running on top of a real-time operating system kernel. The set of applications include the following:

- Name server
- Alias server
- SNMP agent

The set of applications also includes several tasks to manage the following:

- Address assignment
- Routing
- Link initialization
- Fabric initialization
- Link shutdown
- Switch shutdown
- Frame filtering
- Performance monitoring
- Trunking
- Auto speed negotiation
- The user interface

All embedded applications are written in C, except for the SNMP agent (included with the real-time operating system package) and the Web server.

Applications

The 2109 Model F16 software applications implement a variety of functions. Switch applications exist to provide fabric services, such as name server and alias server functionality, to external devices. These particular applications process requests from fabric-attached external devices, and communicate with similar applications running on other switches within the fabric to obtain fabric-wide information to satisfy these requests. The applications present an interface to these standards-based services that provides access to information throughout the fabric while hiding the details of how the information is distributed across switches within the fabric from the external devices.

Other applications running in a switch implement functions used to manage internal fabric operation. One task allows for automatic address assignment throughout a fabric through a distributed algorithm run by participating switches. Another task, used to set up routes within the fabric, communicates with tasks that are running on other switches in the fabric to set up loop-free, lowest-cost routes.

The 2109 Model F16 switch provides an extensive set of diagnostics. A number of comprehensive low-level diagnostics can be used to detect failing switch hardware components by performing hardware-specific tests. In general, these diagnostics must be run when the switch is offline. However, an additional set of high-level diagnostics can be used to exercise individual ports, passing data through external media interfaces and cables. These allow various media, cable, and port faults to be detected while normal switch operation continues on other ports.

New features

The 2109 Model F16 software includes some new features and functionality. The fabric OS enables the 2109 Model F16 to support the new functionality described in the following sections.

Auto-sensing speed negotiation

The 2109 Model F16 ASIC supports link operation at either 2 Gb/s or 1 Gb/s. Auto-sensing negotiation allows easy configuration. Connect the device in and the link speed is negotiated to the highest speed that is supported by the device. Speed selection is auto-negotiated by the ASIC driver on a per-port basis. After the speed is determined, the transmitter and receiver for the port are automatically set. If multiple devices are connected to a port (for example, on an FL_port), the driver auto-negotiates for the highest common speed and sets the transmitter and receiver accordingly.

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 a new feature of the 2109 Model F16 ASIC that enables it 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. After the filter is set up, the complicated function of zoning and filtering can be achieved at wire speed.

Performance Monitoring

Performance Monitoring is a licensed feature that provides error and performance information to manage your storage environment. There are three types of monitoring:

- Arbitrated Loop Physical Address (AL_PA) monitoring: This provides information regarding the number of CRC errors.
- End-to-end monitoring: This provides information regarding a configured source identifier (SID) to destination identifier (DID) pair. Information includes the number of CRC errors for frames with the SID-DID pair, Fibre Channel words transmitted from the port for the SID-DID pair, and Fibre Channel words received for the port for the SID-DID pair.
- Filter-based monitoring: This provides error information with a customer-determined threshold.

Trunking

Trunking is a new feature on the 2109 Model F16 switch 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. The 2109 Model F16 ASIC creates a trunking group (a set of available paths linking two adjacent switches).

Ports in the trunking group are called trunking ports. One trunking port is designated as the trunking master port and is used to set up all routing paths for the entire trunking group. The trunk provides an 8 Gb/s single-aggregate ISL pipe between switches.

Real-time operating system

The 2109 Model F16 real-time operating system consists of a hardware-independent layer and a hardware-dependent section.

The hardware-independent portion of the operating system consists of a third-party real-time kernel plus a number of interfaces. The interfaces provide a structure for handling various layers in the Fibre Channel protocol hierarchy.

In this collection of modules, the FC-PH layer provides FC-2 functionality, supporting reassembly of inbound frames into sequences. This layer also allows for creation of a set of frames to transmit from an internal Fibre Channel sequence description.

The FC-LS layer handles various sorts of Fibre Channel link services, including basic link services and extended link services.

Operations using the Fibre Channel common transport interface, as defined in the FC-GS specification, use the interface provided by FC-CT code in the 2109 Model F16.

Switch-to-switch communications used to manage fabric initialization and routing use the services provided by the FC-SW layer to implement these functions.

Hardware-dependent functions of the real-time operating system contain a number of elements, including the Board Support package. This code is used to provide an interface between VxWorks and the 2109 Model F16-specific hardware related to supporting the 80960VH processor.

Drivers for specific hardware interfaces are also considered part of the hardware-dependent portion of the real-time operating system. A number of drivers support interface hardware that is used for fabric management purposes, such as the Ethernet port and serial port. Other drivers are used for miscellaneous internal functions, including temperature monitoring and power supply control.

Additional drivers, written for the Fibre Channel interfaces of the switch, are managed through two layers. One of these, the port driver, creates a generic interface to the underlying switch hardware, and provides functions common to all switch implementations. Reporting to the port driver are the switch-hardware-specific drivers, which handle the operations of individual types of switch ASICs. Three of these drivers, for the Stitch, Flannel, and Loom chips, are used for IBM's first and second-generation hardware. A fourth module implements the functionality required to drive the Bloom ASIC, which is used in the 2109 Model F16 switch.

Initialization

When the system is started or restarted, the following operations are performed:

1. Early power-on self test (POST) diagnostics are run. POST is run before VxWorks is running.
2. VxWorks is initialized.
3. The hardware is initialized. The system is reset, the internal addresses are assigned to Loom chips, the Ethernet port is initialized, the serial port is initialized, and the front panel is initialized.
4. A full POST is run.
5. The links are initialized. Receiver and transmitter negotiation is run to bring the connected ports online.
6. A fabric exploration is run. This determines whether any ports are connected to other switches. If so, it determines the principal switch.
7. Addresses are assigned. After the principal switch is identified, port addresses are assigned. Each 2109 Model F16 tries to keep the same addresses that it used previously. Previous addresses are stored in the configuration Flash memory.
8. The routing table is constructed. After the addresses are assigned, the unicast routing tables are constructed.
9. Normal Nx_port operation is enabled.

Routing

The embedded processor maintains two routing tables, one for unicast and one for multicast. The unicast routing tables are constructed during fabric initialization. The multicast tables are initially empty, except for broadcast. After the tables have been constructed they are loaded into each ASIC.

The unicast tables change if ports or links come online or go offline, or if some other topology changes occur. When new paths become available, the embedded processor can change some routes in order to share the traffic load.

The multicast tables change as ports register with the alias server to create, join, or leave a multicast group. Each time a table changes it must be reloaded into the ASICs.

Service functions

The ASIC interrupts the embedded processor when a frame arrives that has an error (for example, incorrect source ID), when a frame times-out, or when a frame arrives for a destination that is not in its routing tables. In the latter case, the frame might be addressed to an illegal destination ID, or it might be addressed to one of the service functions that are provided by the embedded processor such as SNMP, name server, or alias server.

SNMP

Simple Network Management Protocol (SNMP) allows network devices to be monitored, controlled, and configured remotely from a network management station running a network manager program.

SNMP agent code in the network device allows management by transferring data that is specified by a Management Information Base (MIB).

The 2109 Model F16 switch agent supports the following:

- SNMPv1 manager
- Command line utilities to provide access to and command the agent
- MIB-II system group, interface group, and SNMP group
- Fabric-element MIB
- IBM-specific MIBs
- Standard generic traps
- IBM-specific traps

Diagnostics

The 2109 Model F16 switch supports a set of power-on self tests (POSTs), as well as tests that can be invoked using Telnet commands. These diagnostics are used during the manufacturing process, as well as for fault isolation of the product in customer installations.

Diagnostic environment

Most diagnostics are written to run in the VxWorks environment. However, as VxWorks does not run without a working SDRAM, a SDRAM/boot EEPROM test is run as part of the pre-VxWorks startup code to verify that the basic processor-connected memories are functioning properly.

Hardware support

Loop-back paths for frame traffic are provided in the hardware for diagnostic purposes. A loop-back path within the ASIC, at the final stages of the Fibre Channel interface, can be used to verify that the internal Fibre Channel port logic is functioning properly, as well as paths between the interface and the central memory.

Additionally, the SerialLink macro within the ASIC includes a serial data loop-back function that can be enabled through a register in the corresponding ASIC.

Diagnostics are provided to allow traffic to be circulated between two switch ports that are connected with an external cable. This allows the diagnostics to verify the integrity of the final stage of the SERDES interface, as well as the media interface module.

Diagnostic coverage

The POST and diagnostic commands concentrate on the Fibre Channel ports and verify switch functionality of the 2109 Model F16 switch.

3.8.4 Interoperability

In the topics that follow, we describe:

- Switch interoperability
- HBA interoperability
- Operating system support

Switch interoperability

The 2109 Model F16 switch supports both 1 Gb/s and 2 Gb/s transmit and receive rates with auto-negotiation. The actual data signaling rate that is used on a port is automatically sensed, and is set to the rate that is supported by a device or devices that are attached to the port. The 2109 Model F16 has been tested and is compliant with the current FC standards. The 2109 Model F16 is compatible with most current-generation switches N_ports, NL_ports, and E_ports, as well as host adapters, Redundant Array of Independent Disks (RAID) storage devices, hubs, and Fibre-SCSI bridge devices, including the 3534 and 2109 series of switches.

Implementation in existing environments

Because the 2109 Model F16 switch has a compatible 1 Gb/s auto-negotiated signaling rate on each port, it can be used as a replacement for current 3534 and 2109 series switches. As newer technology is added to existing systems that support 2 Gb/s signaling, the ports can accept these devices and interoperate with existing 1 Gb/s devices. If the 2109 Model F16 is connected to a third-party device but is unable to negotiate the signaling rate, the 2109 Model F16 allows you to manually set the speed of each port through the management interfaces.

Heterogeneous inter-switch operations

Fabric OS 3.0 supports interoperability for the following functions:

- Basic switch functions:
  - Link initialization
  - Principal switch selection
  - Routing (FSPF)
- Basic services:
  - Simple name service
  - State change notification
  - WWN zoning (typically referred to as soft zoning or name server zoning)

The following facilities are switch-based facilities and will continue to function on any 2109 switch:

- SNMP facilities
- Simple QuickLoops with no zoning
- Translative mode (private target support on fabrics)
- Trunking (will only function between two IBM switches)
- Enhanced performance metrics

The following facilities are IBM value-added facilities that would not be supported in a multi-vendor fabric. Use of these facilities causes the fabric to segment.

- QuickLoop zones
- QuickLoop Fabric assist mode
- Port, protocol, or LUN zoning

IBM is not aware of any areas of non-compliance with any ratified standards at this time.

Host bus adapter interoperability

The 2109 Model F16 has been tested with the following host bus adapters (HBAs) from the following vendors:

- **Emulex**
  - LP6000
  - LP7000
  - LP8000
  - LP850
  - LP952
  - LP9000
- **QLogic**
  - QLA2100
  - QLA2200

- **JNI**
  - FC64-1063
  - FCI-1063
  - FCE-6410
  - FCE-6460
- **Agilent**
  - HHBA-5100
  - HHBA-5101

Operating system support

Fabric OS Versions 2.x and 3.x have no specific OS dependencies. The Fabric OS in the switches allows for any Fibre Channel compliant device to attach to the switches as long as it conforms to the standards for device login, name service, and related Fibre Channel features. Regardless of the operating environment, proper interface to the fabric requires a Fibre Channel HBA with a standards-compliant driver.

The operating systems versions listed in Table 3-1 have been tested (using HBA devices and drivers supplied by QLogic, Emulex, JNI, and Agilent) for interoperability:

Table 3-1 Compatible operating systems

OS	Version
AIX	4.3.3
NT	4.0
Windows 2000	Initial release
Solaris	2.5, 2.5.1, 2.6, 2.8, 7
HP-UX	10.0, 11.0
Linux RedHat	versions 6.2 and 7.0
PTX	
Novel NetWare	NetWare 5.2

3.9 IBM TotalStorage SAN Switch M12

The IBM TotalStorage SAN Switch M12 (2109-M12) provides a highly reliable solution for deploying enterprise-class Storage Area Networks (SANs). By delivering up to 128 ports of connectivity in a single enclosure, the M12 provides unprecedented levels of availability, scalability, manageability, and security for open enterprise storage applications.

With the introduction of the M12, which is an OEM version of the Brocade SilkWorm 12000, IBM continues to extend its IBM TotalStorage SAN connectivity solutions. This high performance, 2 Gb/s Core Fabric Switch is designed to provide the high availability, scalability, manageability and security features to meet your open systems' requirements.The M12 will be available in configurations of a 32-port switch, a 64-port switch, or two 64-port switches in a single, 14U rack mountable enclosure.

Based upon the same next-generation switching technology used in the IBM TotalStorage SAN Switch F16 and F08, this switch supports 1 Gb/s and 2 Gb/s auto-sensing ports as well as advanced fabric services that can simplify the design, administration and management of enterprise SANs. It is designed to provide investment protection to existing customers by being fully backwards-compatible with existing SAN Switches S08, S16, F08, and F16.

High availability features include a fully redundant design and hot-swappable components.

IBM is the first storage provider to offer end-to-end 2 Gb/s solutions designed to exploit next-generation switching technology. Because these end-to-end solutions are up to twice as powerful as previous solutions, they can help reduce the total cost of ownership, simplify SAN management and enable more scalable, larger enterprise SANs.

In Figure 3-11 we show the M12.

Figure 3-11 IBM TotalStorage SAN Switch M12

3.9.1 M12 description

Availability, scalability, and performance are the key attributes required by today's open systems customers. The M12 Switch is designed utilizing high availability features such as:

- Redundant, hot-pluggable components
- Dual-redundant control processors (active/standby)
- Redundant power (four power supplies, two redundant)
- Redundant cooling (three fans, one redundant)

- Automatic path rerouting
- Non-disruptive software upgrades

The IBM TotalStorage SAN Switch M12is designed to support up to eight, sixteen-port Fibre Channel modules (blades) enabling 128 universal (E,F, and FL), full duplex, auto-sensing ports in a single 14U enclosure, each port capable of self-negotiation to the highest speed supported by the attached SAN infrastructure. The M12 will be available in configurations of a 32-port switch, a 64-port switch, or two 64-port switches in a single, 14U rack mountable enclosure. When combined with IBM TotalStorage SAN Switch F16s, S16s and S08s, it is designed to provide a highly scalable core/edge fabric required by our largest enterprise storage customers.

The M12 Switch is designed for high performance, with full duplex, 1 Gb/s throughput on each fibre channel port. Distances up to 10 km are supported using longwave laser transceivers with 9.0u fiber cables. Shortwave laser transceivers support distances up to 500m at 1 Gb/s and up to 300m at 2 Gb/s with 50.0u fiber cables.

Standard features include:

- Rack-mount chassis
- Four power supplies (two redundant) with four rack PDU power cables
- Three fans (one redundant)
- Two control processors (active/standby, with automatic failover)
- Two 16-port 2 Gb/s switch blades provide a single 32-port switch
- Performance Monitoring tools for measuring end-to-end activities.
- ISL-Trunking with up to four links, and up to 8 gigabits per second bandwidth
- Advanced fabric services provided by Fabric OS Version 4.0, IBM SAN Switch Specialist, Advanced Zoning and Fabric Watch

The M12 has a number of selectable options. The Core Fabric Switch includes two 16-port switch blades with the option of either shortwave or longwave Small Form-Factor Pluggable (SFP) optical transceivers and space for up to six additional 16-port switch blades. A mixture of shortwave and longwave SFP optical transceivers may be ordered with a minimum of thirty-two and either 64 or 128 SFP transceivers to completely populate all switch blades.

The M12 Switch requires fiber optic cables for connection to the host systems, and storage systems or devices. These cables can be customer supplied, or ordered with the switch. Additional features available for the M12 Switch include:

- Shortwave SFP transceiver: Provides shortwave optical transceiver for SFP LC media
- Longwave SFP transceiver: Provides longwave optical transceiver for SFP LC media

- 64-port Upgrade: Provides two additional switch blades to create a single 64-port switch
- 128-Port Upgrade: Provides six additional switch blades to create two 64-port switches
- Fabric Manager 3.0: Provides a Java-based application that can simplify management of a multiple switch fabric. IBM SAN Switch Specialist and Fabric Manager run on the same management server attached to any switch in the core/edge fabric. It may also manage up to eight fabrics. It requires a Windows NT/2K or Solaris 7 server with a Netscape or Internet Explorer Web browser.
- Extended Fabric Activation: Provides license key to optimize management of the internal switch buffers to maintain performance on Inter-Switch Links at distances greater than 10 kilometers, and up to 70 kilometers utilizing selected fiber cable extension mechanisms
- Remote Switch Activation: Provides license key to enable the interconnection of two SAN Switches with a pair of CNTs Open System Gateways across an asynchronous transfer mode (ATM) Wide Area Network.

Note: QuickLoop support is not provided with the Core Fabric Switch. IBM TotalStorage SAN Switches with QuickLoop capability may be used to attach private loop devices in a core/edge fabric.

3.9.2 M12 connectivity

The M12 supports Fibre Channel connectivity for the following:

Servers:

- IBM eServer pSeries and selected RS/6000 servers
- IBM eServer xSeries and selected Netfinity servers
- Other Intel-based servers with Microsoft Windows NT and Windows 2000
- Selected Sun and HP servers

Important: IBM eServer zSeries and S/390 G5/G6 servers with FICON channels are *not* supported.

Storage systems:

- IBM Enterprise Storage Server (ESS)
- IBM FAStT Family of Storage Servers
- IBM TotalStorage Enterprise Tape System 3590 and IBM TotalStorage Enterprise Tape Library 3494
- IBM 3583 Ultrium Tape Library and IBM 3584 UltraScalableTape Library
- IBM 3590-A60 FICON and ESS FICON devices are NOT supported

SAN Switches:

- IBM SAN Switch F16 and F08 (Firmware Version 3.0)
- IBM SAN Switch S16, S08 and SAN Managed Hub (Firmware Version 2.6)

For specific availability dates, configuration options, server models, operating systems levels, and attachment capabilities, consult the Web at:

http://www.ibm.com/storage/FCSwitch

3.9.3 Intelligence within the M12

To improve security and manageability, advanced Brocade Frame Filtering intelligence is built directly into the M12 ASIC technology. This design enables new capabilities such as fabric zoning based on Logical Unit Number (LUN), World Wide Name (WWN), or protocol. Administrators can improve end-to-end performance analysis by measuring resource utilization on a fabric-wide basis. They can also track port traffic levels based on source and destination IDs—simplifying the reporting of, and adherence to, service level agreements.

3.9.4 Open SAN management

The M12 simplifies management by networking core and edge switches under the Fabric OS, the embedded real-time operating system. This enables heterogeneous device connectivity, automatic data routing and rerouting, self-healing capabilities, and scalable connectivity. The Fabric Access layer (the Fabric OS API) provides critical functions for integrating applications within the SAN environment. The API enables software vendors to develop feature-rich management applications that leverage the distributed intelligence in IBM SANs.

3.9.5 Seamless upgrades and investment protection

To help protect existing investments, the M12 provides a seamless upgrade path and backward and forward compatibility with IBM TotalStorage switch, midrange, and port aggregation offerings. As SAN technologies evolve, the M12 multi-protocol architecture is designed to integrate with emerging storage networking protocols such as iSCSI, FC-IP, and InfiniBand. The current design is extendable to future 10 Gb/s technologies with a switch module upgrade rather than a forklift upgrade of the chassis.

3.10 INRANGE FC/9000 Fibre Channel Director

The INRANGE FC/9000 Fibre Channel Director (2042-001) is the core product of an IBM and INRANGE Technologies reseller agreement that adds the INRANGE FC/9000 Fibre Channel Director to IBM's growing list of enterprise-class SAN fabric offerings.

To help provide high data availability across the SAN, IBM now offers the INRANGE FC/9000 Fibre Channel Director, which provides the scalability required by rapidly growing e-business and other mission-critical applications. The Director design is based upon S/390 FICON server requirements for a large number of ports in a single Director (128 ports in a single footprint) and upgradability 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 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.

In the topics that follow, we show some of the major components, both hardware and software, that warrant its inclusion in the IBM portfolio.

3.10.1 INRANGE Director product description

The INRANGE Director can currently be configured from its base of 24 ports, and in 8 port increments, up to 64 ports in a single cabinet. All ports are interconnected to provide full non-blocking performance. Each port has a speed of 100 MB/s, full duplex bandwidth enabling industry leading transmission with a latency of 0.6 to 3 micro seconds.

In Figure 3-12 we show a picture of the INRANGE Director.

.

Figure 3-12 INRANGE FC9000 Fibre Channel Director

3.10.2 Supported attachments

The INRANGE Director provides excellent flexibility by supporting the following types of attachment:

- FICON
- FCP
- FC-IP
- Private and Public Arbitrated Loop (including Public-Private translation)
- Cascaded Directors
- Interoperability port for connection of other vendors switched fabrics

3.10.3 Supported port types

The INRANGE 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 (ISL port more commonly known as an E_Port)
- T_Port (not an ISL port but a Switch interoperability port)
- TL_Port (Private to Public Bridging)
- SL_Port (Segmented Private Loop)

All ports, with the exception of SL and TL, are self discovering. INRANGE directors automatically sense the attributes of individual end-nodes, configuring themselves in any combination of loop, 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 2042 adapts to change dynamically — while the network is still up and running.

3.10.4 Availability

The INRANGE Director provides excellent availability with fully redundant components supporting automatic failover, automatic fault detection and isolation, in addition to call home and pager support to enable rapid problem resolution. Other features include:

- Redundant internal pathing
- Redundant power
- Redundant control
- Non-disruptive SW/FW upgrades
- Passive backplane
- Hot swapping for all FRU components

3.10.5 Scalable capacity

To meet the demands of a growing enterprise, the INRANGE 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 currently be scaled, in increments of 8 ports, up to 64 ports.

Additionally, INRANGE have recently announced the capability to expand the 64-port switch up to a 128-port switch within the same footprint, and have pre-announced a 256-port switch.

IN-VSN Enterprise Manager software provides the interface into one or multiple Fibre Channel Switches and can support up to 16 IN-VSN clients for remote management.

Product component overview

In the topic that follows we will describe the main components of the INRANGE Director, along with showing the physical location of the components.

The components can be seen in Figure 3-15 on page 198.

The Fibre Channel I/O card

The INRANGE Director uses Fibre Channel I/O cards (FIO) to provide the physical connection between the INRANGE Director and the external devices being connected. These cards are commonly referred to as blades. Each FIO blade has 8 ports that terminate at INRANGE certified GBIC compliant devices including Copper, Multi Mode fiber Optics, Single Mode Fibre Optics, and FICON. If required 1x 9 connection modules that can be installed using an RPQ.

The INRANGE Director has a maximum port count of 64, and currently the minimum IBM configuration requires 3 FIO blades which means 24 ports. When one or more of the FIO blades is not required, an FIO blank plate 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. Currently, IBM configurations require longwave and shortwave GBICs to be ordered in increments of eight.

All ports are self-configuring and have a full 64 buffer credit 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 (FSW)” on page 196. The base FIO module logic manages the synchronized switch over to the spare module. Each backplane base FIO 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 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 contains a full copy of firmware, operation code, operating parameters and data packets for routing or special handling.

In Figure 3-13 we show the FIO module.

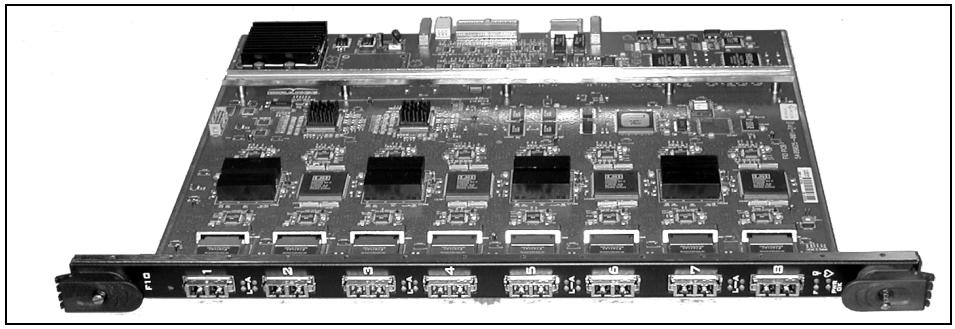


Figure 3-13 8-port FIO module

XCAF FIO Module

The eXtended Credit and Addressing Facility (XCAF) FIO module which will provide support for distances up to 100 Km.

The XCAF blade has the same physical dimensions and GBIC support criterion as the FIO blade.

If the INRANGE Director is configured as an XCAF only device, it is not possible to intermix FIO and XCAF blades. However, if the INRANGE Director is configured as a base FIO system, FIO and XCAF blades can be intermixed.

When running at extended distances there is no requirement to alter the default resource allocation time out values (RA TOV) values.

Fibre Channel Switch Module (FSW)

The Fibre Channel Switch Module (FSW) provides the middle or cross connection architecture of the Director. There are four active and one hot spare FSWs that provide the physical and logical links between the FIOs blades. The hot spare FSW is only provided when the high availability option (feature code 5020) is selected.

FSW modules are hot swapable modules.

Fibre Channel Control Module (FCM)

The remaining module the Fibre Channel Control Module (FCM) provides the command and control interface for the system. It enables the control management software, IN-VSN, to configure, modify and test the INRANGE Director. See Figure 3-14.

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 (feature code 5020) is selected. The FCM modules are hot swapable units.

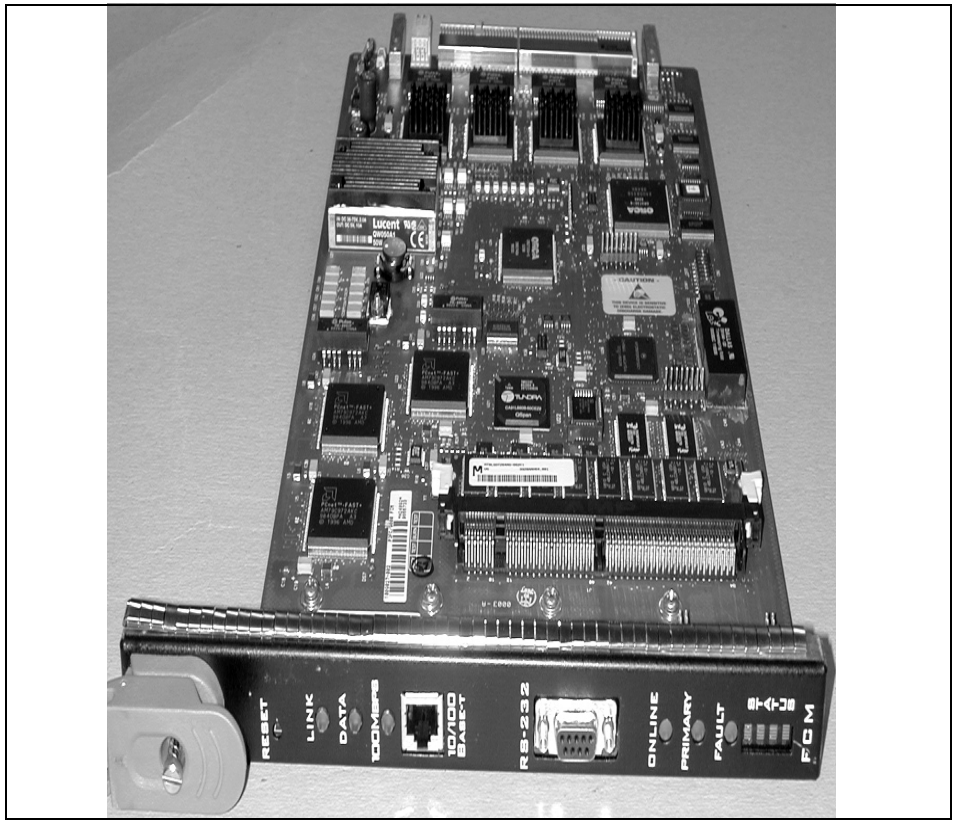


Figure 3-14 FCM module

In Figure 3-15 we show the slot layout for each of the modules described.

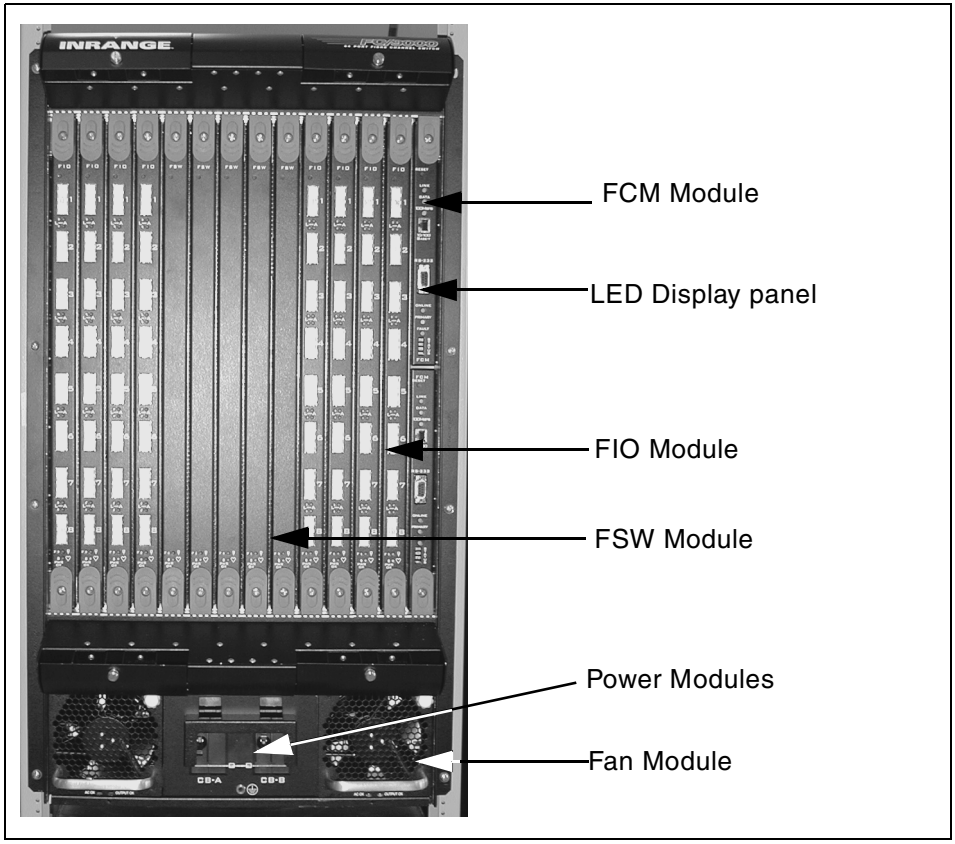


Figure 3-15 Location of modules

The major differences between an FIO blade and FSW module are that FIOs have external ports and have Serializer/Deserializer components; 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. These include:

- Over temperature
- Heart beat
- PWR OK
- F LEDS

For detailed information of the color and LED sequences, refer to the *FC/9000 Fibre Channel Director Maintenance Manual*, 9110774-307 .

Power and fan assembly

There are two hot-swappable power supply assemblies that provide INRANGE Director with full redundancy. 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 VAC nominal; VAC input range is from 180 VAC to 264 VAC.
- Input frequency is 50/60 Hz nominal; frequency range is 47 Hz to 63 Hz
- Output is 48 VDC nominal, plus/minus 5%.

There are four fan modules which either push or draw air to achieve the INRANGE 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 INRANGE Director has a passive backplane that provides the connectivity for all modules.This backplane has the capability to be extended from a 64 port model to a 128 port model.

A backplane upgrade will be a disruptive upgrade. This disruption could be avoided by configuring FIO blades evenly either side of the FSW blades and adding new FIO blades accordingly.

The backplane module also provides the connectivity capability for connecting multiple INRANGE Directors together.

The backplane module is a passive component, and therefore, has no moving components. If physically damaged, it is not a FRU, so the entire INRANGE Director would need to be replaced. The physical replacement of a chassis takes 1 to 2 hours, depending on the number of ports.

ASIC

INRANGE 2042 directors feature a flexible 5th-generation chip 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.

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 2042 is in operation, 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 systems requires additional inter-chassis connections, as well as upgrades to control software, in addition to the second chassis.

The XCA architecture allows the Director to be viewed as a single fabric eliminating the requirement to have any principle and subordinate relationships within the Director.

Cabinet (IBM model 2042-C40)

All INRANGE Directors must be configured within an associated cabinet. A single 40U cabinet can contain up to two INRANGE Director, and the IN-VSN 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.

Management software IN-Vision Enterprise Manager

INRANGE Virtual Storage Network Enterprise Manager (IN-VSN) is browser based software used to manage and control one or more INRANGE Directors.

The IN-VSN suite consists of two components:

- Server software
- Client software

The server communicates with the INRANGE Director, while the IN-VSN client communicates with the IN-VSN server. All user interface is performed by the client software.

IN-VSN management software capabilities include:

- Defining module and port configurations
- Defining zoning parameters
- Monitor alarms and system performance
- Invoking system diagnostics

3.11 McDATA ES-1000 Loop Switch

The McDATA ES-1000 (2031-L00) switch provides an ideal way to consolidate workgroup servers and storage into a seamless, well-managed workgroup or *mini* SAN.

With the ES-1000 switch fabric port, it is possible to create an enterprise storage solution by centralizing data into a single enterprise SAN based on your highly available director-based backbone. In this way, the enterprise storage network is centrally managed from the data center, dramatically reducing management costs. Data is accessible enterprise-wide, and is backed up to central tape libraries, improving data protection. Data center experience and methods now extend from the core to the edge.

3.11.1 Product description

The ES-1000 switch acts as a loop switching hub and a fabric attached switch. The switch provides connectivity between attached Fibre Channel arbitrated loop (FC-AL) devices and a Fabric. This loop function connects workgroup devices into a miniature Storage Area Network (SAN).

We show a picture of the ES-1000 in Figure 3-16.

Figure 3-16 McDATA ES-1000 Loop Switch

The switch also incorporates a bridging function that provides dynamic connectivity between FC-AL devices and McDATA directors participating in a switched fabric. This bridging function allows low-cost or low-bandwidth workgroup (edge) devices to communicate with fabric devices (mainframe servers, mass storage devices, or other peripherals), and ultimately be incorporated into an enterprise SAN environment.

The ES-1000 switch is intended to:

- Implement stand-alone SANs at the departmental and workgroup level. These SANs provide scalability to meet non-disruptive growth requirements, and provide future connectivity to the enterprise SAN.
- Consolidate departmental and workgroup servers to allow centralization of associated storage resources and server communication with the enterprise SAN. These servers are typically low-cost, low-bandwidth devices using the Windows NT or UNIX operating systems.
- Consolidate tape storage devices. Consolidation of workgroup storage and connecting that storage to the enterprise SAN provides better storage resource utilization, increased data protection, and improved data access.

Connectivity

The switch provides device connectivity through 8 hub ports (H_Ports) that attach to device node loop ports. The H-Ports allow for 8 FC-AL ports on the switch. Through the use of cascaded unmanaged hubs, up to 125 FC-AL devices (including hubs) can attach to the switch. The FC-AL standards provides for 127 arbitrated loop physical addresses (AL_PAs). There is a user transparent fabric loop port (FL_Port) and a node loop port (NL_Port) embedded on the switch's control processor card (CTP), each having an AL_PA assigned, leaving 125 for device attachment.

H_Port connectivity is provided through a pluggable fiber optic gigabit interface converter (GBIC) with a shortwave laser transceiver, or a pluggable copper GBIC. GBICs are plugged at front of the switch and they are standard size, fiber cables are connected to GBICs using duplex SC connectors. Copper cables are connected to active copper GBICs with 9 pin DB-9 or 20 pin HSSDC connectors.

The switch also provides connectivity to a switched fabric through a bridge port (B_Port) that attaches to an expansion port (E_Port) of an ED-5000 Director (but not another ES-1000 switch), and through a user transparent FL_Port. This bridge connection forms an interswitch link (ISL) through which a fabric device can communicate with a public loop device attached to the switch. Bridge port connectivity is provided through a pluggable fiber optic GBIC transceiver that can be either shortwave or longwave, and is also located in the switch front panel to the left of the H_Port GBICs.

The switch can be configured to operate in shared mode or switched mode.

Shared mode

When set to shared mode, the switch acts as a hub that implements arbitrated loop topology (although the loop has the physical appearance of a star configuration). When a loop circuit is initialized and established, arbitration protocol ensures only one device attached to an H_Port owns the loop at a time.

The port establishes communication with another device attached to an H_Port (or the B_Port), and half-duplex or full-duplex operation (the default is half duplex) allows the devices to transmit or receive frames at 1.0625 gigabits per second (Gb/s). During frame transmission between these devices, the full bandwidth of the switch is used and no other H_Ports or devices are available for connection. When frame transmission completes, the loop circuit closes and other devices are able to contend for operation (using standard loop arbitration).

Switched mode

When set to switched mode, the switch bypasses full loop arbitration and enables frame transmission through logical connected device pairs. Connections can be established between H_Port pairs, or between an H_Port and FL_Port. Switched mode also allows independent operation of looplets of devices, each connected through an unmanaged hub, and each attached to a single switch H_Port. Because of opportunistic bandwidth sharing, all looplets or connected device pairs operate half duplex or full duplex at 1.0625 Gb/s.

The ES-1000 switch supports connection of public or private fabric loop devices as follows:

Public device

A loop device that can transmit a fabric login (FLOGI) command to the switch, receive acknowledgement from the switch's login server, register with the switch's name server, and communicate with fabric attached devices is a public device. Public devices communicate with fabric-attached devices through the switch's B_Port connection to a director. Public devices support normal fabric operational requirements, such as fabric busy and reject conditions, frame multiplexing, and frame delivery order.

Private device

A loop device that cannot transmit an FLOGI command to the switch, nor communicate with fabric attached devices, is a private device.

Public and private devices are partitioned into two separate address spaces defined in the Fibre Channel address, and the switch's embedded FL_Port ensures private address spaces are isolated from a fabric. The switch does not support any other form of Fibre Channel address conversion (spoofing) that would allow private device-to-fabric device communication.

The switch is controlled by a control processor (CTP) card. The CTP card initializes and configures the switch after the switch is plugged in or a POR is performed, and contains the microprocessor and associated logic that coordinate switch operation. The CTP card provides Intel I960 processor and application specific integrated circuit (ASIC) subsystems that:

- Execute switch firmware and the underlying operating system

- Provide the embedded E_Port and FL_Port that enable communication with a switched fabric and provide fabric services to attached loop devices
- Provide nonvolatile memory for storing firmware, switch configuration information, persistent operating parameters, and memory dump files. Firmware is upgraded concurrently, however, the switch resets during the upgrade, causing Fibre Channel links to momentarily drop and attached FC-AL devices to log out and log back in.
- Provide connections to Fibre Channel ports and enable frame transmission between switch ports without software intervention
- Provide connections to an RS-232 maintenance port and 10/100 Mb/s Ethernet port

The CTP card is not a FRU. If the CTP card fails and cannot be rebooted by performing a POR, the entire switch must be replaced.

3.11.2 High availability features

The FRUs that provide for high availability of the ES-1000 switch are detailed in the following topics.

Power supplies

The switch contains two power supplies that share the electrical operating load. If one power supply fails, the other supply handles the full load. Separate power cord receptacles at the rear of the switch provide facility input power to each supply. For full redundancy, input power for each receptacle should come from a different source.

Fan modules

The switch contains six fans. If one fan fails, the switch can operate indefinitely with the remaining five fans. If two or more fans fail, they must be replaced immediately.

Ports

The switch is delivered with eight H_Ports that support pluggable fiber-optic or active copper GBICs. Any unused H_Port can be used in place of a failed H_Port. To continue device operation, the cable from a failed port is reconnected to an unused operational port. GBICs can be removed, replaced, or relocated without affecting operation of remaining ports. The B_Port is unique and cannot be swapped with another port.

3.11.3 Concurrent firmware upgrades

Since the CTP card provides two memory regions to store firmware, firmware can be upgraded concurrently from the EFC Server. However, the switch resets during the firmware upgrade, causing Fibre Channel links to momentarily drop and any attached FC-AL devices to log out and log back in. Data frames lost during switch reset must be retransmitted.

3.11.4 Serviceability features

The ES-1000 switch, the EFC Manager application, and the ES-1000 Product Manager application provide the following serviceability features:

- Light-emitting diodes (LEDs) on switch FRUs and adjacent to FIbre Channel ports that provide visual indicators of hardware status or malfunctions
- System alerts, event logs, audit logs, link incident logs, and hardware logs that display switch, Ethernet link, and Fibre Channel link status at the EFC Server or a remote workstation.
- Diagnostic software that performs power-on self-tests (POSTs) and B_Port and H_Port diagnostics (internal loopback and external loopback wrap tests).
- Automatic notification of significant system events (to support personnel or administrators) through e-mail messages or the call-home feature
- 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 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.
- Redundant FRUs (GBICs, power supplies, and cooling fans) 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.
- 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.
- Data collection through the Product Manager application to help isolate system problems. The data includes a memory dump file and audit, hardware, and engineering logs.

- SNMP management using the Fibre Alliance MIB that runs on the EFC Server. Up to 12 authorized management workstations can be configured through the EFC Manager application to receive unsolicited SNMP trap messages. The trap messages indicate operational state changes and failure conditions.
- SNMP management using the Fibre Channel Fabric Element MIB (Version 2.0), transmission control protocol/internet protocol (TCP/IP) MIB-II definition (RFC 1213), or a product-specific MIB that run on each switch. Up to 12 authorized management workstations can be configured through the Product Manager application to receive unsolicited SNMP trap messages. The trap messages indicate switch operational state changes and failure conditions.

3.11.5 ES-1000 zoning

The switch supports a name server zoning feature that partitions attached devices into restricted-access groups called zones. For public loop connectivity, this feature is implemented in conjunction with zoning for a fabric director. FC-AL and fabric-attached devices in the same zone can recognize and communicate with each other through port-to-port connections. Devices in separate zones cannot recognize name server information or communicate with each other.

Name server zoning for the arbitrated loop switch is implemented by device WWN only. This contrasts to a fabric director, where zoning is implemented by domain ID and port number or WWN. Zoning of switch H_Ports is not implemented because fabric directors only recognize the arbitrated loop physical address (AL_PA) of the switch's embedded FL_Port. Directors cannot recognize H_Port AL_PAs because:

- The AL_PA assigned to each switch port number is information stored on the switch's CTP card and is not accessible by the fabric.
- The AL_PA assigned to each switch port number is dynamic and can change each time the arbitrated loop initializes.

If an attempt is made to implement ES-1000 zoning by domain ID and port number (by explicitly defining such a zone or merging zone sets), the switch segments from the attached fabric.

3.12 McDATA ES-3016 and ES-3032 Fabric Switches

The McDATA Fabric switches provide an entry point for building a highly available, extensible SAN. It is focused on providing to a mid-range environment, a flexible solution capable of handling current workload and being able to merge into a larger fabric.

Note: In addition to IBM plans to re-market McDATA ED-6064 2 Gb/s (200 MB/s) port cards and upgrade features, IBM also intends to re-market McDATA ES-3232 (2031-232) and ES-3216 (2031-216) 2 Gb/s, 32-port and 16-port Fabric Switches. This will enable IBM to offer a complete range of McDATA 2 Gb/s core-to-edge switching solutions. These 2 Gb/s Fabric Switch models will not be upgradable from the current 1 Gb/s Fabric Switch models.

3.12.1 Product description

The McDATA ES-3016 switch, IBM 2031-016 is shown in Figure 3-17, and provides 16 Fibre Channel Generic ports for attachment to device ports or director expansion ports through fiber optic links. The switch provides full duplex, bidirectional data transfer for all ports.

The ES-3032, IBM 2031-032 switch provides 32 ports with the same characteristics.

Figure 3-17 McDATA ES-3016 switch (top) and ES-3032

The switches have small form factor (SFF) transceivers that are hot pluggable. Fiber optic cables are attached using LC connectors.

There are short wave and long wave laser transceiver available and these can be intermixed as needed.

The 16 Generic ports (G_Ports) on the ES-3016, or the 32 Generic ports on the ES-3032, are available in the front panel. Any port can function as an F_Port when connected to a device or as an E_Port when connected to another switch.

These switches do not support direct connection of arbitrated loop devices, however these devices can communicate with the switches using bridge devices like the McDATA ES-1000 switch.

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 nonvolatile 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.

The CTP is not a FRU and, if it fails, the entire switch must be replaced.

3.12.2 High availability features

The features that ensure high availability of the ES-3016 and ES-3032 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

The switches have 16 or 32 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.

3.12.3 Setup configuration

The switch can be installed in one of three configurations:

- Table or desk top version
- Fabricenter equipment cabinet: One or more switches come pre-installed in a McDATA supplied cabinet.
- Customer supplied equipment rack: One or more switches and the required mounting hardware are shipped to be installed in a customer supplied 19” rack.

The height of the ES-3016 switch is 1 EIA unit (1.75 inches), the height of the ES-3032 is 1.5 EIA unit (2.6 inches).

If the switch is to be attached to the LAN, or to an existing EFC Server LAN, the network address must be set. The network address is set through the RS232 port.

If a new EFC Server is being installed with the switch, the EFC Server setup is the same as for the ED-5000 Director.

Once the ES-3016 or ES-3032 switch is connected and configured to the EFC Server, the switch configuration can be done using the ES-3016 or ES-3032 Product Manager. Where there is no EFC Server available, the switch configuration can be done using the embedded Web server application.

A PC platform running Netscape Navigator 4.6 or higher or Microsoft Internet Explorer 4.0 or higher is required. This PC and the Ethernet LAN segment where the switch is attached must have connectivity through the customer network.

The following configuration tasks can be performed from the Web server:

- Configure the switch ports.
- Configure the switch identification, date and time, operating parameters and network addresses.
- Configure SNMP trap recipients.
- Configure user passwords.

For installations where these switches will coexist with ED-5000 Directors, special consideration should be given to the small form factor LC connectors. There are adapters that can be ordered so existing fiber cables with dual SC connectors can be attached to the new SFF transceivers.

3.12.4 Management software

Management access to the switches is provided through an Ethernet LAN connection to the CTP card.

Management

Management is achieved through the EFC Manager, EFC Fabric Manager, and ES-3016 or ES-3032 Product Manager applications residing in the EFC Server. EFC Manager at release Level 3.01 or higher is required to configure an ES-3032 switch. The ES-3016 and ES-3032 switches do not provide inband management capabilities.

Web management

The ES-3016 and ES-3032 switches also have an imbedded Web server application that provides management capabilities if an EFC Server is not available. This interface supports configuration, statistics monitoring, and basic operation of the switch, but does not offer all the capabilities of the ES-3016 or ES-3032 Product Manager application. The Web server is accessed from any PC attached to the same network and running an Internet browser. Pointing the PC browser to the IP address of the switch, a login screen is presented. Once a valid username and password is entered, the PC browser becomes a management console.

3.12.5 Serviceability features

The ES-3016 and ES-3032 switches provide 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.
- RS232 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.
- SFF transceivers that are removed and replaced without affecting other ports operation.
- Beaconing for quick identification of a switch or specific port by a flashing LED without affecting operation.
- Data collection through the Product Manager application to help isolate problems.
- Unsolicited SNMP trap messages indicating operational state changes and failure conditions sent to authorized workstations.

3.13 McDATA ED-6064 Director

The McDATA ED-6064 Enterprise Fibre Channel Director (2032-064) offers the same kind of enterprise level availability and performance characteristics as the ED-5000 Director, but with double the number of ports and reduced size.

Up to four ED-6064 Directors and the EFC Manager can be installed in a single FC-512 cabinet.

The ED-6064 Director is focused on providing better scalability characteristics in order to cover the growing requirements of today's enterprise level SANs.

3.13.1 Product description

The ED-6064 is a second-generation, 64-port director that provides dynamic switched connections between Fibre Channel servers and devices in a SAN environment.

Note: The McDATA ED-6064 Enterprise Fibre Channel Director (2032-064), resold by IBM, now offers options that enable 2 gigabit (Gb) Fibre Channel technology. The base ED-6064 is now 2 Gb/s capable with the required firmware and control processors (CTP2) standard on all new directors.

New customers can take advantage of 2 Gb/s technology by ordering the new Universal Port Modules (UPM) 2 Gb/s 4-port cards. Existing ED-6064 Directors are upgradable to 2 Gb/s technology with upgrade kits. This upgrade capability provides investment protection for customers by allowing existing customers to upgrade with minimal impact to their SAN.

A total of five new features are being announced for the McDATA ED-6064 Enterprise Fibre Channel Director (2032-064). These features are designed to:

- Introduce 2 Gb/s Fibre Channel technology
- Offer new customers the capability to operate at 2 Gb/s
- Offer investment protection to existing ED-6064 (2032-064) customers who want to migrate to 2 Gb/s capability

For specific configuration support dates and other details on availability, server models, operating system levels, and attachment capabilities, visit:

http://www.ibm.com/storage/mcdata

We show a picture of the ED-6064 and the FC-512 cabinet in Figure 3-18.

Figure 3-18 ED-6064 (left) and FC-512 cabinet

The director implements Fibre Channel technology that provides high performance scalable bandwidth (one gigabit per second), highly available operation, redundant switched data paths, long transmission distances (up to 20 kilometers), and high device population.

The director provides high performance port connections to end devices such as servers, mass storage devices, and other peripherals in a Fibre Channel switched network. Up to 64 Fibre Channel connections are provided through generic ports (G_Ports). The McDATA ED-6064 Director software configures and supports any to any port connectivity.

The McDATA ED-6064 Director offers the following performance and redundancy characteristics:

- Any-to-any non-blocking connections
- High bandwidth: All ports provide full duplex serial data transfer.
- High availability: Redundant configuration of critical FRUs with automatic fault detection and notification.

- Low latency: Less than 2 microseconds between frame transmission at source port and reception at the corresponding destination port.
- Hot FRU replacement
- Concurrent firmware updates
- Service Class 2, Class 3, and Class F support

3.13.2 Attachment

The McDATA ED-6064 Director supports attachment of Open Systems FCP and S/390 FICON servers and devices.

The director supports both point to point and multi-switch fabric topologies, and indirectly supports arbitrated loop topology.

Point-to-point topology provides a single direct connection between two device N_Ports. This topology supports bidirectional transmission between source and destination ports. Through dynamic switching, the director configures different point to point transmission paths. In all cases, connected N_Ports use 100% of the available bandwidth.

A multi-switch fabric topology provides the ability to connect directors (and other McDATA switch elements) through E_Ports and ISLs to form a Fibre Channel fabric. Director elements receive data from a device; and, based on the destination N_Port address, route the data through the fabric (and possibly through multiple switch elements) to the destination device.

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 through the McDATA ES-1000 switch. This switch connects to the director through a bridge port (B_Port).

3.13.3 Planning for 2 Gb/s

Each Director provides up to 64 ports of non-blocking fibre channel switching capability. The minimum configuration of the ED-6064 (2032-064) stays the same, 24 ports using six 4-port modules (now either FPMs and UPMs). Scalability continues to be in increments of four ports, utilizing 4-port modules, up to the maximum of 64 ports. The three new UPMs are 2 Gb/s, 4-port modules and are available in these combinations:

- Four shortwave optical ports
- Four longwave optical ports
- Three shortwave and one longwave optical ports

New customers who purchase a McDATA ED-6064 (2032-064) need only configure these new 2 Gb/s UPMs to enable the Director to run at 2 Gb/s. The base machine is now 2 Gb/s capable, because the required firmware and new control processors (CTP2) come standard with the ED-6064s. To run at 2 Gb/s, all the 4-port modules installed in the ED-6064 must be 2 Gb/s UPMs. The 2 Gb/s ports are auto-sensing and will negotiate to operate at the speed (1 Gb/s or 2 Gb/s) of the attached device, on a port-by-port basis.

For existing customers who have already purchased a McDATA ED-6064 (2032-064), 2 Gb/s operation requires that all of the existing 1 Gb/s FPM 4-port modules be upgraded to the new 2 Gb/s UPM 4-port Modules. Operation at 2 Gb/s also requires installation of either Upgrade Kit depending upon configuration of the director.

Important: Both 1 Gb/s (FPM) and 2 Gb/s (UPM) port cards can be intermixed in the McDATA ED-6064 Enterprise Fibre Channel Director. To enable 2 Gb/s operations, all port cards must be UPM cards. If there are any 1 Gb/s FPM 4-port modules installed, the entire Director will operate at 1 Gb/s.

Operation of the ED-6064 at 2 Gb/s requires:

- ECFM must be at release 6.0 or higher.
- Firmware must be at release 2.0 or higher.
- CTP cards must be replaced with CTP2 cards.
- All 1 Gb/s 4-port cards must be replaced with 2 Gb/s 4-port cards.

3.13.4 Port types

When connected to a device, a G_Port behaves as a fabric port (F_Port). When connected to another director (or other managed McDATA product) in a multi-switch fabric, a G_Port behaves as an expansion port (E_Port).

Up to 32 ports can be used as expansion ports, that is half the maximum total number of ports.

Each Fiber Port Module (FPM) card provides four Fibre Channel connections through duplex small form factor (SFF) pluggable fiber optic transceivers.

Shortwave laser transceivers are available for transferring data over multimode fiber optic cable. Longwave laser transceivers are available for transferring data over single mode fiber optic cable. Transceivers in a single FPM card can be mixed as needed.

Fiber cables attach to the SFF transceivers on the FPM card using duplex LC connectors.

3.13.5 Scalable configuration options

The director is configured from a minimum of six FPM cards (24 ports) to a maximum of 16 FPM cards (64 ports).

Two options are available for fiber port module (FPM) cards:

- Four duplex small form factor pluggable optic transceivers. These can be all shortwave laser transceivers, all longwave laser transceivers, or a mixture of each.
- Four duplex fixed optical transceivers. These transceivers can all have the same parameters, such as cable connector and optics (shortwave or longwave), or each transceiver may have different parameters.

High availability features

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 FPM cards.

Power supplies

The director contains two power supplies that share the electrical operating load. If one power supply fails, the other supply handles the full load. Separate receptacles at the rear of the director provide facility input power to each supply. For full redundancy, input power for each receptacle should come from a different source.

Fan modules

The director contains two fan modules, each containing three fans (six fans total). If one or more fans in a module fail, the redundant fan module provides cooling until the failed module is replaced. If the second fan module fails, software shuts off power to the director to prevent system damage.

CTP cards

The director is delivered with two CTP cards. The active CTP card initializes and configures the director after power on and contains the microprocessor and associated logic that coordinate director operation. A CTP 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 CTP card without switching off power or affecting operational fiber optic links.

Each CTP 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.

Each CTP 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 CTP card provides nonvolatile memory for storing firmware, director configuration information, persistent operating parameters, and memory dump files.

Director firmware is upgraded concurrently (without disrupting operation). The backup CTP card takes over operation if the active card fails. Failover from a faulty card to the backup card is transparent to attached devices.

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.

FPM cards

The director is delivered with a minimum of eight FPM cards (32 ports). Any unused Fibre Channel port of the same type can be used in place of a failed port.

Pluggable transceivers can be removed and replaced concurrently with other port operations.

Note: Spare port cards do not automatically fail over and provide link operation after a port card failure. To continue device operation, the fiber optic cable from a failed port is reconnected to an unused operational port. When storage ports are moved, the operating system may need to have the path reconfigured.

An FPM card is a concurrent FRU and can be added or replaced while the director is powered on and operating.

Power module assembly

The power module contains two AC power connectors and the power circuit breaker. Included in this module is also a 9-pin serial connector used for a local terminal or remote dial in attachment for maintenance purposes. This serial interface is also used to setup director network addresses.

The module is a non-concurrent FRU, and the director must be powered off prior to scheduled removal and replacement.

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 ED-6064 Director is ready to support 2 Gb/s operation. So when 2 Gb/s becomes available, it will only require replacing FPM cards and GBICs.

Management software

The McDATA ED-6064 Director provides for outband management access in the following ways:

- Through the EFC Server attached to the director's CTP card.
- 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.

Inband management console access (through a Fibre Channel port) is provided by enabling user specified features that allow Open Systems or FICON host control of the director. The features are mutually exclusive; only one can be installed at a time.

Web management

The embedded Web server interface provides a GUI accessed through the LAN (locally or remotely) to manage the McDATA ED-6064 Director.

This interface is available with director firmware Version 1.2 (or later) installed, and does not replace nor offer the management capability of the EFC Manager and Product Manager applications (for example, the Web server does not support all director maintenance functions). In addition, the Web server interface manages only a single director.

Web server users can perform the following:

- 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, 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 embedded Web server interface can be opened from a standard Web browser running Netscape Navigator 4.6 or higher, or Microsoft Internet Explorer 4.0 or higher.

4

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.

4.1 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, “Identifying your business and technology goals” on page 3.

TCO/ROI

The TCO and ROI are two important considerations in your SAN design. We also cover those aspects in Chapter 1, “Identifying your business and technology goals” on page 3.

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.

4.2 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.

4.2.1 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 meshed 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.

4.2.2 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?

4.2.3 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 100 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), which operates at either 132 MB/s or 264 MB/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 achieve the required performance, since two would overrun the bus speed. If you attach a second HBA, it should only be for redundancy purposes.

Our recommendation is to install one adapter per system bus.

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.

4.3 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.

4.3.1 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 colored by the degree of risk which you associate with the implementation of an enterprise wide SAN today. The technologies are still relatively new; industry standards in some key areas are still to be agreed; not all server platforms can easily participate in Fibre Channel configurations, and the rate of change is extremely rapid. However, in the last few years substantial progress has been made.

The fact is that the majority of SANs which have been implemented today are 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 practice then, the bottom up approach seems to be the most pragmatic. Solve specific application needs now, to deliver value to your organization. This does not mean that you should not establish some common guidelines or standards regarding the purchase of equipment within the enterprise. This will facilitate inter-operability in the future, and avoid dead end investments which cannot be integrated in a larger SAN environment as you expand the topology in the future.

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.

A strong trend is towards switched fabric environments. This is because a fabric offers the greatest flexibility and scope for the future. You may choose to install small FC-AL topologies now, for reasons of cost, or because the application being addressed is small scale today. If so, there is good logic in selecting hardware such as a hub. This gives you flexibility for the future, such as more sophisticated functionality, manageability, and upgrade-ability, and compatibility within a family of fabric devices,.

Remember that FC-AL is not designed for high performance. 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, but is more expensive at the outset.

4.3.2 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.

4.3.3 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 across the SAN it needs one hop to get from switch to switch. This time is defined as the latency of the link. This time is approximately 2 microseconds across the switch. This time usually includes the physical path from one switch to the another.
- 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. With this type of over-subscription we may see this in a meshed fabric. 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 really in place. That means, for example, more servers try to talk to the storage device over the same ISL. In this case, multiple servers are contending for bandwidth. Because the link has limited bandwidth, the servers will be throttled down to the total 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. We cover the technical details about this protocol in Chapter 2.7, “Fabric Shortest Path First” on page 78.

  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.

4.3.4 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 4-1.

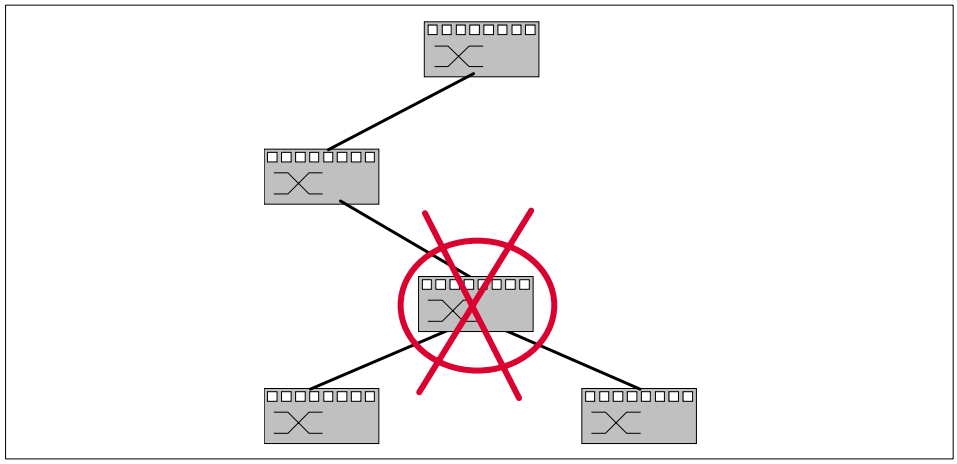


Figure 4-1 Single fabric non resilient

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 4-2.

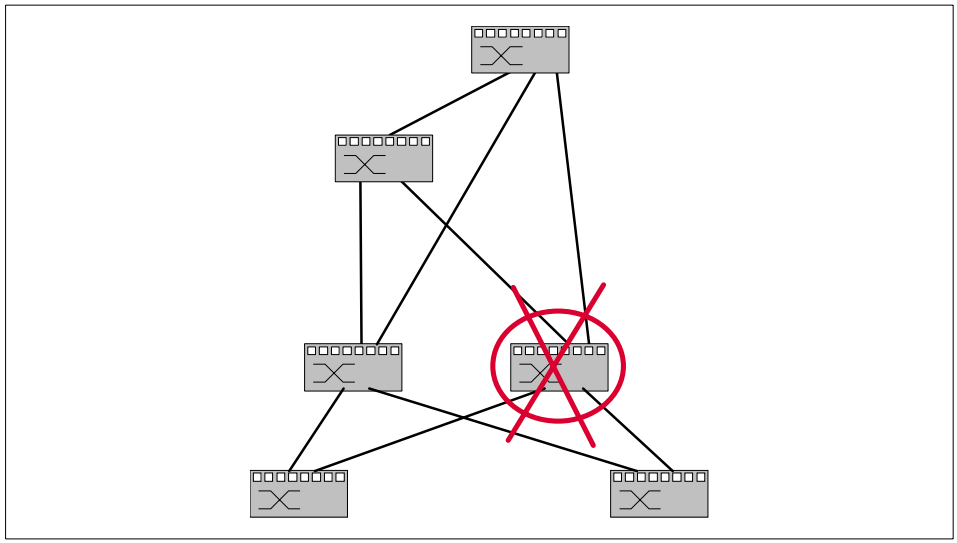


Figure 4-2 Single fabric - resilient

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 4-3.

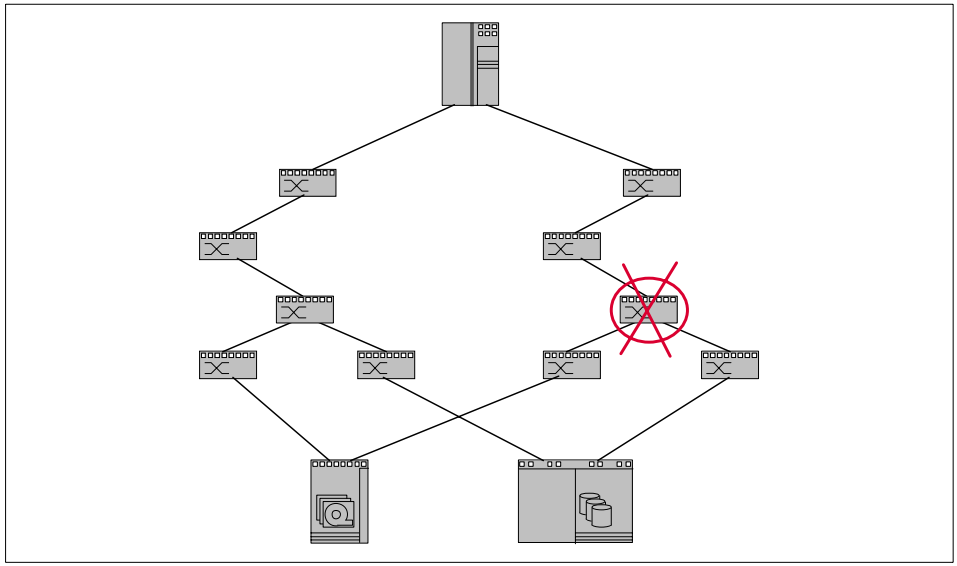


Figure 4-3 Redundant fabric non resilient

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 4-4.

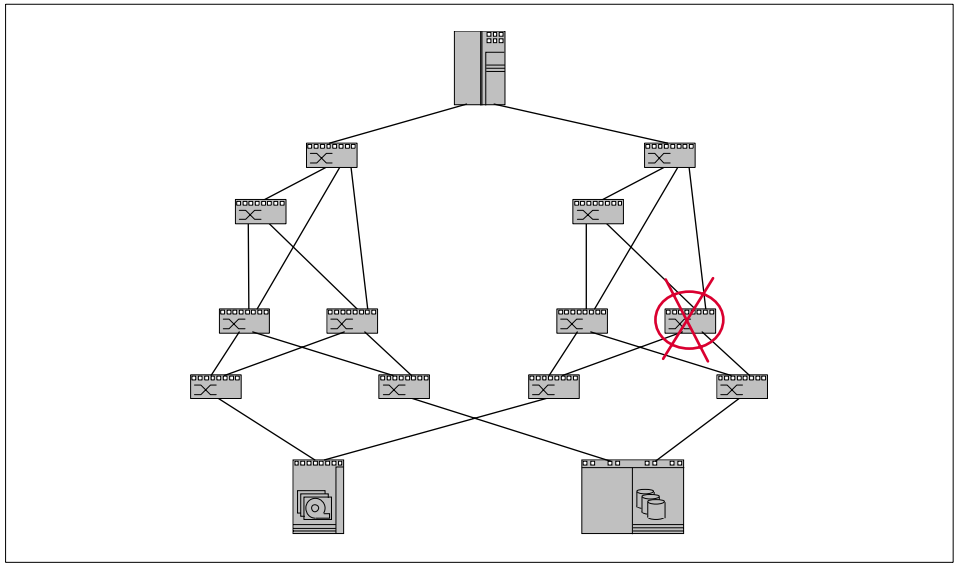


Figure 4-4 Redundant fabric resilient

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.

4.4 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 6227 supported for IBM RS/6000 servers is supplied by Emulex. This chapter will still provide value to readers of these platforms, as we provide an overview of each HBA and detail error diagnostics tips that would still apply to these platforms.

We provide an overview of the IBM supported HBAs and highlight any unique functions the card may have.

4.4.1 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

Product specifics

In the topics that follow, we look at three vendors that are associated with the IBM portfolio of HBAs.

4.4.2 Emulex

IBM currently supports the Emulex LP7000E and LP8000 Fibre Channel HBAs with Emulex Port driver (NT Version 1.27A3 and for W2K 1.27A5). The port driver version 2.10 and miniport driver Version 4.53 conform to the SNIA API HBA standards.

LP7000E

The Light Pulse LP7000E, a second generation Fibre Channel PCI host bus adapter, uses the Emulex Superfly chipset, a 266 MIPS onboard processor and high speed buffer memory. The LP7000E features a 32–bit PCI interface.

The 32-bit memory acts as a frame buffer and enables the LP7000E to achieve its high performance throughput.

The 1 Gb/s LP7000E provides features, including switched fabric support using F_Port and FL_Port connections, full-duplex data transfers, high data integrity features, support for all Fibre Channel topologies, and support for service classes 2 and 3.

LP8000

This is the third generation Fibre Channel PCI host bus adapter. The LP8000 uses the Dragonfly ASIC with a 266 MIPs onboard processor. The major difference between the LP7000E and LP8000 is that the 8000 card uses a 64-bit PCI interface and provides significantly higher performance.

Similar to the 64 buffer credit associated with a longwave port in a switch or director, the 64-bit interface enables the LP8000 to sustain high performance over a distance of up to 10 KM. This performance benefit will only be realized for a direct server to device attachment or by using a switch or Director that supports 64 buffer credits. Like the LP7000E, this buffer capability improves the performance of the card.

Emulex special features

In addition to the 64-bit interface, other features unique to either of the Emulex HBAs are included here.

Persistent binding

This function, available with the port driver and all UNIX drivers, allows a subset of discovered targets to be bound between a server and device. Binding can be by WWNN or WWPN. Once a configuration has been set, it will survive reboots and hardware configuration changes, as the information will be held in the registry of the server.

Note: This feature is especially important for Windows operating systems. For example, if you lose one device in a redundant SAN configuration, the order of devices at next reboot can be completely different. Because of this all applications which rely on a particular device order will stop functioning.

For example, this function may be useful for legacy tape software that expects to see its tape devices at the same SCSI target ID at all times. By binding the tape device's WWN to a SCSI target ID we are able to satisfy this criteria.

LUN mapping

This function allows LUNs that are beyond NT's LUN range to be bound permanently to an NT LUN number.

4.4.3 JNI

IBM currently supports the JNI Fibre Star FCI-1063-N and FC64-1063-N Fibre Channel HBAs.

FCI-1063-N PCI HBA

The JNI FCI-1063 is a 32-bit PCI-to-FC Adapter with integrated non-OFC short-wave fiber optic interface. The FCI-1063 comes with EZ Fibre, JNIs proprietary software for managing and configuring a Fibre Channel installation.

FC64-1063-N SBus HBA

The JNI FC64-1063-N is a 64-bit SBus-to-FC Adapter with Integrated Non-OFC Short wave Optical Interface. The FC64-1063-N provides a full-duplex Fibre Channel connection between SBus Sun servers and SAN devices.

LUN level masking

With the JNI EZ Fibre software you have the ability to implement LUN-Level Zoning at the host bus adapter level. The LUN-Level Zoning option comes standard on all JNI adapter cards and offers the following advantages:

- Shorter boot-up time by controlling device discovery process in multiple CPU environments
- Flexibility (dynamic allocation, the ability to change drives on the fly)

- Allocation of backup resources (one can allocate backup within a tape array itself)
- 0% performance loss (since the zoning has been pre-configured, the operating system does not incur the overhead of determining resource availability)
- Enhanced security JNI's host-based LUN-Level Zoning can be used to.

Boot BIOS

The FC64-1063 and FCI-1063 do not support external boot.

4.4.4 QLogic

IBM currently supports the QLogic QLA2100F and QLA2200F Fibre Channel HBAs, with limited support for the QLA2300. QLogic Fibre Channel HBA products have achieved SANMark certification, which is the industry standard for device compatibility. The HBAs are based on a single chip architecture, providing high reliability and low power consumption. In addition, they are able to boot to an external FC storage device, either on a local loop or through a switched fabric.

QLA2100/SANblade 2100

QLogic's first generation of Fibre Channel HBA products is the QLA2100 family. These cards are currently available as either fixed copper (QLA2100/66) or optical (QLA2100F/66) nodes, IBM only supports the optical version. The /66 indicates the maximum supported PCI bus speed. These are 64-bit PCI cards that also function in 32-bit PCI environments. The HBAs use the ISP2100 ASIC.

The 2100 HBAs operate at 1 Gb/s data rate on the Fibre Channel medium. These products support either FC-AL or switched fabrics using an FL_Port connection. IBM only supports this HBA for legacy attachments.

QLA2200/SANblade 2200

QLogic's second generation of performance optimized Fibre Channel HBAs is the QLA2200 series. These boards are based on QLogic's ISP2200 ASIC. As with the 2100 series, these are PCI cards that operate in 33 and 66 MHz as well as 32 and 64-bit environments. The 2200 series supports FC-AL as well as switched fabric via F_Port and FL_Port connections. Additionally, the QLA2200 series is able to support IP protocol. The QLA2200 series supports Class 2, 3, and FC Tape. IBM supports the QLA2200F for connections to the ESS and FAStT.

QLA2300/SANblade 2300

QLogic's third generation of performance optimized Fibre Channel HBAs is the QLA2300 series. These boards are based on QLogic's ISP2300/ISP2310 ASIC. As with the 2100 and 2200 series these are PCI cards that operate in 33 and 66 MHz as well as 32 and 64-bit environments. The 2300 series supports FC-AL as well as switched fabric via F_Port and FL_Port connections. Additionally, the QLA2300 series is able to support IP protocol. The QLA2300 series supports Class 2, 3 and FC Tape. The HBA will auto negotiate Fibre Channel bit rate to 1 Gb/s or 2 Gb/s.

IBM supports the QLA2300F for connections to the ESS and FAStT on selected operating platforms and systems only. The supported adapters may change often. Therefore, it is important that *before* purchasing adapters or any SAN component, refer to the Supported Server and HBA Version matrices available for FastT and ESS products at:

`http://ssddom02.storage.ibm.com/techsup/webnav.nsf/support/home`

For a complete list of QLogic drivers refer to:

`www.qlogic.com`

4.4.5 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 4-5.

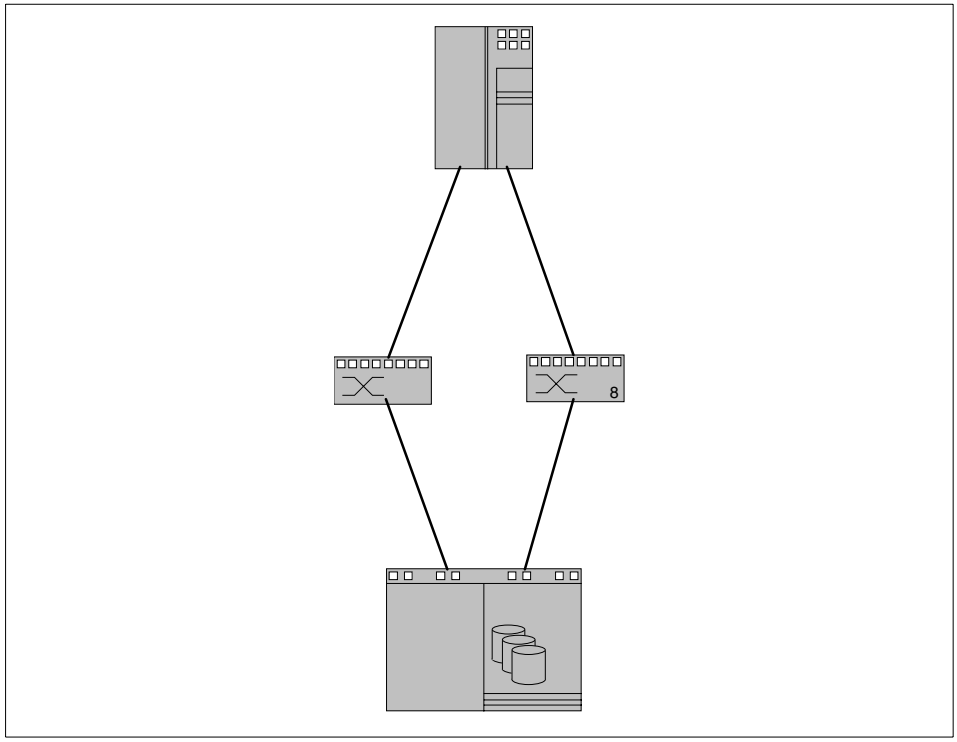


Figure 4-5 Multiple paths to the same LUN

As shown in Figure 4-5, 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 4-6.

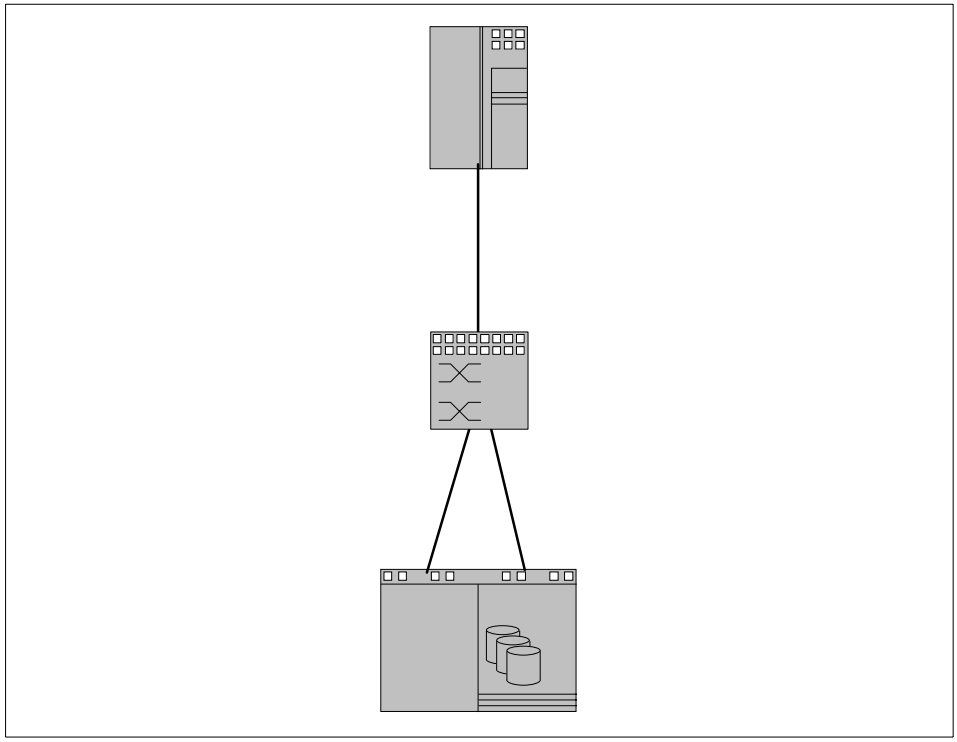


Figure 4-6 Multipath in single fabric SAN

The reason why you have two paths in a single fabric SAN is explained in Chapter 2.11, “Multipathing” on page 99.

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.

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.

Some of the most popular multipathing products are:

1. IBM SDD: Subsystem Device Driver, available for free with IBM Enterprise Storage Server, and supported on multiple platforms:
   - AIX
   - Windows NT/2000
   - Linux
   - SUN Solaris
   - HP UX
2. Veritas Logical Volume Manager includes support for multipathing. It is a priced, add on product for:
   - HP-UX
   - Windows NT/2000
   - Linux
   - SUN Solaris
3. QLogic HBA High Available Package includes support for multipathing and failover. It is a priced, add on product for:
   - Windows NT/2000
   - Novell Netware

4.4.6 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.

4.4.7 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 like:

- Tivoli Storage Network Manager
- SANavigator from McDATA
- SANSurfer from QLogic
- SANPoint from Veritas
- Fabric manager from Brocade

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

4.5 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 4-7.

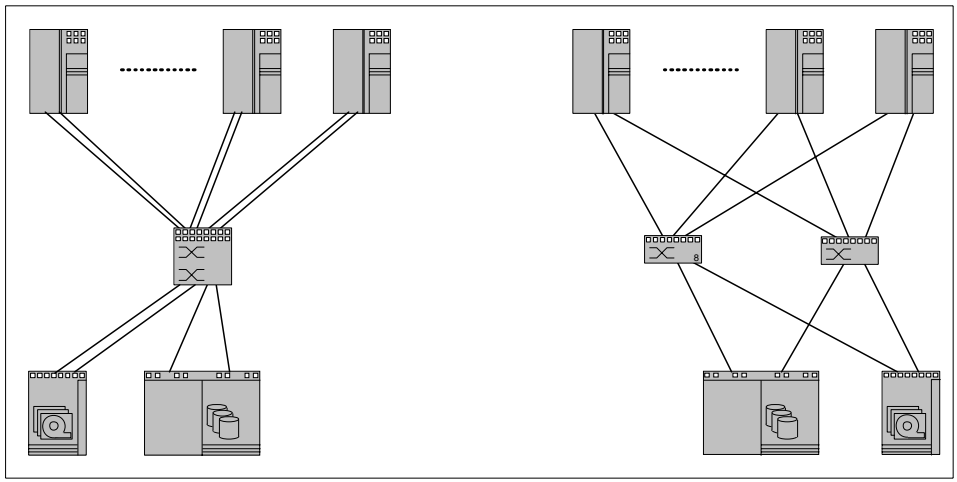


Figure 4-7 Director class or switch dilemma

In this simple scenario, we will try to outline the important parameters which can affect the SAN design you will ultimately chose. Let us assume that we identified the amount and the type of fabric ports as described in Chapter 4.2, “Existing resources needs and planned growth” on page 225, 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 4-7, 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 product 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. Because 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 allow you to upgrade their firmware at the level of an individual ASIC.
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 outbound 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. Today, director class products and some switches 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 4-8.

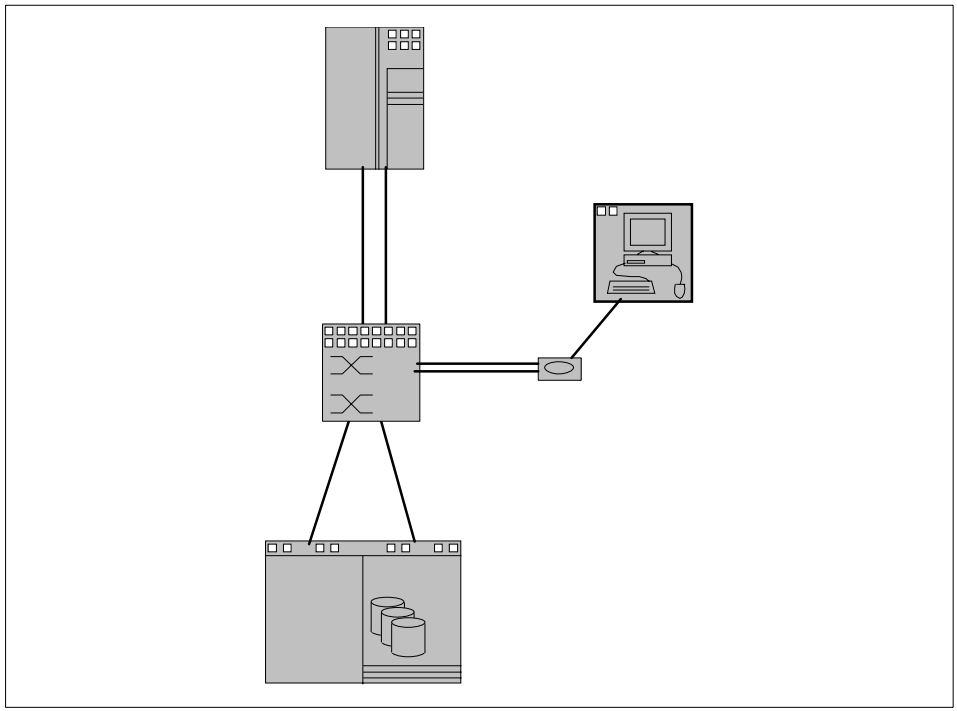


Figure 4-8 External managing of director class product

9. For director class products or switches 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 4-9.

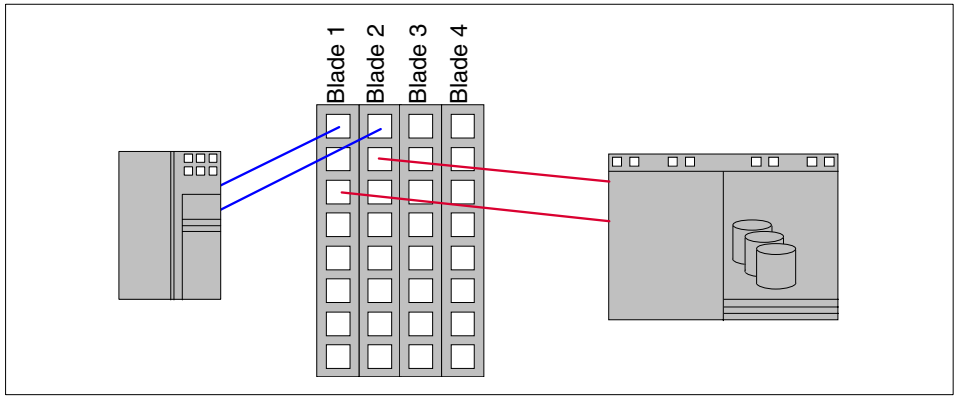


Figure 4-9 Ports on different blades

11. For performance reasons, it is recommended that you group the devices which will talk to each each other on the same blade if possible.
12. When you are using director class products in a similar manner as that shown in Figure 4-9, 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 4-10.

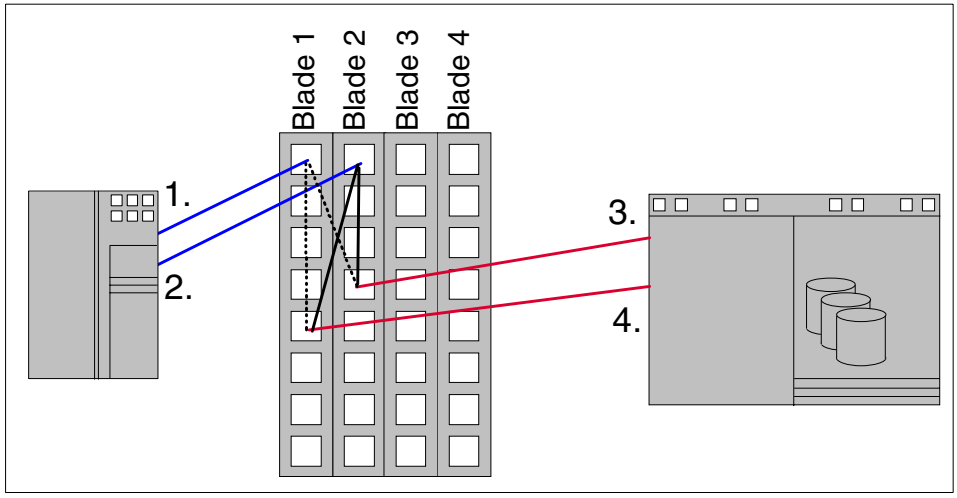


Figure 4-10 Routes in director class product

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 - 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 4-11.

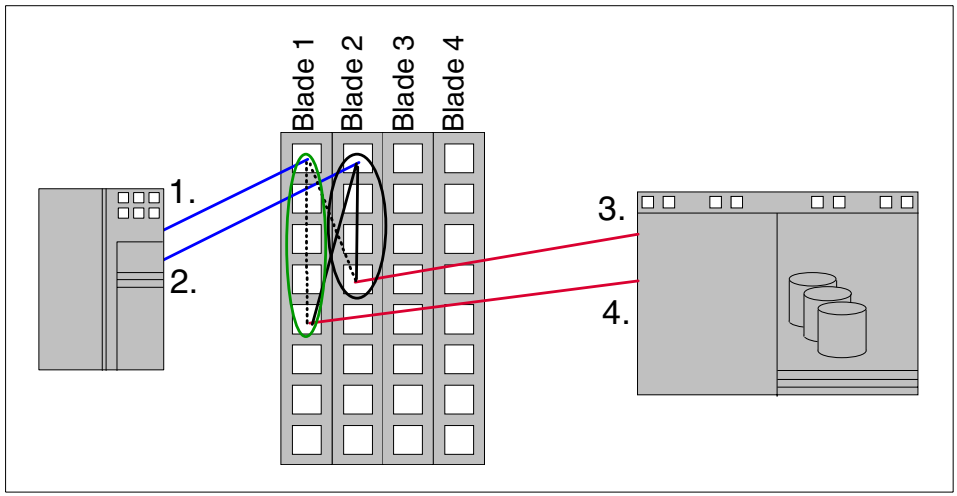


Figure 4-11 Blade zoning

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 comes along because director class products have a single backplane.

13. Older director class products and switches also sometimes use fixed ports (they are soldered) on the blades. In such a case, if one port fails, the whole blade has to be replaced as opposed to the hot pluggable port design, where you just replace the failing port.

14. Director class products usually have an option for *in box* 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.)
- More of a complex to understand topology
- More power, heating
- 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)

In Figure 4-12, we show the comparison between a 64 port director class product and a 72 port full meshed fabric built from 16 port switches.

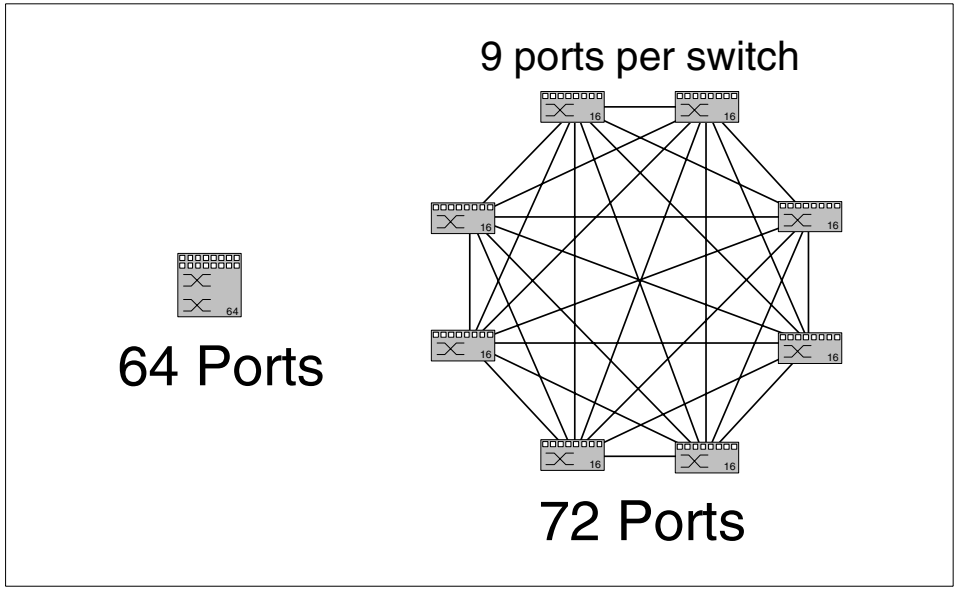


Figure 4-12 Director class product versus full meshed switch fabric

As you can see in Figure 4-12, we have 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 2.7, “Fabric Shortest Path First” on page 78, all communication will go through one ISL, and this gives a 9:1 over-subscription ratio.

In Figure 4-13, we show an example of how to provide 64 ports using six 16-port switches.

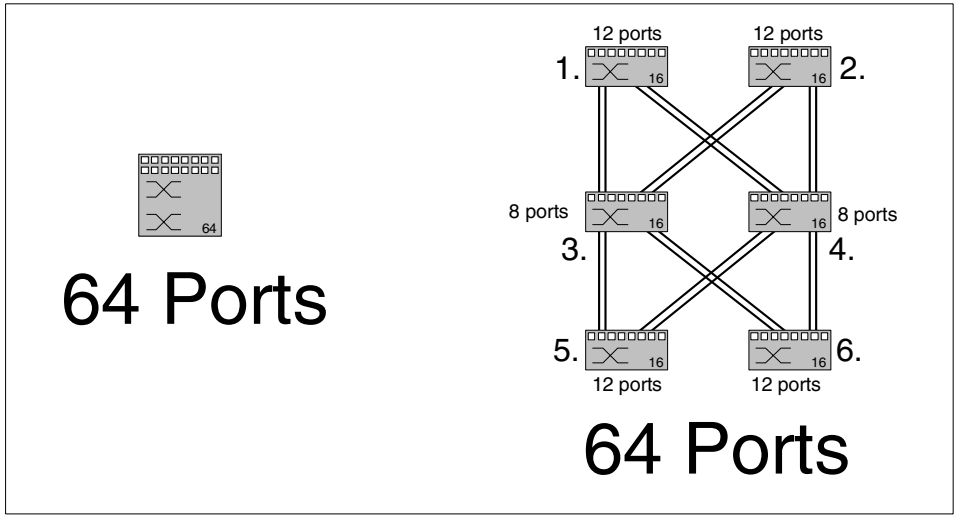


Figure 4-13 Director class 64 ports against 64 ports switch fabric

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.

15. 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 4-14, we show such a solution with the introduction of loop switches to support FC-AL storage devices.

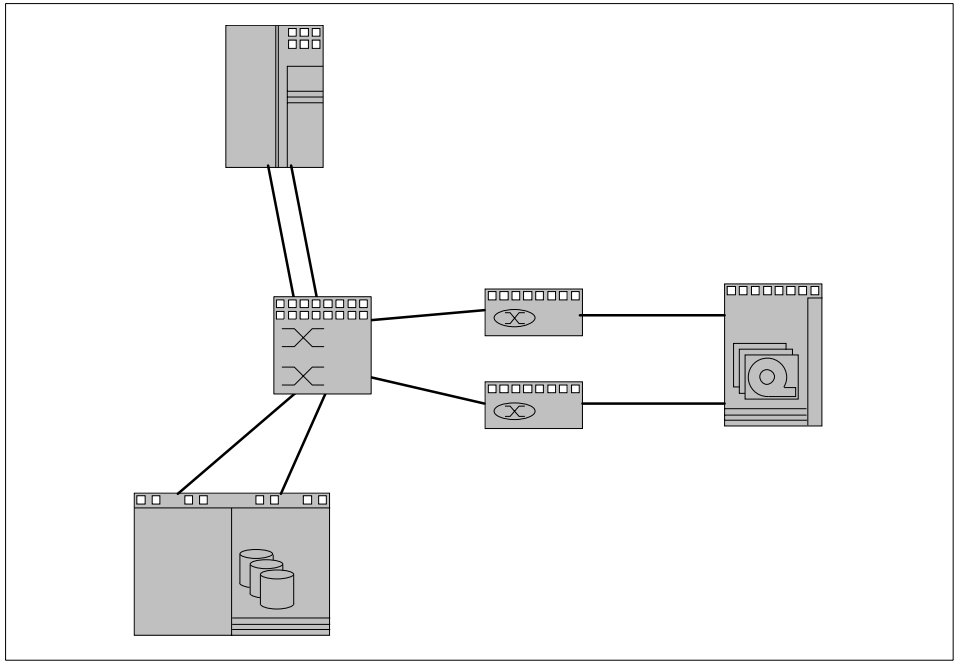


Figure 4-14 Adding tapes to the director class SAN

In the example in Figure 4-14, we are assuming that the tape device is only FC-AL capable. Today, almost all native tape attachments are FC-AL, but 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 4-15.

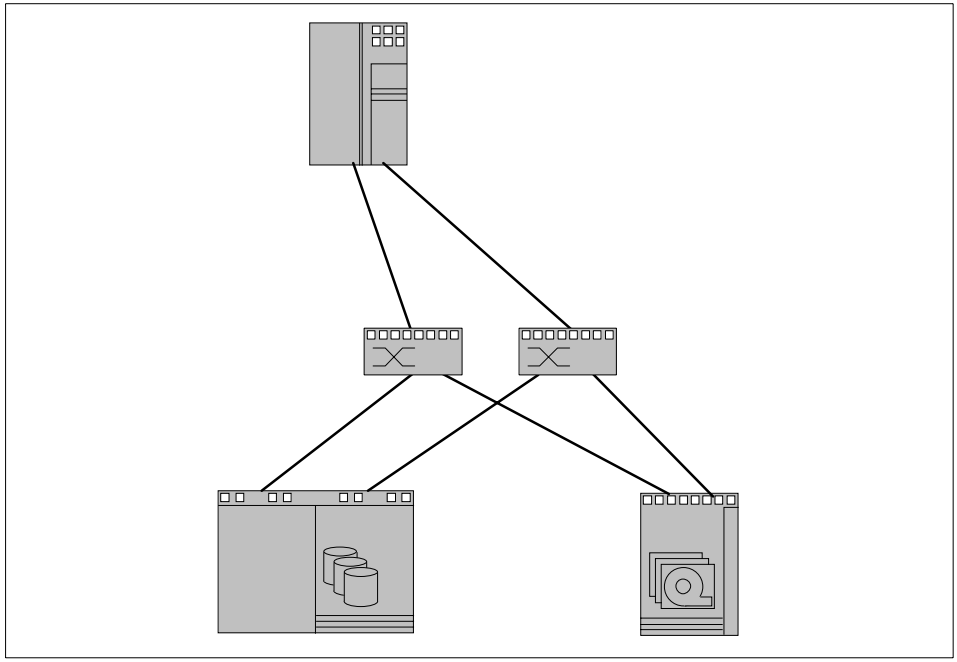


Figure 4-15 Switches with loop support

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.

16. 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.

17. 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.

18. In Figure 4-16, we show an example of implementing an ISL between two redundant fabrics.

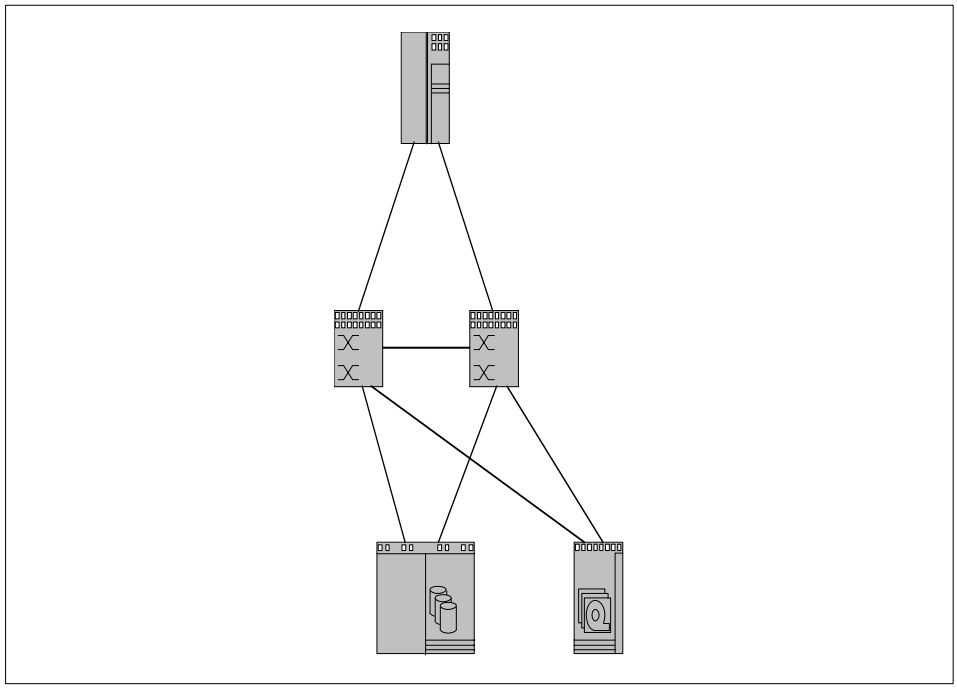


Figure 4-16 ISL between two redundant fabrics

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. This type of zoning can be accomplished by using soft zones for tape connectivity and hard zones for disk connectivity.

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 4-17.

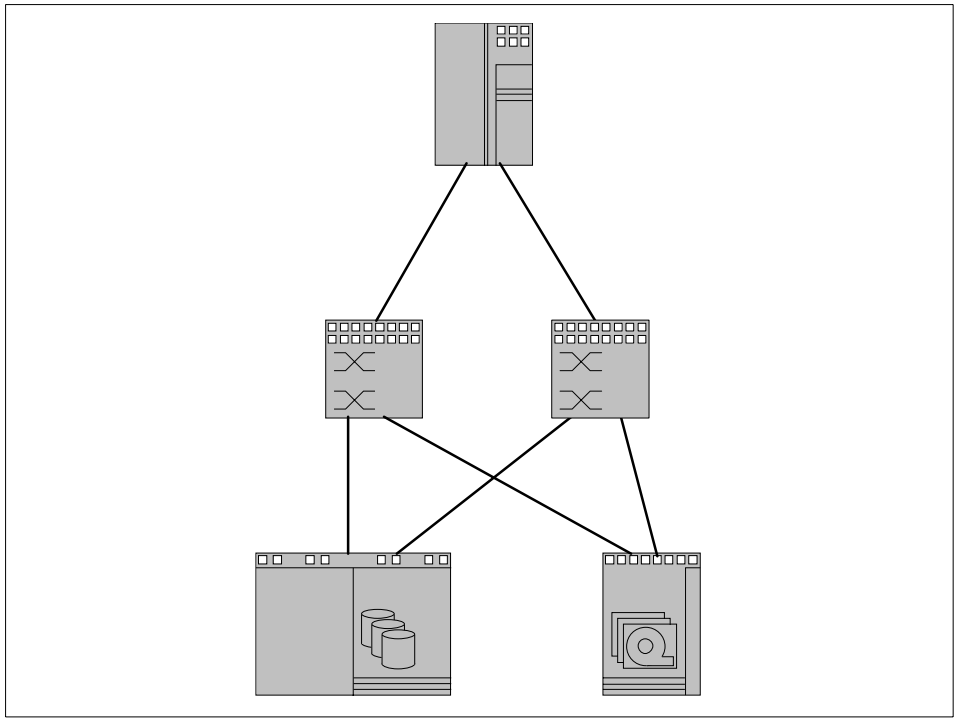


Figure 4-17 Director class with FC-AL support solution

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 director class products

We show this design in Figure 4-18.

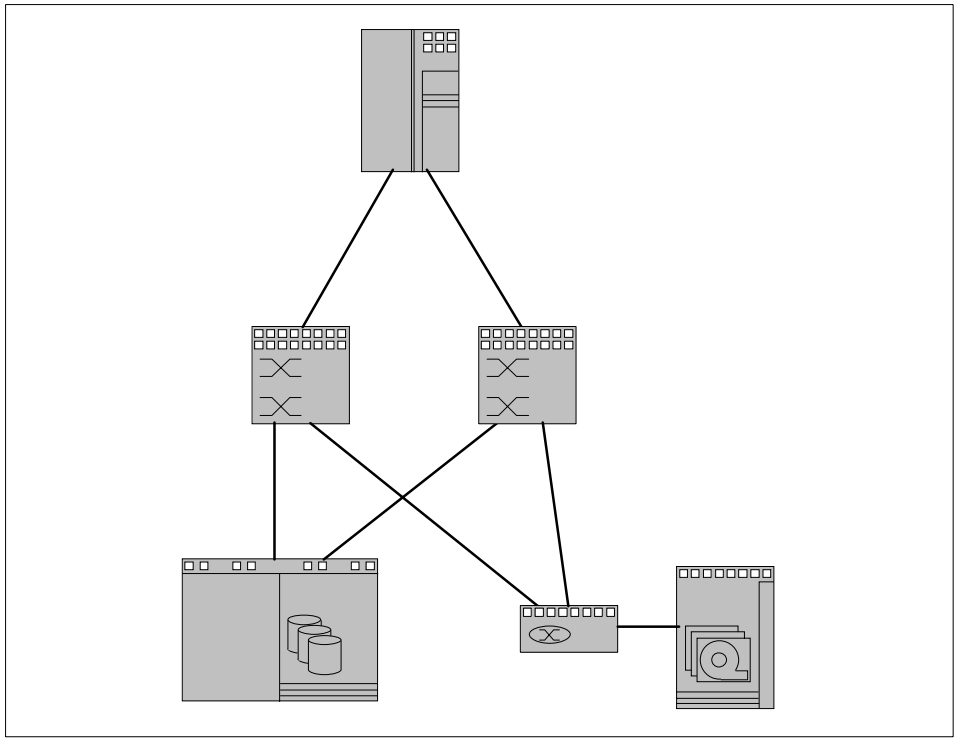


Figure 4-18 Director class without FC-AL support

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 4-19.

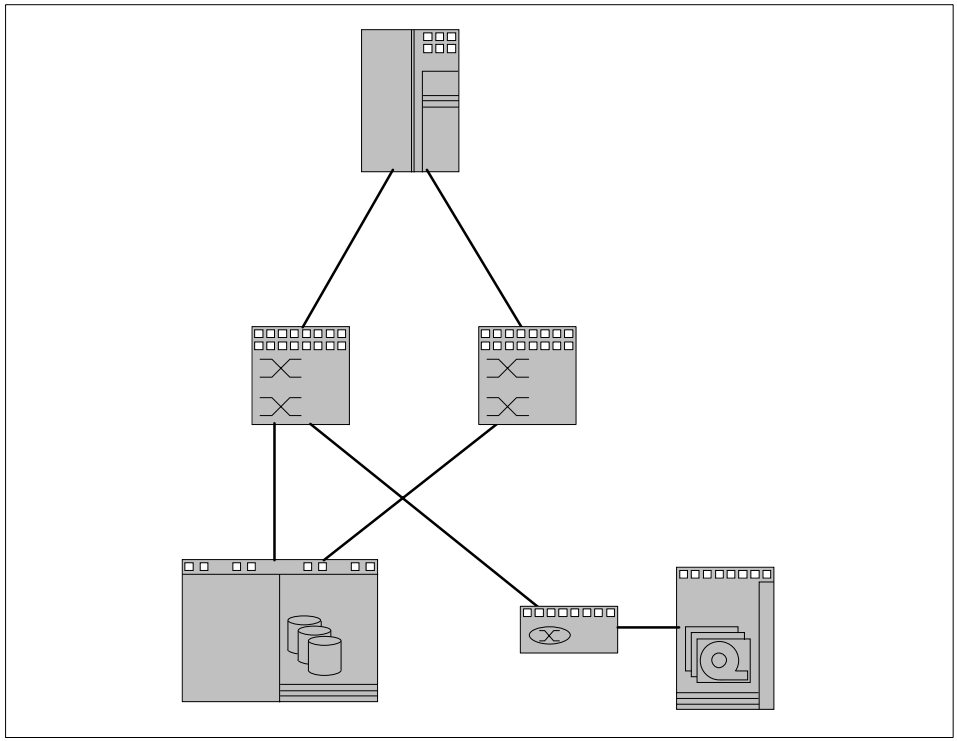


Figure 4-19 Edge switch with only one connection

Design with one director class product supporting FC-AL

We show this design in Figure 4-20.

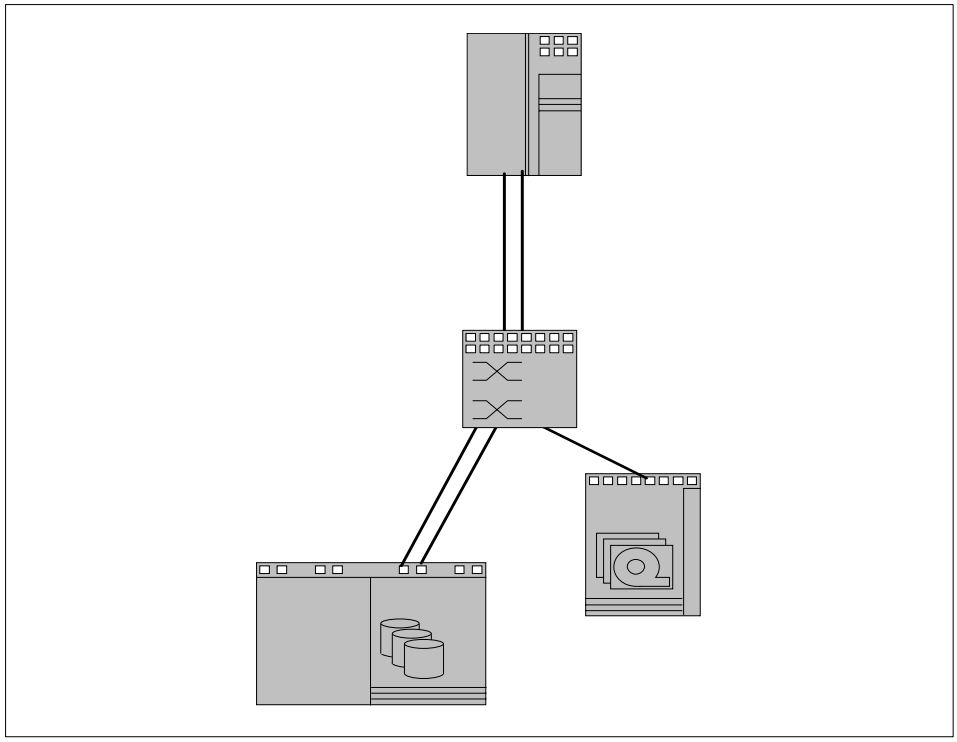


Figure 4-20 One director class product solution

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. The design has the following attributes:

- Single point-of-failure in the SAN
- 99.999% available components
- Support for FC-AL devices on all 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

Design with one director class product without FC-AL support

We show this design in Figure 4-21.

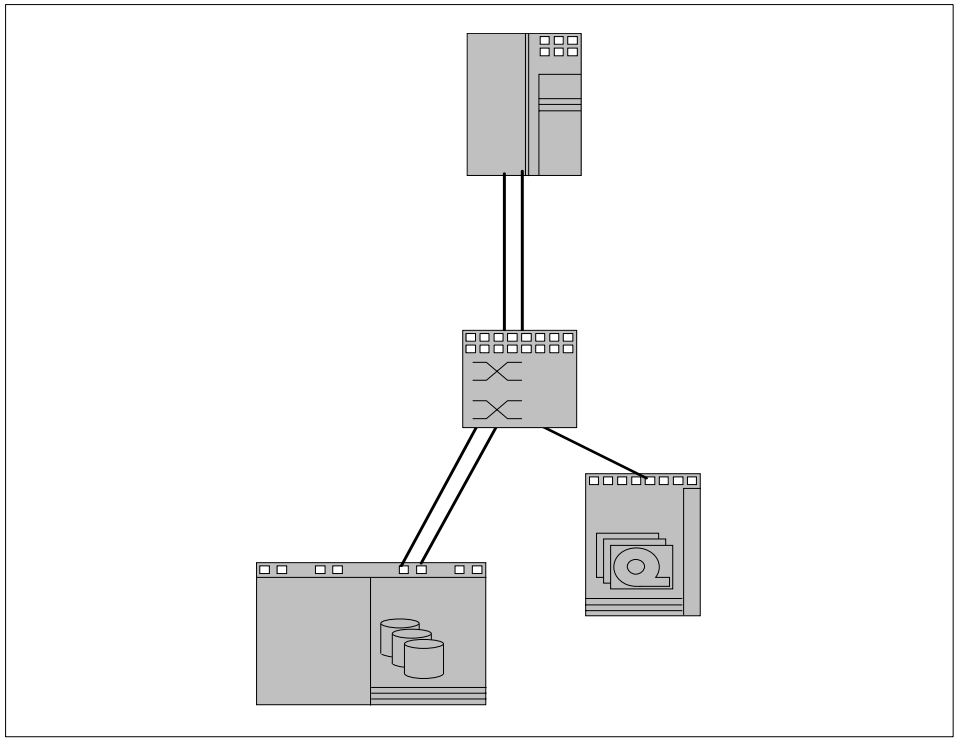


Figure 4-21 One director class product solution without FC-AL support

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 4-22.

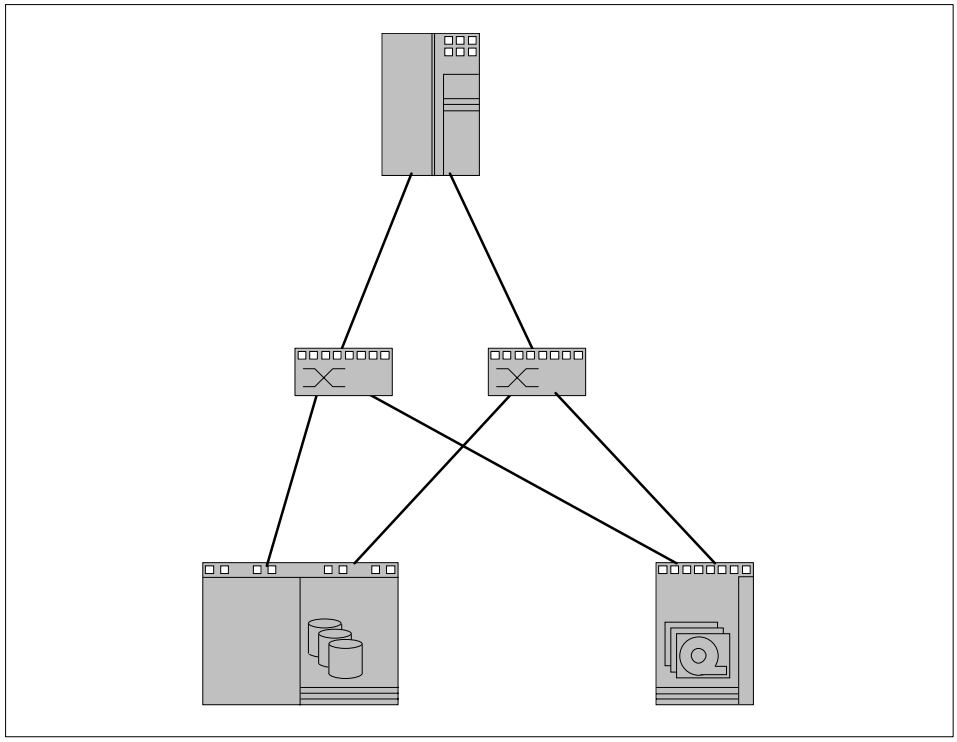


Figure 4-22 Two switch solution with FC-AL support

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 4-23.

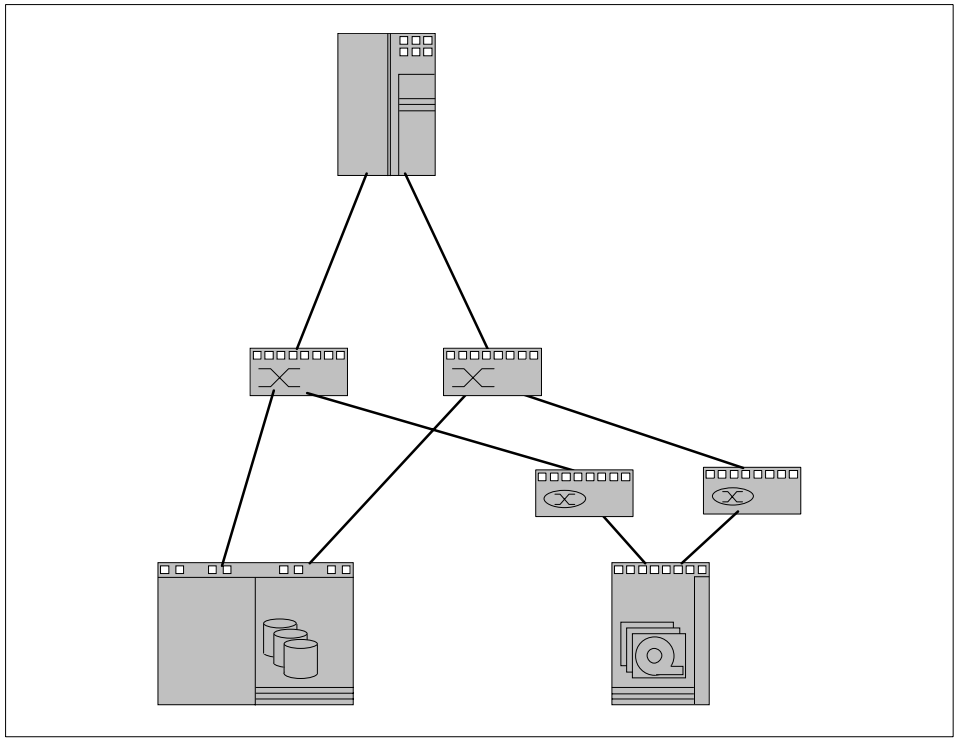


Figure 4-23 Two switch solution without FC-AL support

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 4-24.

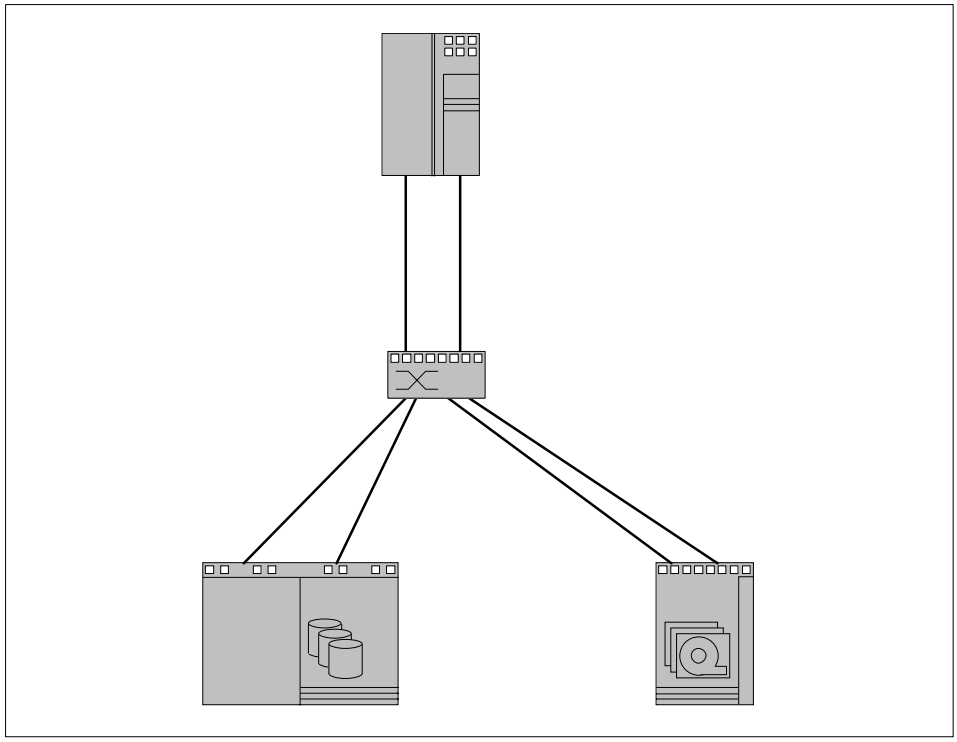


Figure 4-24 Single switch design

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

4.6 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.

4.6.1 Ports and ASICs

When designing the SAN you should also take into account how the SAN ports on the components are internally connected to the ASICs, which basically perform all the routing. You will get the best performance if you put two devices on the same ASIC.

Of course, all switch or director class vendors will claim that the delay in the switch is insignificant to the delay caused by the storage devices and servers. Nevertheless, it is still valuable to detail information as to how the ports are grouped together around the ASIC and use this information to optimize the design even further.

The following represents the information as to how ports are grouped together for each vendor product:

- IBM 3534-1RU: Four ports per ASIC. Ports 0-3 on one ASIC and ports 4-7 on another ASIC.
- IBM 3534-F08: Four ports per ASIC. Ports 0-3 on one ASIC and ports 4-7 on another ASIC.
- IBM 2109-S08: Four ports per ASIC. Ports 0-3 on one ASIC and ports 4-7 on another ASIC.
- IBM 2109-S16: Four ports per ASIC. Ports 0-3 on first ASIC, ports 4-7 on second ASIC, ports 8-11 on third ASIC and ports 12-15 on fourth ASIC.
- IBM 2109-F16: Eight ports per ASIC. Ports 0-7 on one ASIC and ports 8-15 on another ASIC.
- IBM 2109-M12: Four ports per ASIC, four ASICs per blade. Ports 0-3 on first ASIC, ports 4-7 on second ASIC, ports 8-11 on third ASIC and ports 12-15 forth ASIC.
- McDATA ES-3016 16 port switch: Four ports per ASIC. Ports 0-3 on first ASIC, ports 4-7 on second ASIC, ports 8-11 on third ASIC and ports 12-15 forth ASIC.

- McDATA ES-3032 32 port switch: Four ports per ASIC. Ports 0-3 on first ASIC, ports 4-7 on second ASIC, ports 8-11 on third ASIC and ports 12-15 forth ASIC and in the same sequence for the rest of the ports.
- McDATA ED-6064 64 port director: Four ports per ASIC, one ASIC per blade.
- INRANGE FC/9000: Two ports per ASIC, four ASICs per blade. Ports 0 and 1 on first ASIC, ports 2 and 3 on second ASIC, ports 4 and 5 on third ASIC and ports 6 and 7 on the fourth ASIC. The same is applied for all blades.

4.6.2 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.

4.6.3 Domain IDs

The number of available domain IDs in the fabric should also be a consideration point. Different vendors will offer different numbers as to the switches/directors which can be interconnected in the same fabric:

- IBM/Brocade, 31 domain IDs with SilkWorm 1000 series, and 239 domain IDs with SilkWorm 2000 series and newer
- McDATA, 31 domain IDs, but practical limit is 8
- INRANGE:
  - 56 total domains, if only 64-port FC/9000 in fabric
  - 48 total domains, if 128-port FC/9000 in fabric
  - 32 total domains, if 256-port FC/9000 in fabric
  - 16 total domains, if FC/9000-8 or FC/9000-16 in fabric

The number of available domain IDs could limit you in your SAN design.

4.6.4 Zoning

Your SAN design should also be adapted to use zones. In the following sections we will try to outline how the zones behavior will impact your design.

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.

Various vendors offer different implementations of zoning.

IBM TotalStorage SAN Switch zoning

In the topic that follows we show this switch's features and characteristics.

Zone types:

- **Hard zones**

  In a hard zone, all zone members are specified as switch ports; any number of ports in the fabric can be configured to the zone. When a zone member is specified by port number, only the individual device port specified is included in the zone.

Hard zones are position-dependent, that is, a device is identified by the physical port to which it is connected. Switch hardware ensures that there is no data transfer between unauthorized zone members. However, devices can transfer data between ports within the same zone. Consequently, hard zoning provides the greatest security possible. Use it where security must be rigidly enforced.

- **Soft zones**

  In a soft zone, at least one zone member is specified by WWN. A device is included in a zone if either the node WWN or port WWN specified matches an entry in the name server table. 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 zones are returned. Other devices are hidden from the name server query reply.

  When a WWN is specified, all ports on the specified device are included in the zone. Soft zones are name server-dependent and therefore provide more flexibility — new devices can be attached without regard to their physical location. However, the switch does not control data transfer so there is no guarantee against data transfer from unauthorized zone members. Use soft zoning where flexibility is important and security can be ensured by the co-operating hosts.

- **Broadcast zone**

  The broadcast zone controls the delivery of broadcast packets within a fabric. Used in conjunction with IP-capable Host Bus Adapters, a broadcast zone restricts IP broadcast traffic to those elements included in that zone.

  A broadcast zone is hardware enforced by the switch irrespective of the type of element in the zone (that is to say, port or WWN). Broadcast zones are independent of any other zones, hard or soft, in force between the source and destination elements, that is to say, a broadcast is sent to all ports defined in the broadcast zone even though a port is protected by a hard zone.

In Table 4-1, we compare zone types.

Table 4-1 Comparison of 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

Feature	Hard zone	Soft zone	Broadcast zone
Hardware Enforced Data Transfers	Yes	No	Yes
State Change Notification (RSCN)	State changes on any devices within the same zones.		NA
Eligible Devices	All elements must be Physical Fabric Port Numbers	Fabric Port Numbers or World Wide Names. Soft zone automatically specified whenever any element is a WWN.	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.		

Note: Whatever type of zoning you are using, ISLs are transparent to all zones. This means that you cannot separate ISL traffic across different ISLs, but all the paths will be automatically allocated by FSPF. To dedicate a path, static routing could be employed.

McDATA zoning

McDATA implements soft zoning, which is also called name server zoning. It is done by authorizing or restricting access to name server information associated with device ports. These are some important characteristics of zoning:

- Each device port that belongs to a zone is called a member of the zone.
- 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.

There is a maximum a 4096 zone members, but the exact number of the zone members that can be defined is bounded by the available nonvolatile random access memory (NVRAM) in the director 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 reconfigured with the new WWN.

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 getting access to devices that you did not intend them to have access to, and losing access to the correct devices.

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 and up to 64 zone sets can be defined.

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.

It is always wise to be careful when activating zone sets as any one of the following 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 the default zone is disabled then no device can communicate.
- Activating a new zone set while one is active, the new set will replace the currently active.
- Deactivating the currently active zone set will make all devices members of the default zone.

- 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.
- Zone sets are saved in a zone set library. Any zone set in the zone set library can be selected and made the active zone set.

> **Note:** When you use zoning with McDATA switches/directors, ISLs are transparent to all zones. This means that you cannot separate ISL traffic across different ISLs, but all the paths will be automatically allocated by FSPF.

INRANGE zoning

There are three types of zones that the INRANGE Director supports and may be defined:

- Hard zoning
- Name server zoning
- Broadcast zoning

Hard zoning

Hard zoning follows physical boundaries within a single-stage switch chassis, and limits the communication of a port to only other ports in the same hard zone. Hard zoning, in certain circumstances, is the only way to provide the required additional level of security, but careful consideration should be applied prior to activating any hard zones, as it may be possible to isolate devices.

Hard zoning rules

The purpose of hard zoning is to guarantee that there is no possible way for information in one zone to be confused with information in another zone. Allowing multiple hard zones to overlap in the same port, even a T_Port, would violate this level of security. For a lower level of security where port overlapping is required, software zoning should be implemented.

If a lower level of security is desired, then soft zoning is recommended where overlapping is allowed, so T_Ports can exists in multiple zones.

There are a number of rules that have to be followed to allow you to implement hard zoning successfully:

- You can define a maximum of 16 hard zones in an INRANGE fabric.
- When a hard zone is created it must be in a granularity of four ports.
- Ports in the same zone must be adjoining.
- When a hard zone is defined, all used and unused ports must be assigned to one hard zone or another.

- Hard zoning is supported for multiple INRANGE Directors connected in a fabric. All ports used and unused within that fabric must be assigned to one hard zone or another. The T_Port connection used to connect the INRANGE Directors together must be the same physical port number on both Directors.
- Any update to hard zone settings will cause all ports to perform a fabric login. Hard zone changes should be restricted to initial setup and at maintenance slots.

Note: In the case of Hard zoning with INRANGE directors, ISLs are included in the zone. This means that you *can* use hard zoning to separate the traffic.

Name server zoning

Name server zoning, or software zones, allow the division of the fabric (one or more switch chassis) into 16 zones. With the latest level of firmware this will increase to 256 possible zones.

INRANGE implement soft zoning by using the World Wide Port Name (WWPN) address.

The fabric-wide zones (across one or many INRANGE Directors) define which ports receive name server information. A particular WWPN may be defined in one or more of these name server zones.

If a WWPN has been defined in a hard zone the same WWPN cannot be used in a name server zone.

The name server zones WWPN are selected from the Fabric Topology window and apply to the entire fabric. If a WWPN is defined to a zone, a port will receive name server information for all ports in the same name server zone (or zones) in which the port is defined.

All ports not defined as being part of any enabled name server zone are name server zone orphans. Name server zone orphans are all placed in the same name server orphan zone.

Note: When you use soft zoning with INRANGE directors, ISLs are transparent to all zones. This means that you cannot separate ISL traffic across different ISL, but all the paths will be automatically allocated by FSPF.

Broadcast zoning
You can currently have up to 16 zones and, like name server zoning, this will increase to 256 with the next version of firmware. Broadcast zones allow the division of the fabric into an area of broadcasts. A particular port may be placed in one or more of these broadcast zones. A port will broadcast to all ports in the same broadcast zone (or zones) in which the port is defined. If hard zones are enabled, broadcast zones may not cross the defined hard zone boundaries.

Broadcast zones are typically used for transferring TCP/IP addresses for clustered server configurations.

Because in general, ISLs are used across all zones (except in the case of INRANGE hard zoning) the number of ISLs should not increase when you plan to use zones in your design.

4.6.5 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.

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 are 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 2.12.9 “Buffers and credits” on page 114.

4.7 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 what were 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 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 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.

4.7.1 Certification and support

As we already mentioned in the 4.7 “Interoperability issues in the design” on page 271, 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.

4.7.2 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

4.8 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.

4.9 Management

In this section, we will discuss aspects of how to manage the SAN.

4.9.1 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 4-25.

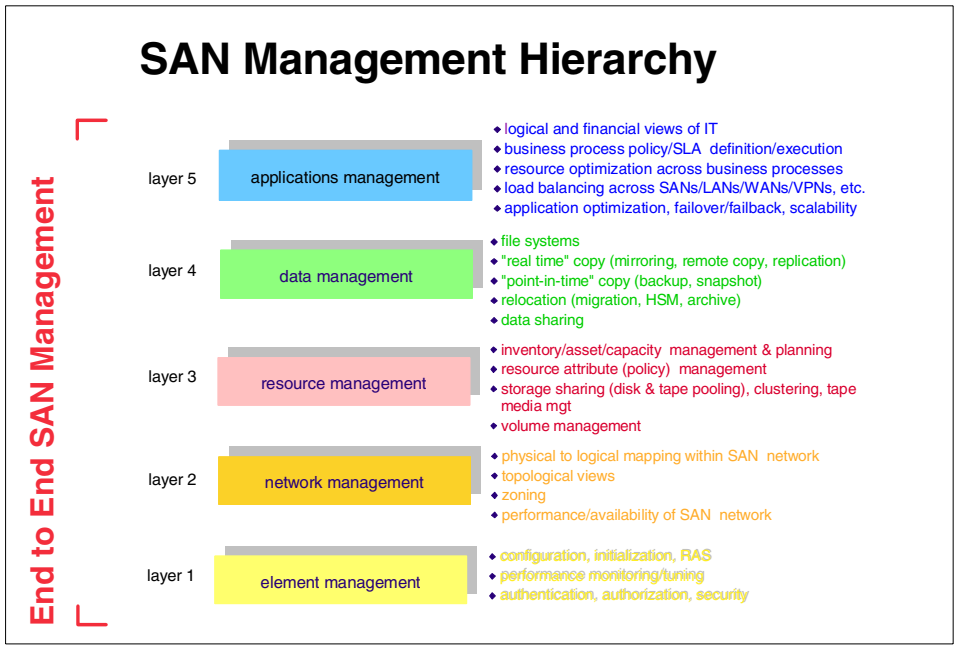


Figure 4-25 SAN management hierarchy

These can be implemented separately, or potentially as a fully integrated solution to present a single interface to manage all SAN resources.

4.9.2 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) sponsored by Microsoft, or application programming interfaces (API), such as those proposed by the Jiro initiative, sponsored by Sun Microsystems, and others by SNIA and SNMWG.

Figure 4-26 illustrates a common interface model for heterogeneous, multi-vendor SAN management.

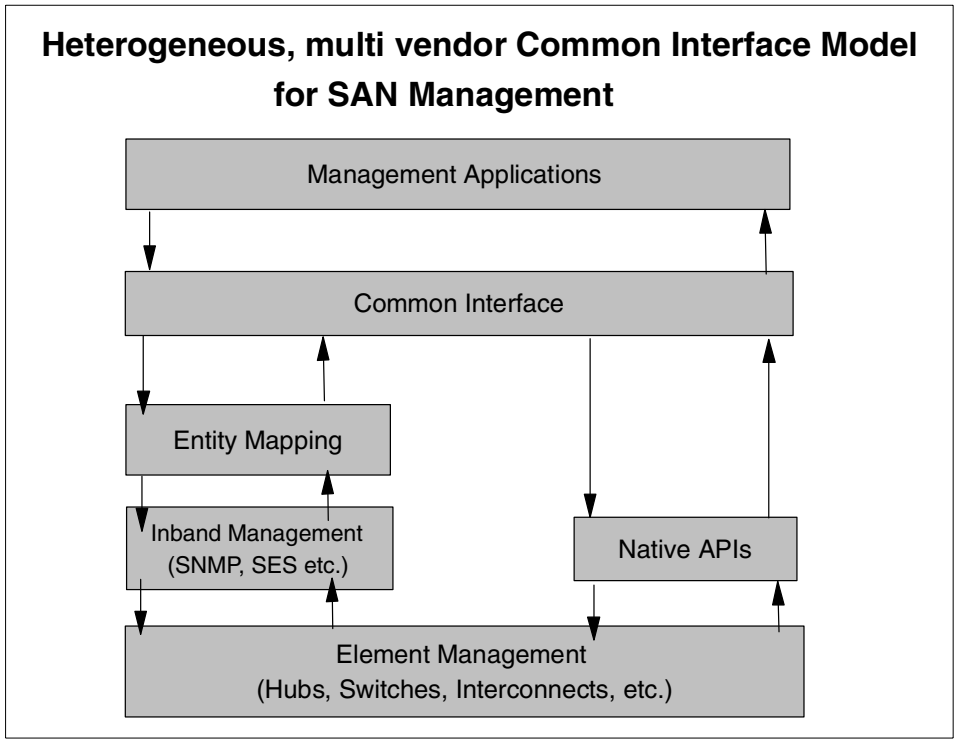


Figure 4-26 Common interface model for SAN management

4.9.3 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.

4.9.4 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, and this makes the manual consolidation of resource administration information impractical, 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

4.9.5 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, as shown in Figure 4-27.

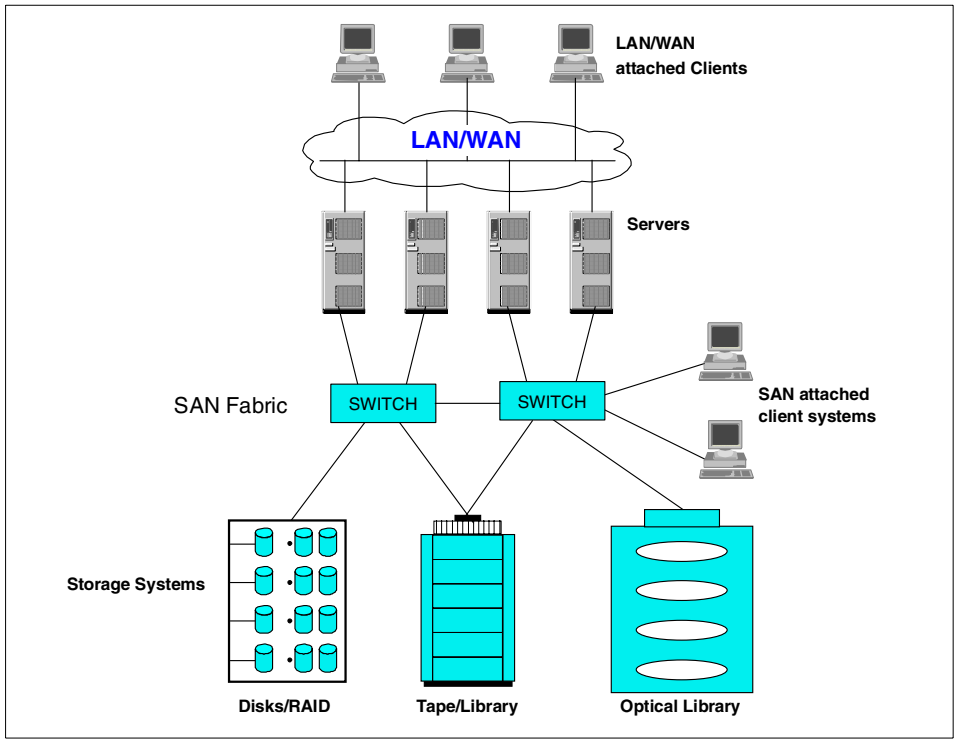


Figure 4-27 Typical SAN environment

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. This is especially in terms of physical connectivity mapping, fabric zoning, performance monitoring, error monitoring, and predictive capacity planning.

4.9.6 Element 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 individual elements, usually comprising software, firmware and hardware elements such as those shown in Figure 4-28.

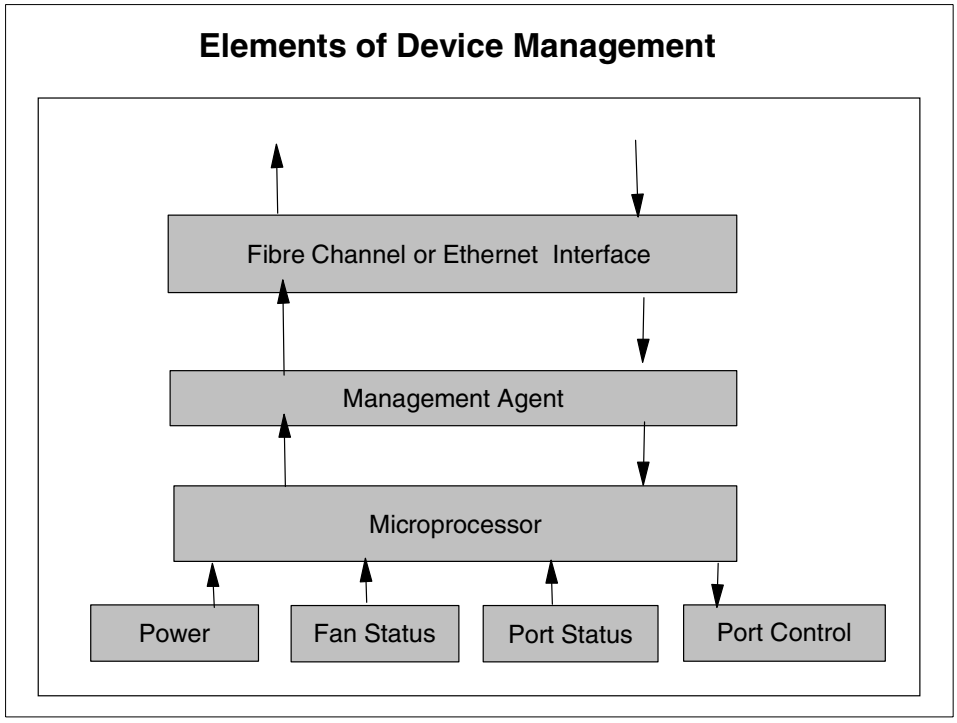


Figure 4-28 Device management elements

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.

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.

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 don’t 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.

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 SNIA. When these are defined and adopted, multi-vendor SANs can be managed by common commands and queries.

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.

4.9.7 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.

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 inband and outband in 2.14.4 “Management” on page 130.

There are several access methods for managing a switch or director.

In Table 4-2 we summarize the management access methods available.

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.

Table 4-2 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 outband Ethernet port, and one for inband Fibre Channel network access.

Part 2

Case studies and solutions

In this part of the book, we introduce our real world case studies and the solutions that we have designed based upon our experience and the concepts that we have discussed in the first part of this book.

Once we have shown our designs, we show some best practices that should be employed to optimize your SAN infrastructure.

Remember, a SAN is not just a collection of fabric components.

5

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, McDATA, and INRANGE. 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.

5.1 Case Study 1: Company One

We will detail the background and requirements of the company for this case study.

5.1.1 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.

5.1.2 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.

5.1.3 Current infrastructure

The existing network infrastructure will be used which is 100BaseT switched Ethernet.

5.1.4 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 2000 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 4.0 SP6, but will migrate to Microsoft Windows 2000 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 IBM RS/6000).
- 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.

5.1.5 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 4, “SAN design considerations” on page 221 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 2000 Server operating system.
- Two servers running Microsoft Windows NT 4.0 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.

5.2 Case Study 2: Company Two

We will detail the background and requirements of the company for this case study.

5.2.1 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.

5.2.2 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

5.2.3 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 IBM RS/6000 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 6.5.1) running independently but using two way database mirroring with Informix 7.2 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 5GB 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 5-1 we show the inventory.

Table 5-1 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 D Class	HP/UX 10.01	70GB	4KB	4000	8	Getwell
HP 9000 D Class	HP/UX 10.01	300GB	4KB	450	3	Feelinbad
SGI Origin 2000	IRIX 6.5.12	400GB	32KB	10,000	2	Getwell
IBM RS/6000 H50	AIX 4.3.1	300GB	8KB	2000	10	Getwell
DEC Alpha 4100	OpenVMS 7.1	40GB	2KB - 4KB	400	6	Getwell
DEC Alpha 3000	Tru64 V.4.0	30GB	2KB - 4KB	600	4	Getwell
Compaq Proliant 6500	MS/NT SP5	60GB	2KB - 8KB	3800	10	Getwell
Compaq Proliant 6500	Netware 5.1	60GB	4K B - 8KB	3200	10	Getwell
Compaq Proliant 6500	Netware 5.1	120GB	4KB - 8KB	990	3	Feelinbad

Hardware	OS	Total Storage for all	Block Size Range	Total I/O/S for all	Quantity	Location
Compaq Proliant 6500	Netware 4.11	40GB	4KB - 8KB	660	6	Getwell
Compaq Proliant 6500	Netware 4.11	80GB	4KB - 8KB	400	2	Feelinbad

The average read/write ratio across the whole server complex is 75:25.

In Figure 5-1 we show the server schematic for Case Study 2.

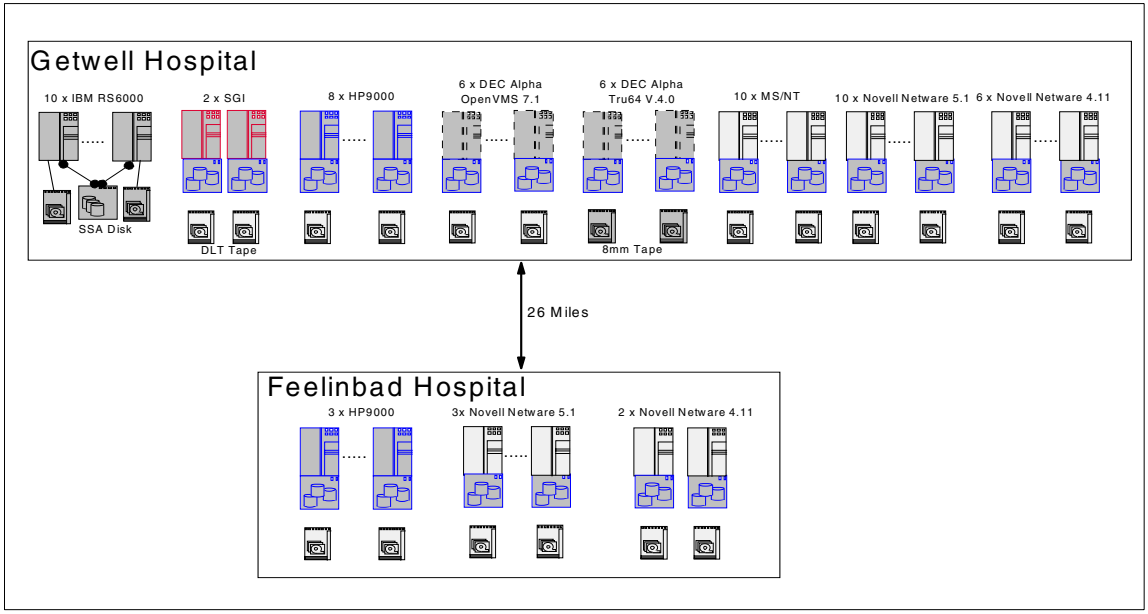


Figure 5-1 Case Study 2: Server Schematic

5.2.4 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).

5.2.5 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 4, “SAN design considerations” on page 221, 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 500GB. 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; refer to 2.7.6, “100 MB/s” on page 84. 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 RS/6000. 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 5-2, we show a schematic representation of dividing ISLs for data access and for data replication.

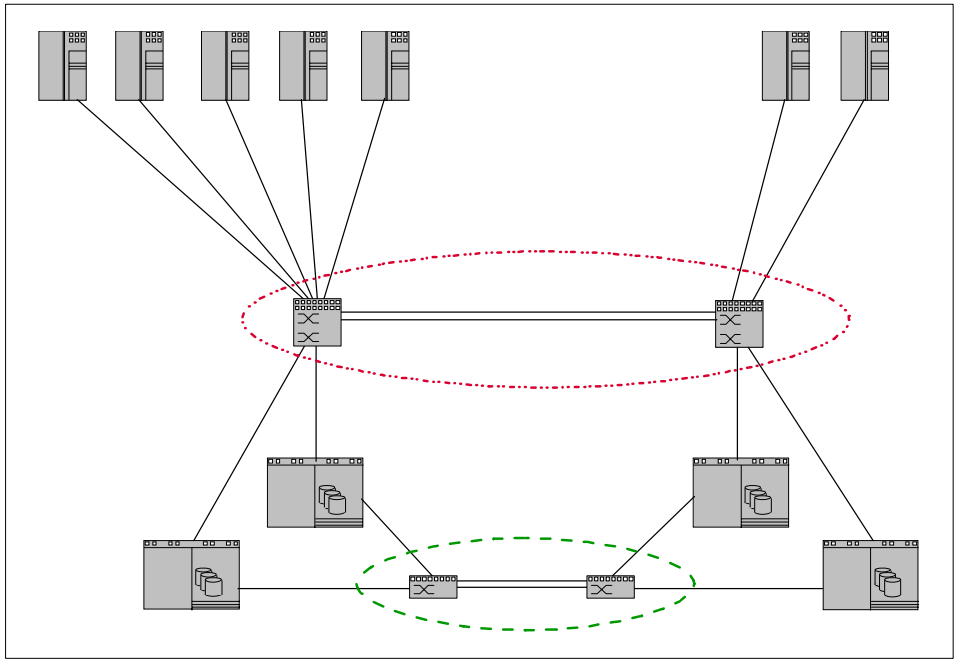


Figure 5-2 Different ISL for data access and for data replication

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 ESCON links for the data replication between two storage devices. If the distances are too long for standard ESCON links we could use ESCON directors or 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 ESCON. We will use an adequate number of ESCON links. Because the sites are 26 miles apart we will use ESCON 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.

5.3 Case Study 3: Company Three

We will detail the background and requirements of the company for this case study.

5.3.1 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.

5.3.2 High-level business requirement(s)

- Implement a consolidated storage solution to satisfy growth of NT platform which is nearing capacity.

5.3.3 Current infrastructure

Seven servers: 5 Microsoft Windows NT Server (Dell), 1 RS/6000 (R40) and an AS/400 (720). Windows NT environment performs Microsoft Exchange and file/print server functions and must be available during office hours (Monday through Friday, 8am - 5pm). The RS/6000 needs to process data that is created on the AS/400 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 AS/400 runs a proprietary core business application and needs to be available 24x7 for processing (batch and online). The cost to the business of the AS/400 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, RS/6000 has 8GB, and AS/400 has 36GB. In Figure 5-3 we show the server schematic.

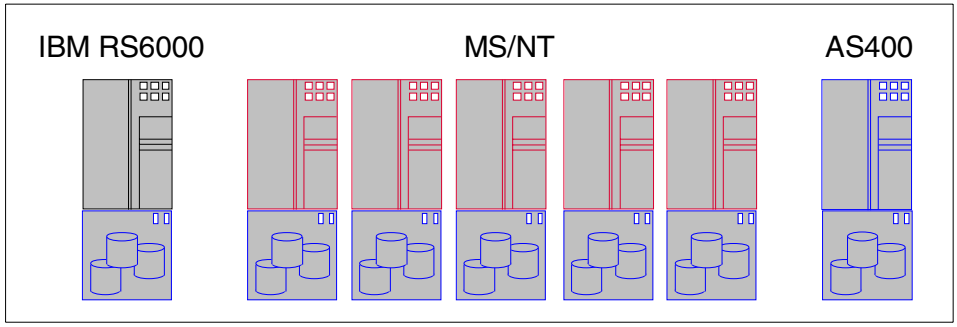


Figure 5-3 Case Study 3: Server Schematic

5.3.4 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.

5.3.5 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 4, “SAN design considerations” on page 221 for explanation of how various factors can affect your SAN design.

In our example we have the following requirements:

- Five Windows NT servers with 57GB of storage, one RS/6000 server with 8GB of data and AS/400 with 36GB 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.
- At the time of writing this, AS/400 supported only FC-AL connections to the SAN and only one target for connection which has to be in the FC-AL mode. If OS/400 is not on the right level (5.0 or up) it will not support Fibre Channel connections, thus leaving the AS/400 out of the SAN design. In our example we assume that we have the correct level of the OS/400 which is 5.1. Because of the target which has to be in the loop mode we will dedicate two separate storage ports for AS/400 storage. Those ports will be configured to operate in FC-AL mode.
- There is a need to transfer 20 MB of data every night from the RS/6000 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.
- The ports for AS/400 have to be able to support FC-AL connections. Also for these ports we have to have the option to group one server port with and one storage port, and isolate them from all other ports.
- 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.

5.4 Case Study 4: Company Four

We will detail the background and requirements of the company for this case study.

5.4.1 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 OS/390, 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.

5.4.2 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.

5.4.3 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 S/390 G5 4-way 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 RS/6000 S80 servers (production), 2x RS/6000 H70 (test/QA) servers (all running AIX 4.3.2), 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 RS/6000's use IBM's HACMP, and the Sun servers rely on Veritas for this function.

The production RS/6000 servers run DB2 UDB 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 RS/6000 servers have a similar configuration to that of production (to enable full *functional* testing of product) but only use 200GB 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 (8KB 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 OS/390 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 OS/390: The same OS/390 system is ESCON attached via 8 paths to an IBM 3494 tape library with a VTS-B20 (Virtual Tape Server) with sixteen 3590E 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 5-4 we show the case study schematic.

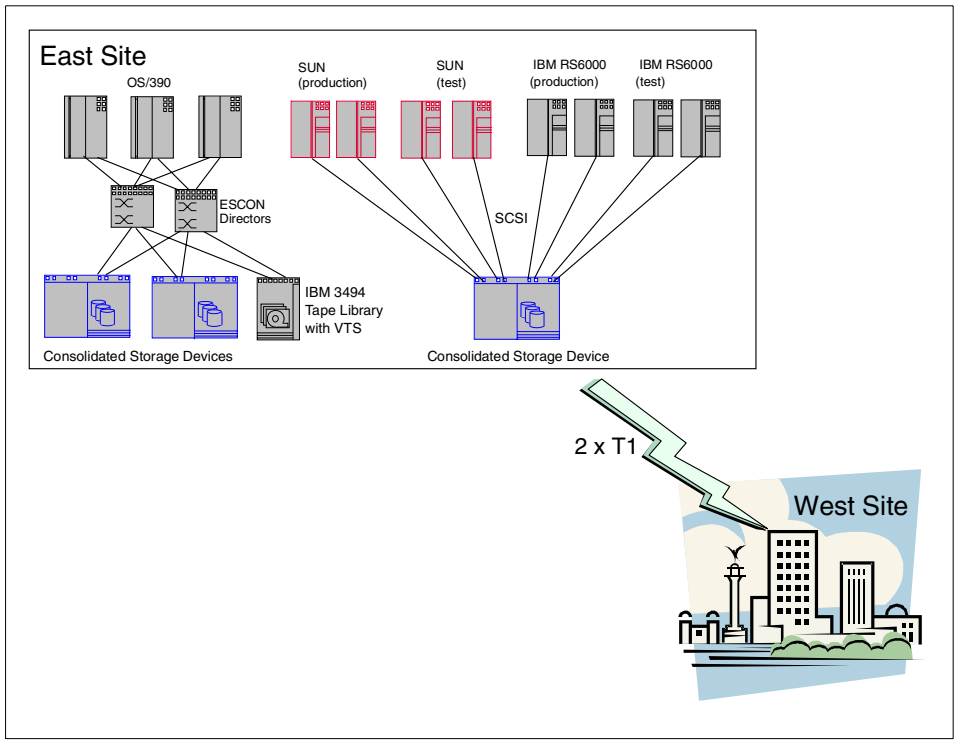


Figure 5-4 Case Study 4: Server schematic

5.4.4 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 3 years.

  As Company Four must replace the existing equipment, they would like to take advantage of Fibre Channel 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 500 miles (804.7 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.

5.4.5 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 4, “SAN design considerations” on page 221, 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 S/390 storage for two S/390 servers in the east location. The I/O peak throughput is 218 MB/s
- Two RS/6000 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 RS/6000 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 1 Gb/s SAN port per server. But because we have to implement the redundant fabric without single points of failure we will use two 1 Gb/s SAN ports per each server. This will give us 4 x 2 = 8 SAN ports for the open systems servers in the east location. To accommodate the storage device bandwidth we will use two 1 Gb/s SAN ports for open systems servers.

For the west servers we will have 2 x 2 = 4 1 Gb/s SAN ports for open systems servers and two 1 Gb/s 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 user synchronous 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 TSM server on the S/390. The backup will be performed over dedicated network backbone and only at one site.

For S/390, we will use FICON connectivity for storage devices. We will have three FICON channels per server (recommended ESCON to FICON ratio is 4: 1). We will use the same IBM ESS as for the open systems servers. For those six FICON connections we will have six FICON connections on IBM ESS.

For the disaster recovery site we will implement XRC for S/390 systems data replication from east to west site, using FICON over ATM channel extenders. For the XRC implementation we will add an additional four FICON adapters to the local ESS. On the remote site we will have an additional four FICON 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 FICON connections for XRC will be attached directly from the IBM ESS at the east site to a S/390 at the recovery site (west).

We will keep the ESCON connections for the S/390 tape backup.

This gives us a total, in the east site, of:

- Ten SAN ports for open system servers and storage
- Twelve FICON SAN ports for S/390 systems and their storage

For the west site we have total of:

- Six SAN ports for open system servers and storage
- Twelve FICON SAN ports for S/390 systems and their storage

We will accommodate six ESCON channels for the open systems' PPRC.

5.5 Case Study 5: Company Five

We will detail the background and requirements of the company for this case study.

5.5.1 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.

5.5.2 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.

5.5.3 Current infrastructure

Server environment consists of 3 Dell servers. Two of the servers are running Microsoft Windows NT, the third is running Redhat Linux 7.1. 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 5-5 we show the server schematic.

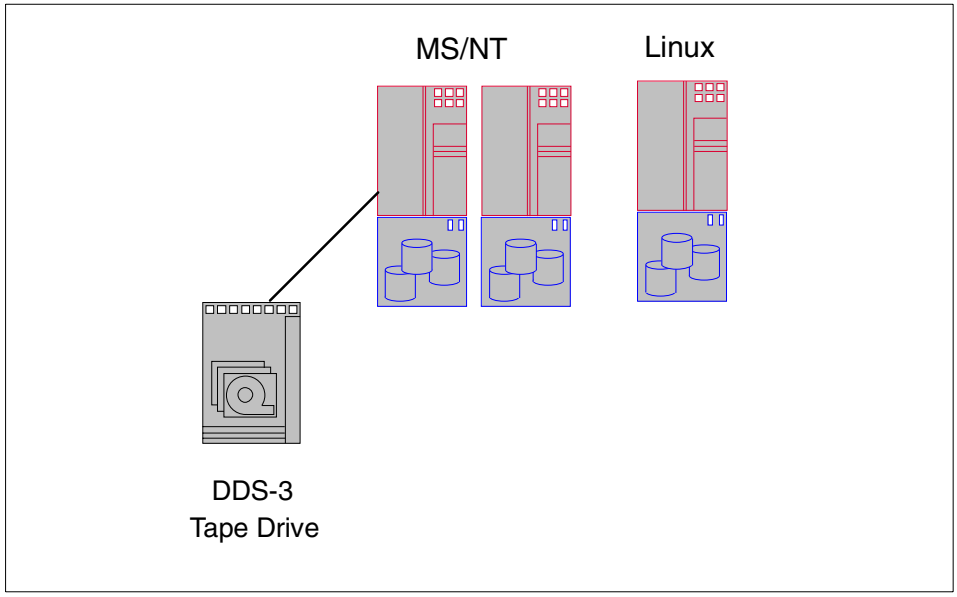


Figure 5-5 Case Study 5: Server Schematic

5.5.4 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.

5.5.5 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 4, “SAN design considerations” on page 221, 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 1 Gb/s SAN port. Because we want to build redundant connections from servers to the SAN fabric, we will use two 1 Gb/s SAN ports per server.

Bases on the bandwidth and storage requirements, we will introduce a storage device with two 1 Gb/s 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 1 Gb/s SAN ports for the backup server and two FC-AL ports for Fibre Channel tape drives in an IBM 3584 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 1 Gb/s SAN ports for all servers
- Two 1 Gb/s SAN ports for storage
- Two 1 Gb/s FC-AL SAN ports for tape drives

5.6 Case Study 6: Company Six

We will detail the background and requirements of the company for this case study.

5.6.1 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.

5.6.2 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

5.6.3 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 p640. 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 1 Gb/s switched ethernet backbone.

The read/write ratio is 70:30 across the applications.

A direct SCSI connect DLT library with two DLT7000 drives daisy chained is available for use between the two servers for backup purposes. It requires a manual process to connect the library to each server when backups need to be taken.

In Figure 5-6 we show the server schematic for Case Study 6.

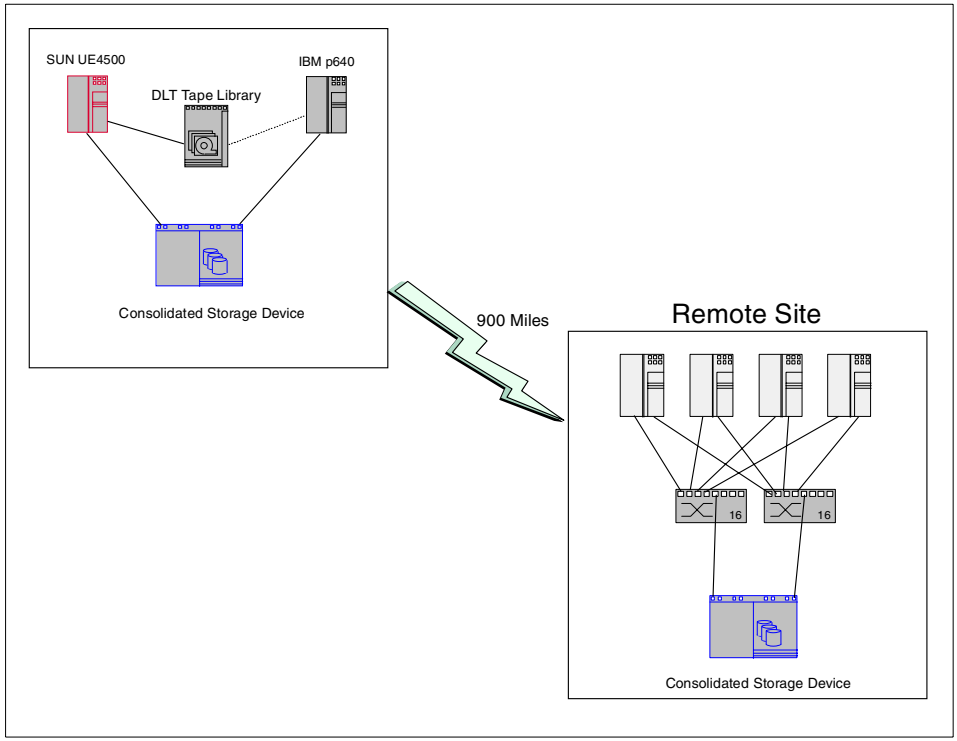


Figure 5-6 Case Study 6: Server Schematic

5.6.4 Detailed requirements

- To increase data availability and accommodate the storage requirements for another UNIX server.

  Company Six recently experienced a GBIC 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 40GB 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 200GB 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 with no manual intervention.

  With the proposed introduction of another server, Company Six would like to avoid the current manual process of connecting the DLT library. They would like to continue using the same library.

5.6.5 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 4, “SAN design considerations” on page 221, 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 46.88MB/s. Currently, this server is only using one Fibre Channel adapter, which is directly connected to the storage device.
- One IBM p640 server with 800MB 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 800KB/s.
- Because of recent failure, we will design a redundant SAN with two ports in each server. This will give us six 1 Gb/s 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 1 Gb/s 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 RS/6000. Existing servers will use LAN-Free clients to backup their data through the SAN. The existing tape library will be connected to the SAN using a SAN Data Gateway Router. The existing DLT tape library will be verified for support in this environment.

To summarize we have the following SAN ports requirements at the primary site:

- Six 1 Gb/s SAN ports for servers
- Two 1 Gb/s SAN ports for the storage device
- Two 1 Gb/s SAN port for the tape device
- Two ISLs for data replication

At the secondary site we have the following requirements:

- Six 1 Gb/s SAN ports for UNIX servers
- Eight 1 Gb/s SAN ports for Windows NT servers
- Two ports for storage device
- Two ports for ISLs for data replication

6

IBM TotalStorage SAN Switch Solutions

In this chapter we will describe solutions that are based upon the IBM TotalStorage SAN Switch range of products.

6.1 Case Study 1: Company One

If we consider the company and its requirements, as detailed in 5.1 “Case Study 1: Company One” on page 288, we will propose two designs. In one design, we will use director class products and in the other we will use switches.

6.1.1 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 6-1.

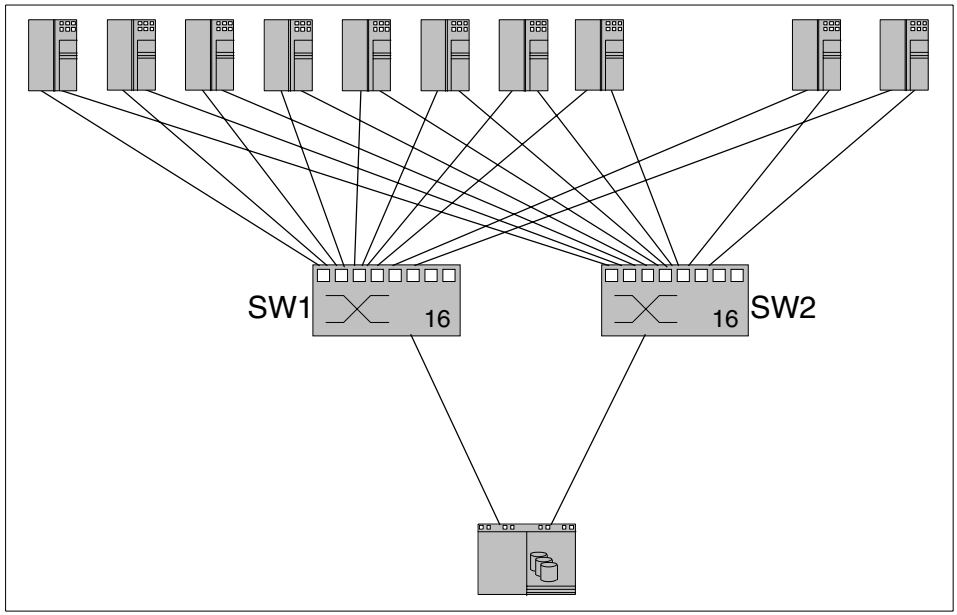


Figure 6-1 Core SAN design

In Table 6-1 we show the number of ports used for the design.

Table 6-1 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 6-2, you can see that we have added an additional two switches to the SAN.

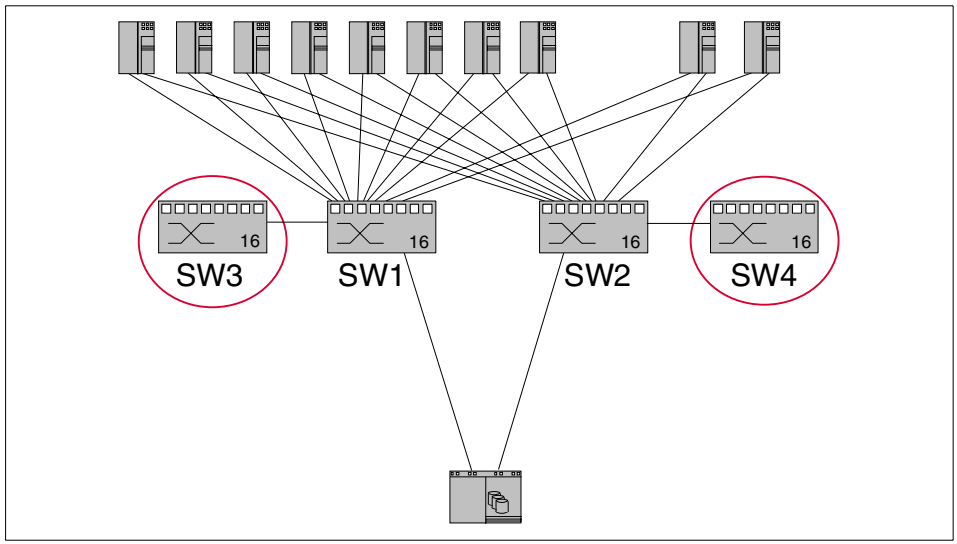


Figure 6-2 Adding additional two switches to the SAN

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 6-3, we show how new servers in the second year can be connected to the SAN.

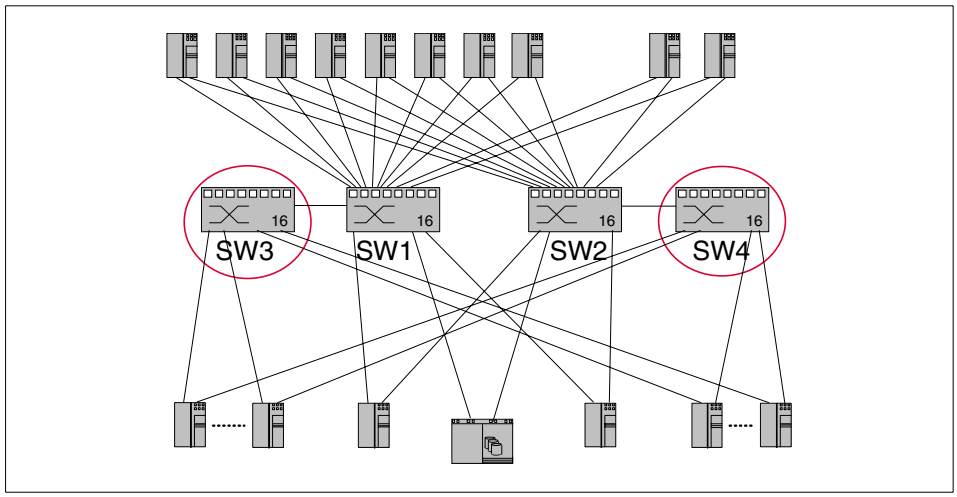


Figure 6-3 SAN after first year of expansion

In Table 6-2 we show the port usage.

Table 6-2 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 6-4.

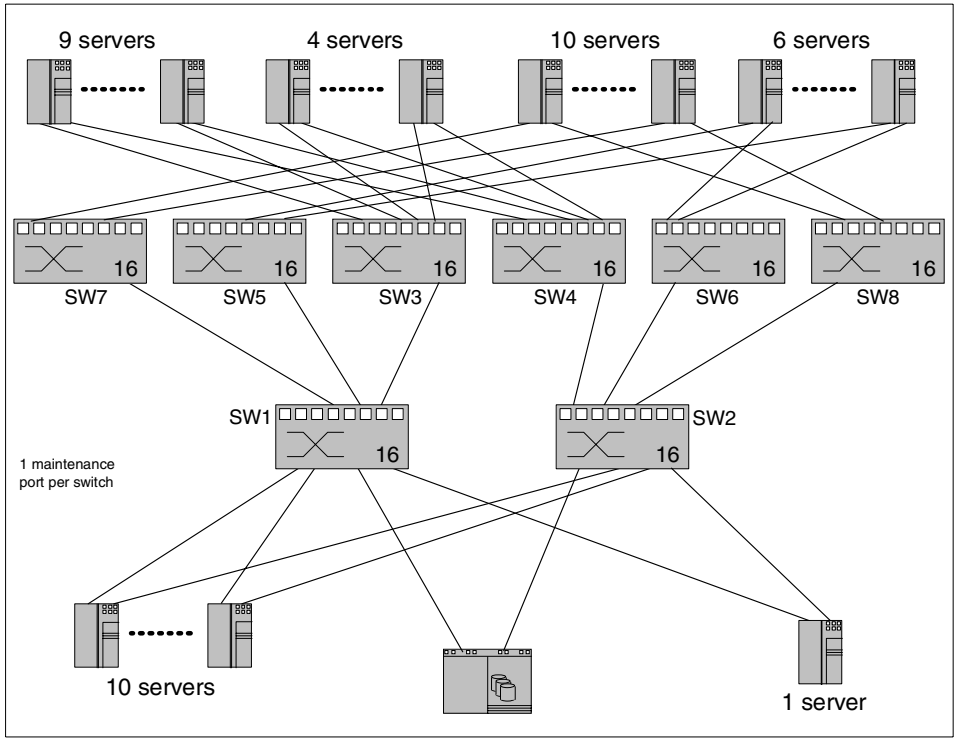


Figure 6-4 SAN design after three years of operation

In Table 6-3, we show the number of ports used for the solution.

Table 6-3 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 6-4 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 6-5, and this rearrangement will not cause any further expense for the purchase of new equipment.

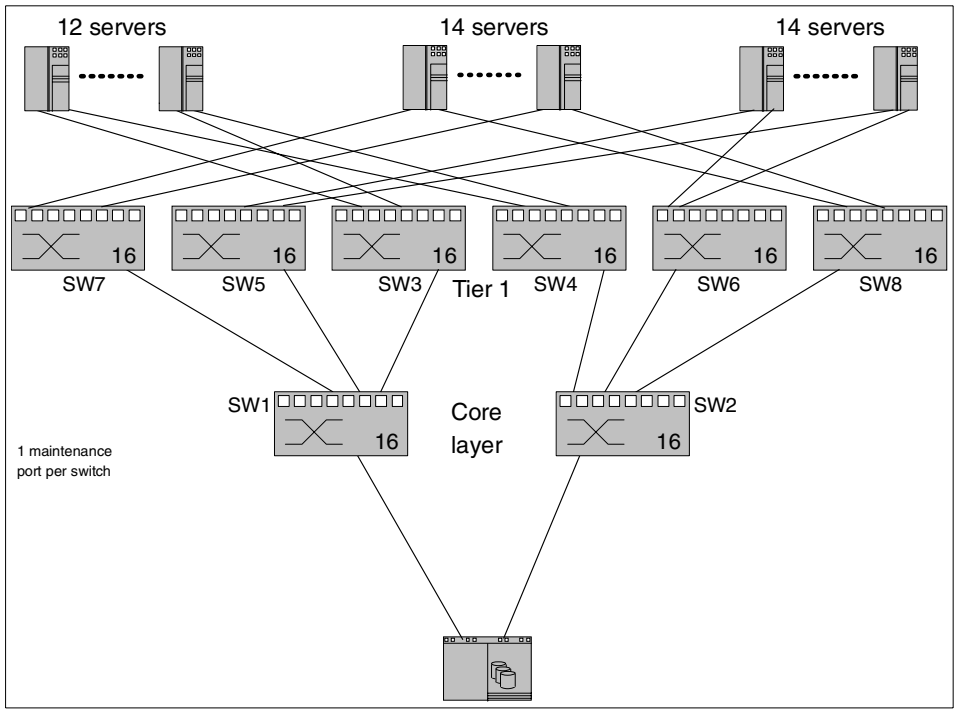


Figure 6-5 Core edge design

In Table 6-4 we show the number of ports used for the solution.

Table 6-4 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 6-5.

Note: All the Fibre Channel switches used in this design can be either 1 Gb/s or 2 Gb/s capable. However bandwidth requirements dictate that 1 Gb/s should be sufficient.

In the following sections, we will outline some aspects of the design.

6.1.2 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 8.32MB/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 16.64MB/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 fibre cable is 5 microseconds.

In our example, the whole implementation can be done using 1 Gb/s technology.

6.1.3 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.

6.1.4 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 NT/2000). 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.

6.1.5 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 GBICs as the servers are within a radius of 500 m. It is possible to move some servers further away by using longwave GBICs 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.

6.1.6 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 6-5 on page 326.

This will give us ample opportunity to grow as we can attach additional pairs of switches to core switches SW1 and SW2.

6.1.7 “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 GBIC, 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.

6.1.8 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 6-6.

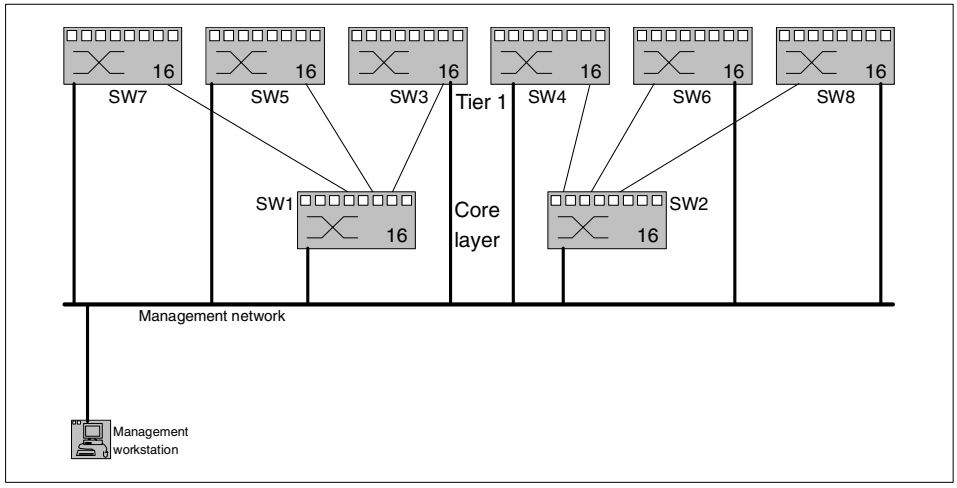


Figure 6-6 Management network

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.

6.1.9 Core switch design

For the core switch design, we decided to use two 24 port core switches. You can see the design in Figure 6-7.

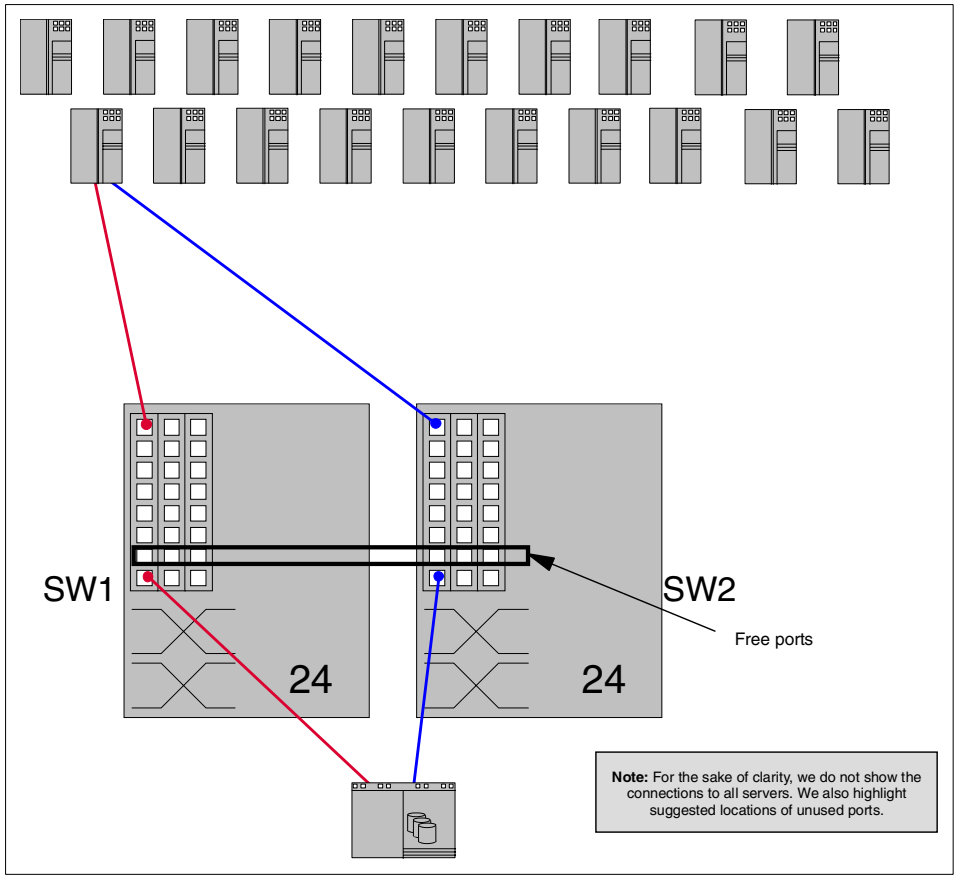


Figure 6-7 Core switch design

In Table 6-5, we show the number of ports used for the design.

Table 6-5 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 6-7 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 6-8.

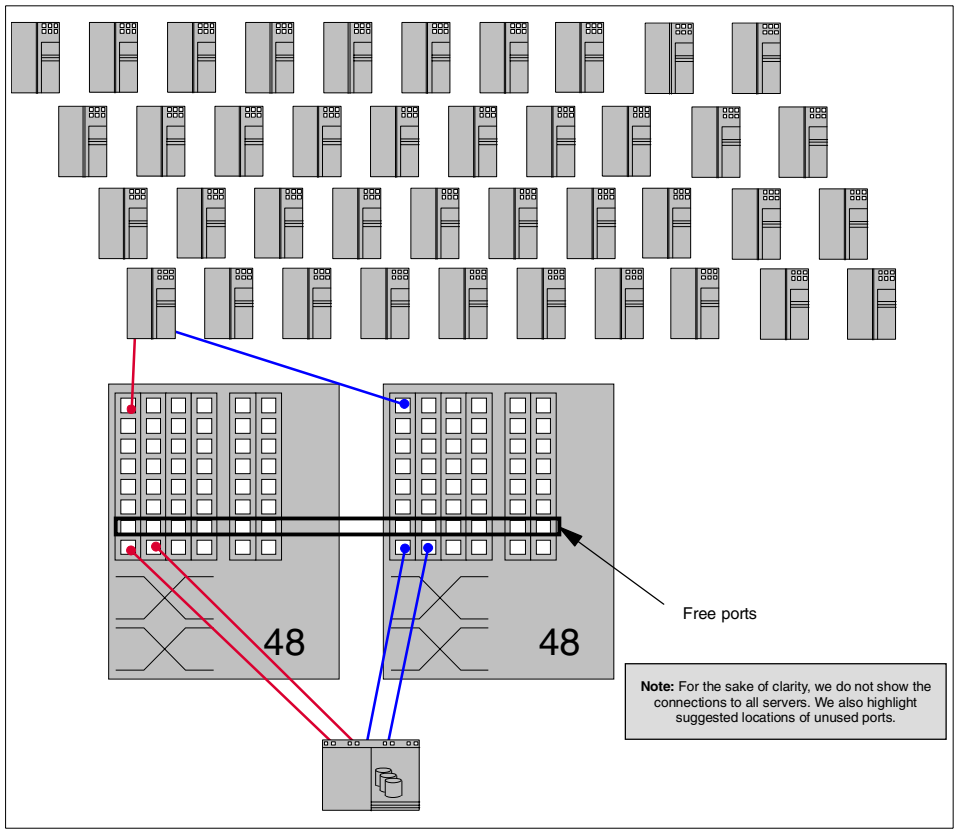


Figure 6-8 Expanding to 40 servers using core switch technology

In Table 6-6 we show the number of ports used for the design.

Table 6-6 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 6-8, we simply added additional port blades inside the existing core switches. No need to allocate ISLs as in the switch design.

6.2 Case Study 2: Company Two

If we consider the company and its requirements as detailed in 5.2 “Case Study 2: Company Two” on page 291, we will propose the following designs.

6.2.1 Design

Considering the requirements as detailed in “Analysis (ports and throughput)” on page 294, we will propose two designs: In one design we use director class products, and in the other we use switches. Because we use 2 Gb/s HBAs in all servers we will use 2 Gb/s components in both designs.

Switch design

In Figure 6-9 we show the initial design for the Getwell site.

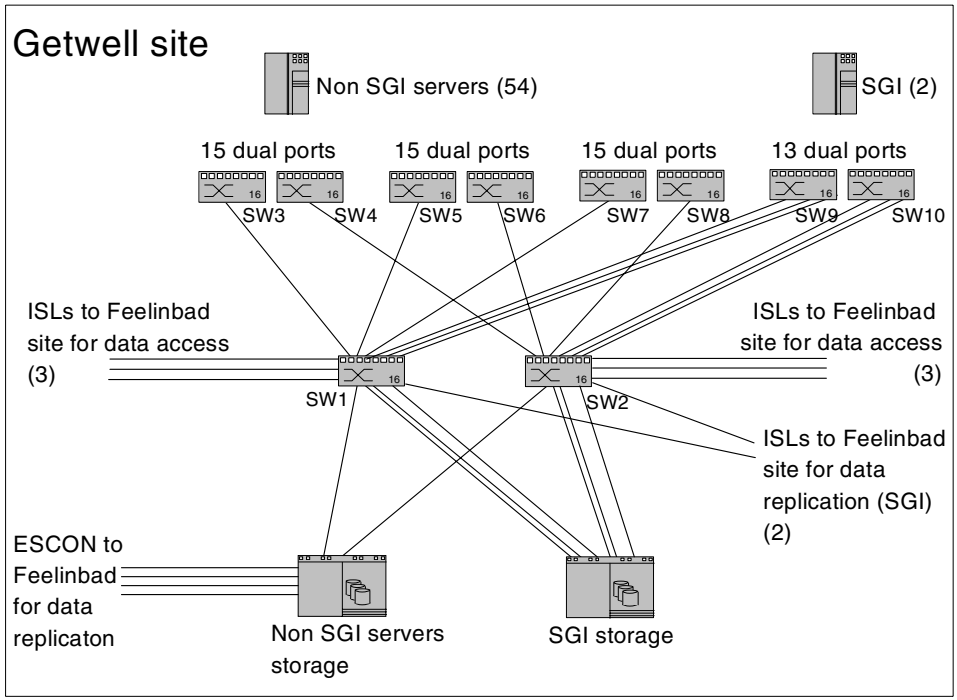


Figure 6-9 Getwell site initial design

In Table 6-7 we show the number of ports used for our solution.

Table 6-7 Number of used ports

Ports	Servers	Storage	ISLs	Spare
SW1	0	4	10	2
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
- ESCON 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 6-10.

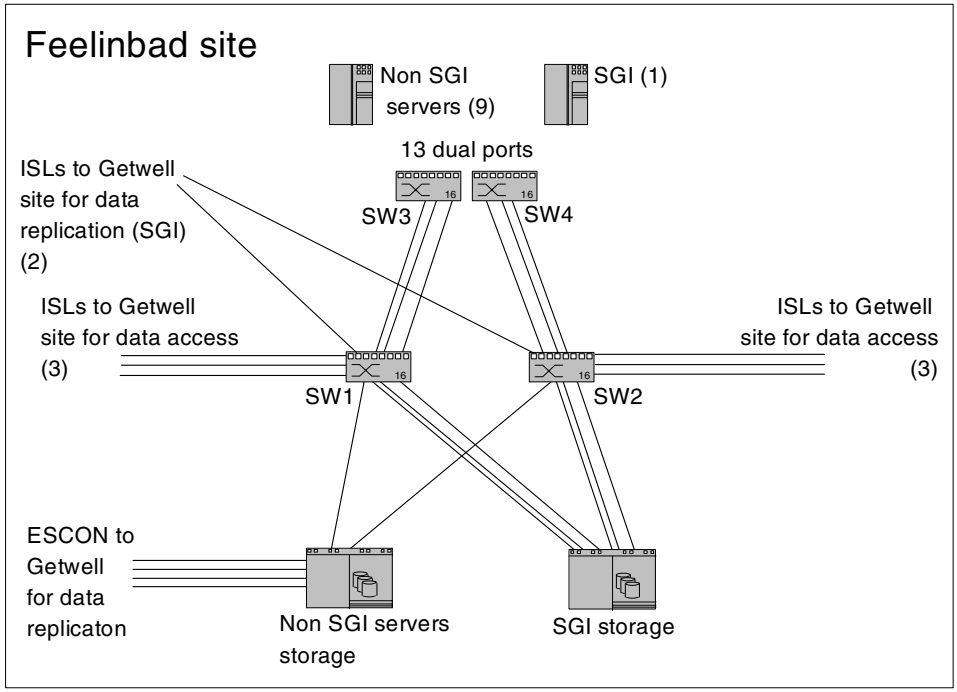


Figure 6-10 Feelinbad site initial design

In Table 6-8 we show the number of ports used for our solution.

Table 6-8 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 ESCON 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.

6.2.2 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 6-11.

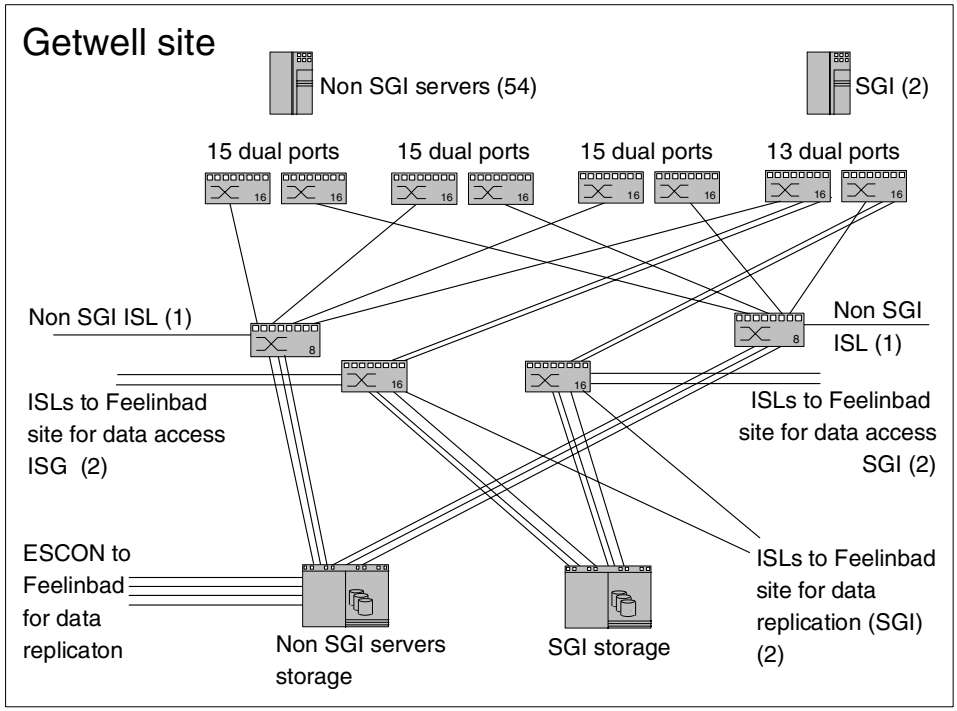


Figure 6-11 Adding additional storage ports for non-SGI servers

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.

In our example the whole implementation can be done using 2 Gb/s technology.

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. We explain this behavior in 2.7 “Fabric Shortest Path First” on page 78.

In our example we have decided to use trunking on all multiple ISLs. In Figure 6-12 we show where we applied trunking.

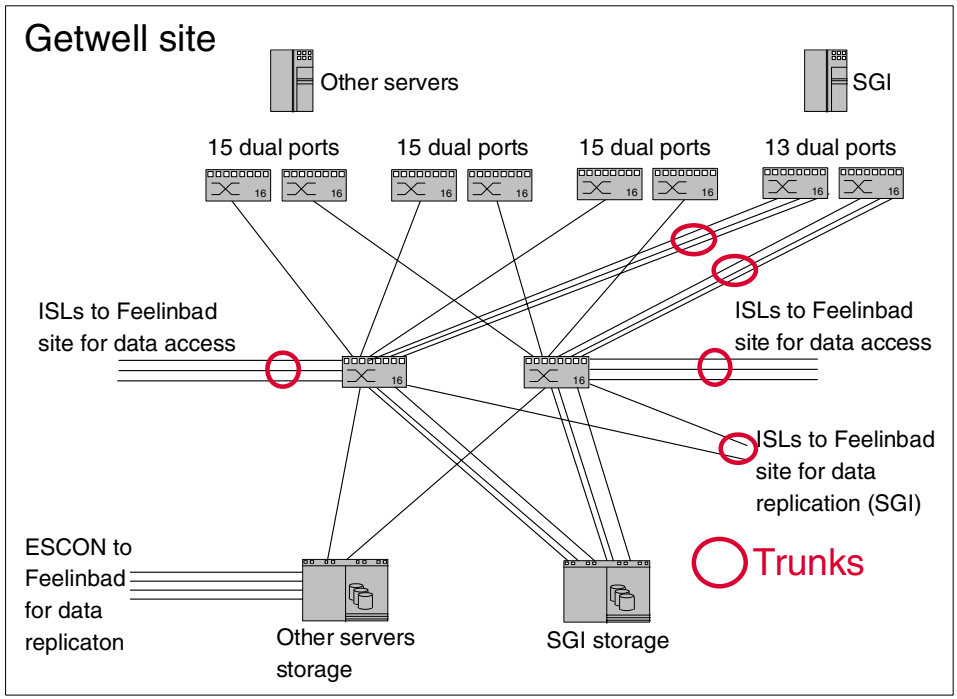


Figure 6-12 Trunking in Getwell Center

Trunking will give us better utilization of the ISLs.

6.2.3 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.

6.2.4 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.

6.2.5 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 GBICs as the servers are in a radius of 500 m. It is possible to move some servers further away by using longwave GBICs 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 GBICs 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.

We discuss this in greater detail in 2.12.9 “Buffers and credits” on page 114.

6.2.6 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. By choosing 2 Gb/s technology, 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 switches.

6.2.7 “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 GBIC, 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.

6.2.8 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 6-13.

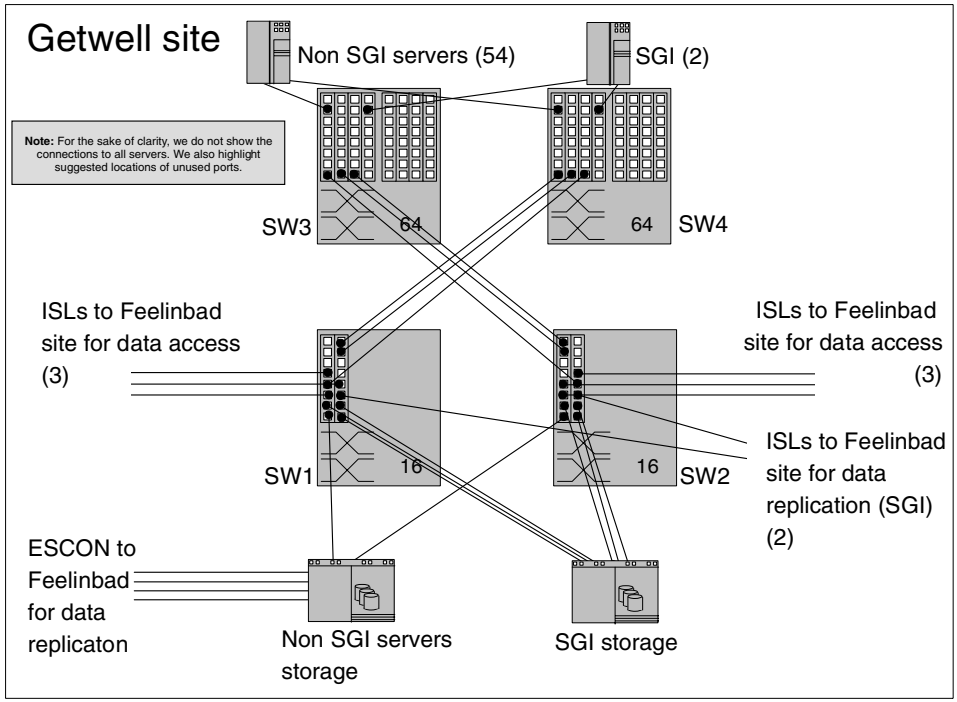


Figure 6-13 Core switch design for Getwell site

In Table 6-9 we show the number of ports used for solution.

Table 6-9 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 6-14.

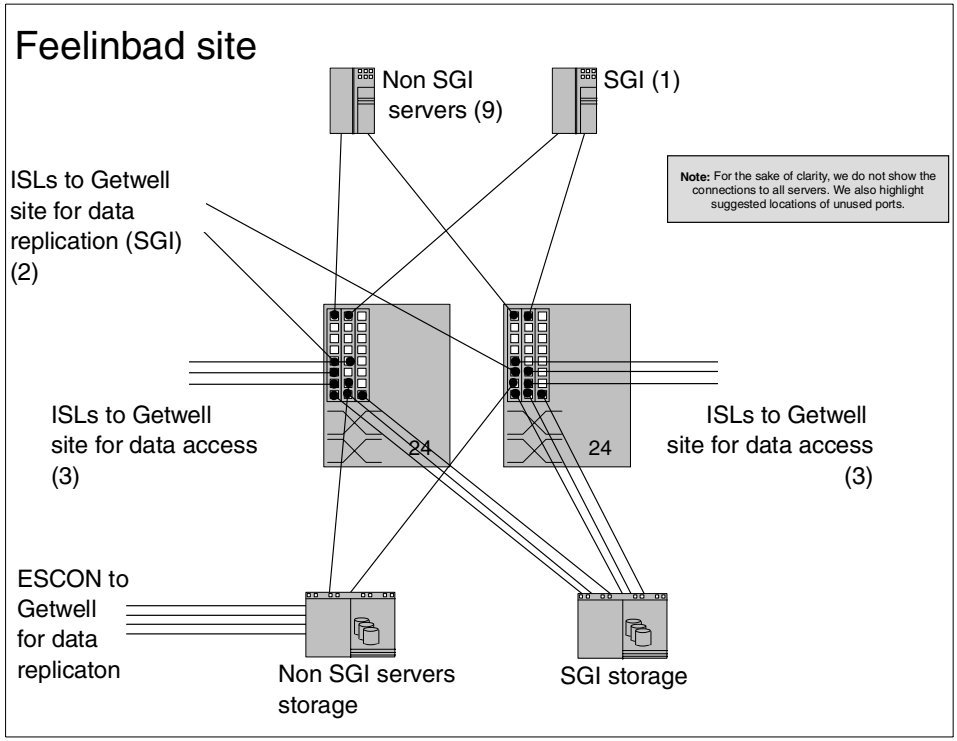


Figure 6-14 Core switch design for Feelinbad site

In Table 6-10 we show the number of ports used for our solution.

Table 6-10 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 6-11 on page 337. With this we would use an additional two ports in each core switch.

6.3 Case Study 3: Company Three

If we consider the company and its requirements as detailed in 5.3 “Case Study 3: Company Three” on page 299, we will propose this design.

6.3.1 Design

In considering the requirements as detailed in 5.3.5 “Analysis (ports and throughput)” on page 301, we decided to use two 1 Gb/s 16-port Fibre Channel switches for our design.

We show the proposed design in Figure 6-15.

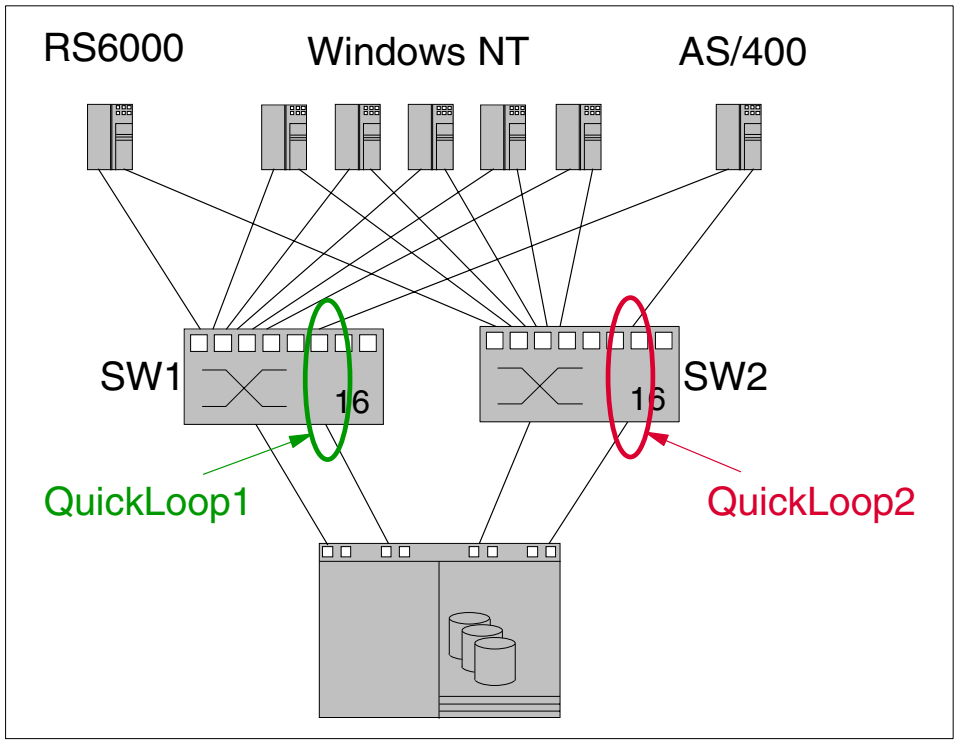


Figure 6-15 Proposed design for Company Three

In Table 6-11 we show the number of ports used for our solution.

Table 6-11 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 AS/400 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.

6.3.2 Performance

In the initial design, we have a 6:1 fan-out ratio for all servers except for AS/400. This means that six servers are accessing one storage port. The fan-out ratio for AS/400 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 AS/400 platform, we need to provide private FC-AL communication between the AS/400 server and his storage port. This means that we need to give the AS/400 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 M12) 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. More information about QuickLoop can be found in “QuickLoop” on page 127.

6.3.3 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 AS/400. The AS/400 does not support load balancing and failover of the traffic in the 5.1 release. So in case of maintenance or upgrades of the switches, we need to manually switch the used path within the operating system. This also means that the AS/400 access to the storage is not transparently redundant, because in the case of a failure on the active path, we will have to recover manually.

6.3.4 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 AS/400 traffic from other servers, because AS/400 is not capable of sharing storage devices with other systems at the time of writing this book. Because of this, we will introduce the following zones on both switches:

  - Zone for AS/400 server and its storage port
  - Zone for all other servers and their storage port

  Note: If you do not implement the zone for AS/400, it will not be able to access the storage. The zone for other servers is not necessary, but it is recommended.

6.3.5 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 GBICs.

6.3.6 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.

6.3.7 “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 GBIC, 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: Failover operations of AS/400 have to be done manually.

6.3.8 Manageability and management software

The management techniques are described in 6.1.8 “Manageability and management software” on page 329.

6.4 Case Study 4: Company Four

If we consider the company and its requirements as detailed in 5.4 “Case Study 4: Company Four” on page 303, we will propose two designs.

6.4.1 Design

If we consider the requirements as detailed in 5.4.5 “Analysis (ports and throughput)” on page 306, we have decided to split our solution into two parts. Because the 2109 family of switches do not support FICON, we will design open systems and S/390 systems separately. By breaking the design into two separate entities, we realize that at this stage we are not fulfilling one of the requirements which was to integrate S/390 and open systems into one SAN.

Open system SAN

For the open system SAN we decided to use two 1 Gb/s 16-port Fibre Channel switches for our design in both sites.

We show the proposed design for open systems at the East site in Figure 6-16.

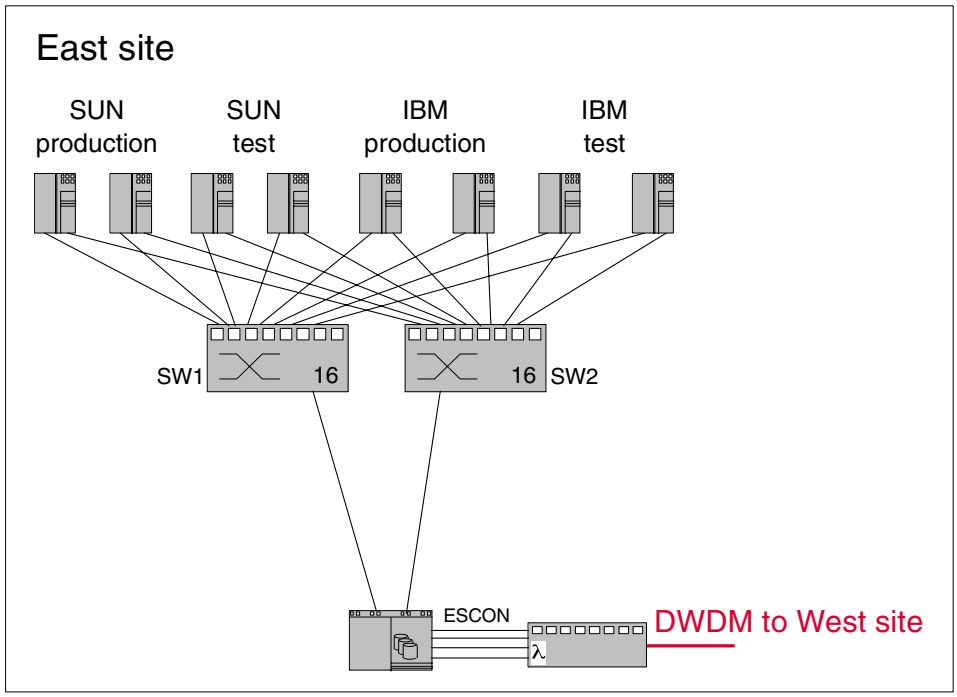


Figure 6-16 Open systems SAN for East site

In Table 6-12 we show the number of ports used for our solution.

Table 6-12 Number of used ports

Ports	Servers	Storage	ISLs	Spare
SW1	8	2	0	6
SW2	8	2	0	6
Total	16	4	0	12

The West site design is shown in Figure 6-17.

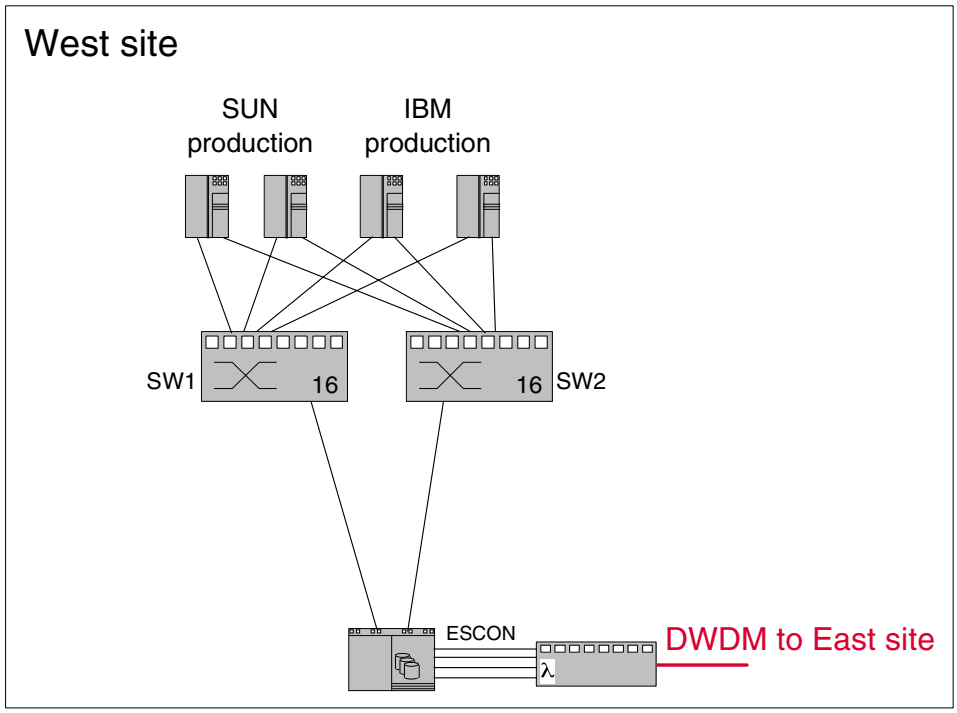


Figure 6-17 Open systems SAN for West site

In Table 6-13 we show the number of ports used for our solution.

Table 6-13 Number of used ports

Ports	Servers	Storage	ISLs	Spare
SW1	4	2	0	8
SW2	4	2	0	8
Total	8	4	0	16

S/390 design

In Figure 6-18 we show the design for the S/390 platform.

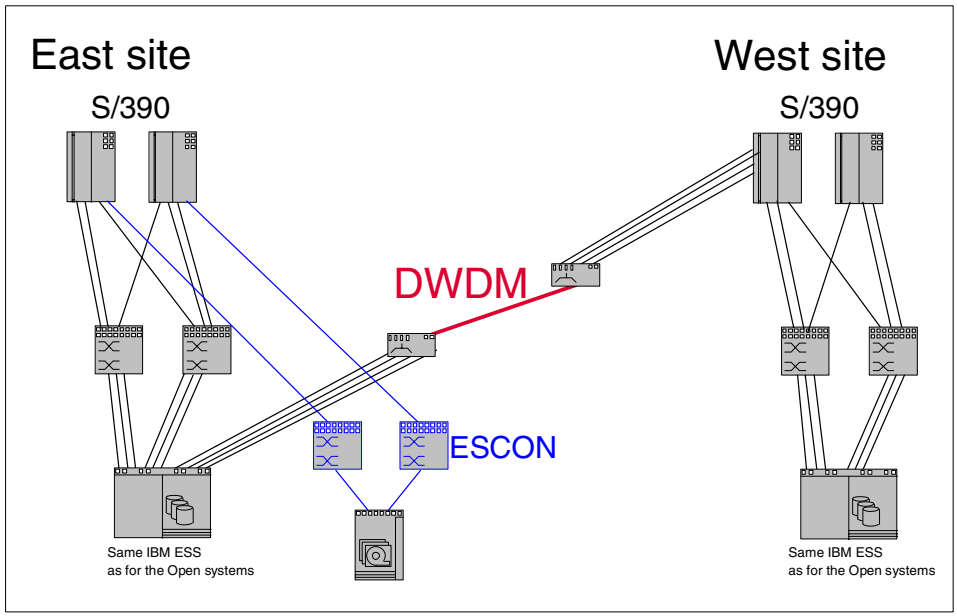


Figure 6-18 S/390 design

In the following sections we will only cover the open systems SAN.

With this 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. Because we are using a redundant SAN we can introduce additional switches without downtime on the existing servers.

For storage data replication we are using ESCON links over a DWDM connection. The peak bandwidth used by the servers is 11.58 MB/s. As only 30% of this peak are writes (3.48MB/s), we will allocate two ESCON channels for the data replication.

In the following sections, we will outline some aspects of the design.

6.4.2 Performance

In the initial design, we have a 6:1 fan-out ratio for all servers in 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 through the switches, where the latency is no greater than 2 microseconds.

6.4.3 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.

6.4.4 Security

When implementing the switches we need to take care of 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**

  In our example, we will not implement any zoning.

6.4.5 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 GBICs.

We are not introducing any ISLs from site to site.

6.4.6 Scalability

Within these 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 be probably 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.

6.4.7 “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 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 it using a hot-pluggable GBIC, 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.

6.4.8 Manageability and management software

The management techniques are the same as in Case Study 1, and we described them in 6.1.8 “Manageability and management software” on page 329.

You also need to consider to integrate the management of ATM products into your management infrastructure.

6.5 Case Study 5: Company Five

If we consider the company and its requirements as detailed in 5.5 “Case Study 5: Company Five” on page 308, we will propose the following design.

6.5.1 Design

In considering the requirements as detailed in “Analysis (ports and throughput)” on page 311, we have decided to use two 1 Gb/s 8-port Fibre Channel switches for our design.

We show the proposed design in Figure 6-19.

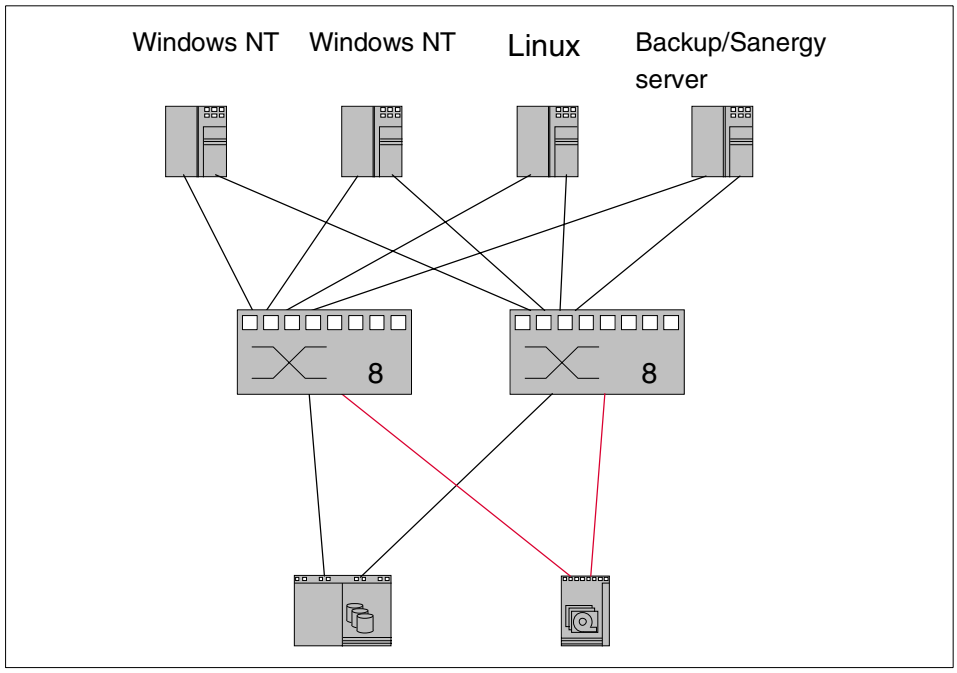


Figure 6-19 Proposed design for Company Five

In Table 6-14 we show the number of ports used for our solution.

Table 6-14 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.

6.5.2 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.

6.5.3 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.

6.5.4 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.

6.5.5 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 GBICs.

6.5.6 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.

6.5.7 “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 GBIC, 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.

6.5.8 Manageability and management software

The management techniques are the same as in Case Study 1, and are described in 6.1.8 “Manageability and management software” on page 329.

6.6 Case Study 6: Company Six

In considering the company and its requirements as detailed in 5.6 “Case Study 6: Company Six” on page 313, we will propose this design.

6.6.1 Design

Taking into account the requirements as detailed in 5.6.5 “Analysis (ports and throughput)” on page 315, we have decided to use two 1 Gb/s 16-port Fibre Channel switches for our design. You can see the proposed design for the Primary site in Figure 6-20.

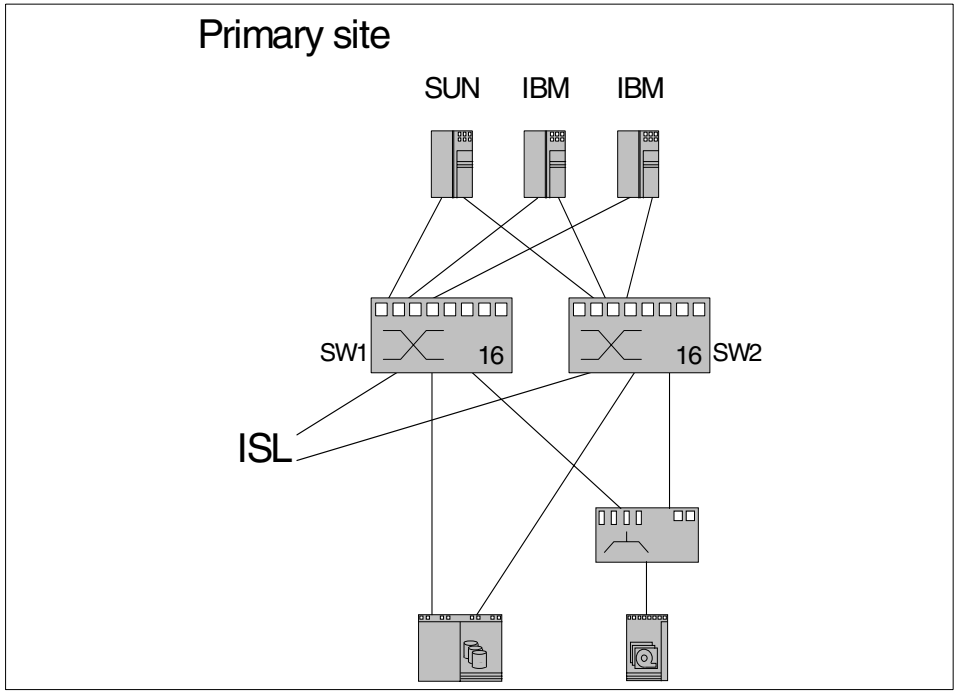


Figure 6-20 Proposed design for Primary site

In Table 6-15 we show the number of ports used for our solution.

Table 6-15 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 6-21.

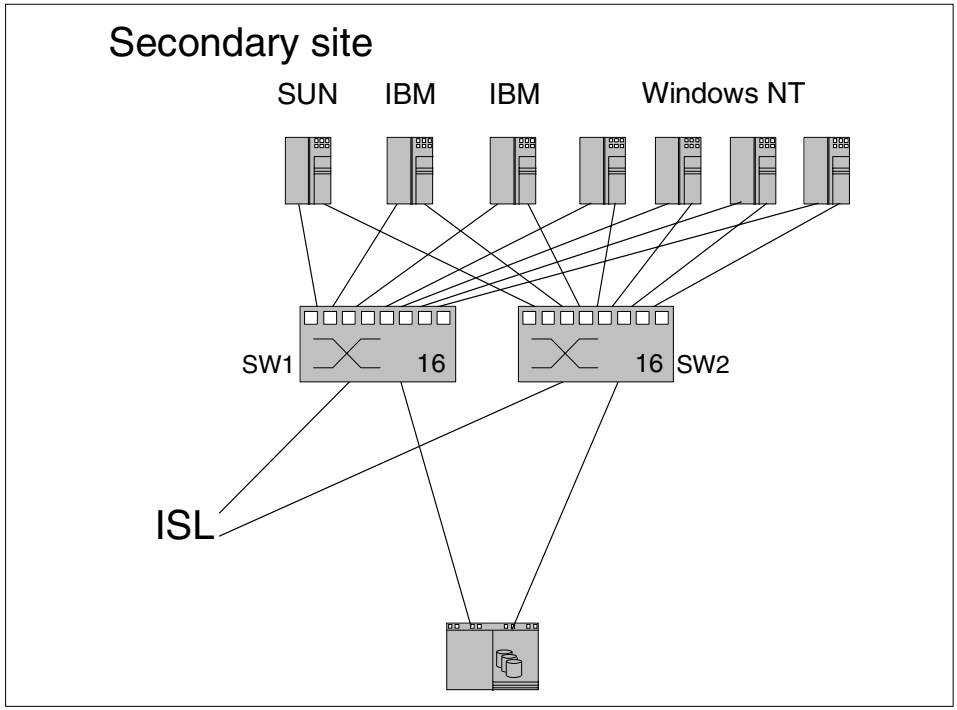


Figure 6-21 Proposed design for Secondary site

In Table 6-16 we show the number of ports used for our solution.

Table 6-16 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 6-22.

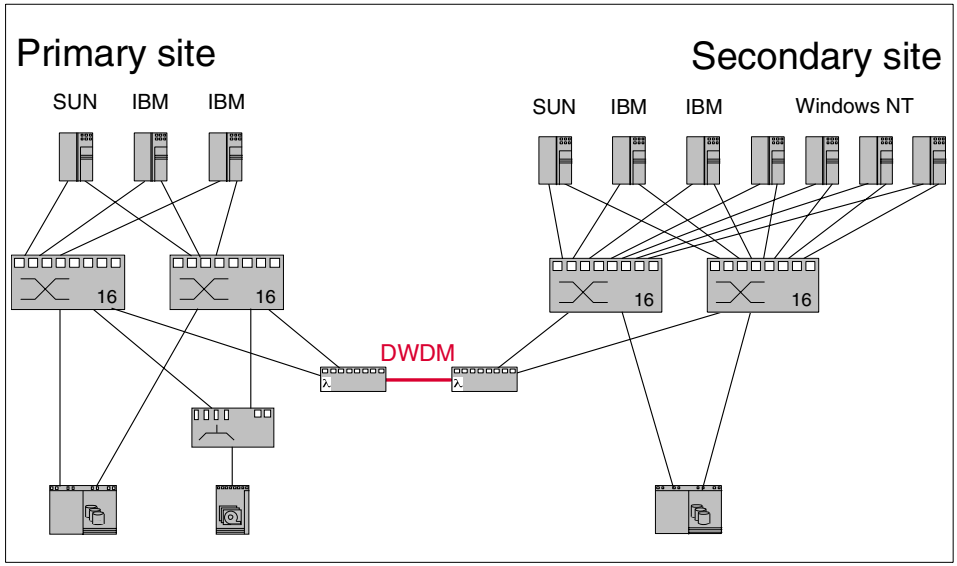


Figure 6-22 DWDM connection between sites

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.

6.6.2 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.

6.6.3 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.

6.6.4 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. 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 in separate then 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.

6.6.5 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 GBICs.

For connections from site-to-site, we use Longwave GBICs 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.

We describe this more in 2.12.9 “Buffers and credits” on page 114.

6.6.6 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.

6.6.7 “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 GBIC, 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.

6.6.8 Manageability and management software

The management techniques are the same as in Case Study 1, and we describe them in 6.1.8 “Manageability and management software” on page 329.

You will also need to consider to integrate the management of DWDM products into your management infrastructure.

7

INRANGE solutions

In this chapter we will describe solutions that are based upon the INRANGE FC/9000 Fibre Channel Director.

7.1 Case Study 1: Company One

If we consider the company and its requirements, as detailed in 5.1 “Case Study 1: Company One” on page 288, we will propose the following solution.

7.1.1 Design

When considering the requirements as detailed in 5.1.5 “Analysis (ports and throughput)” on page 289, we selected a solution based around the INRANGE FC/9000. 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 the discussions in Chapter 1, “Identifying your business and technology goals” on page 3. 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 7-1.

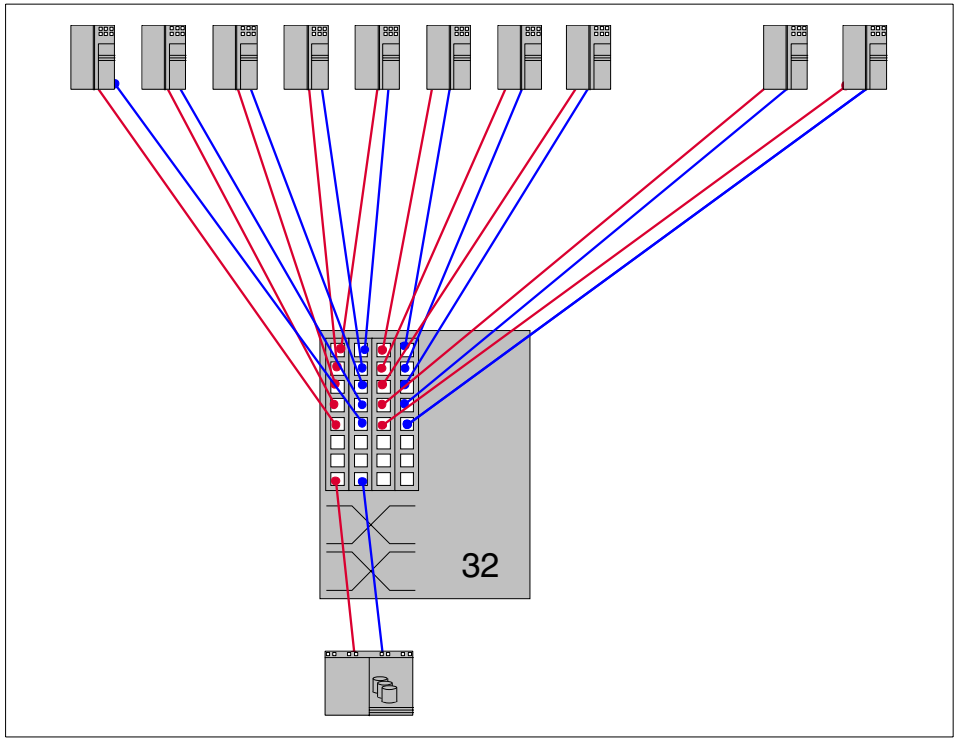


Figure 7-1 Single Director solution

There is a summary of the ports and their usage in Table 7-1.

Table 7-1 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 INRANGE FC/9000 allows us to upgrade firmware non disruptively.
- Possible growth without impact on the production. The INRANGE FC/9000 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 7-2.

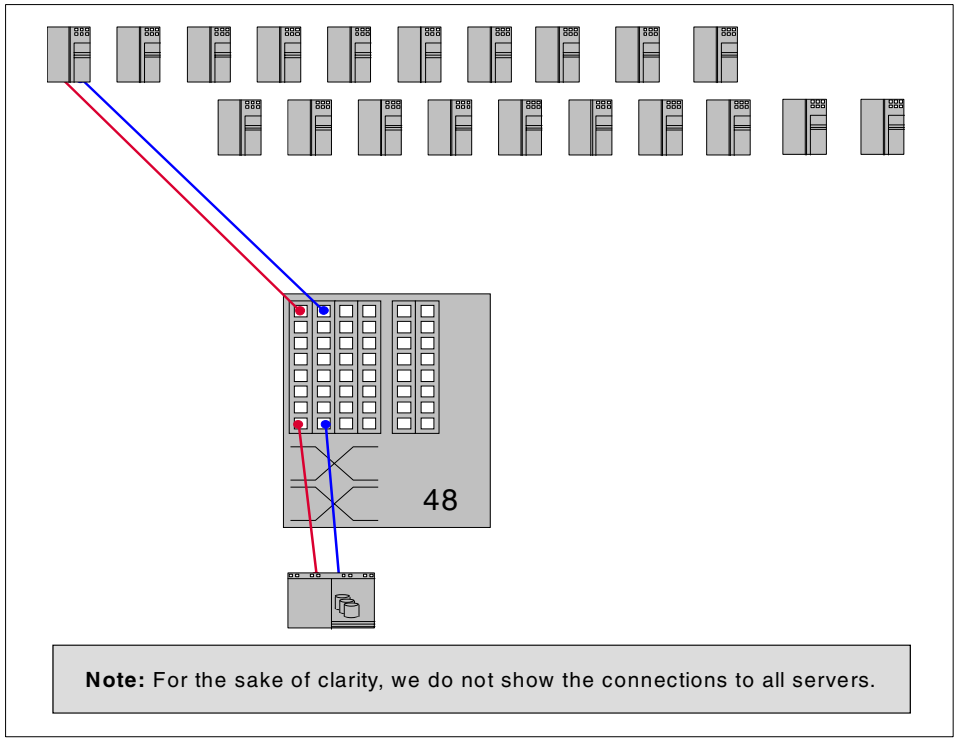


Figure 7-2 The 20-server solution

There is a summary of the ports and their usage in Table 7-2.

Table 7-2 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 7-3.

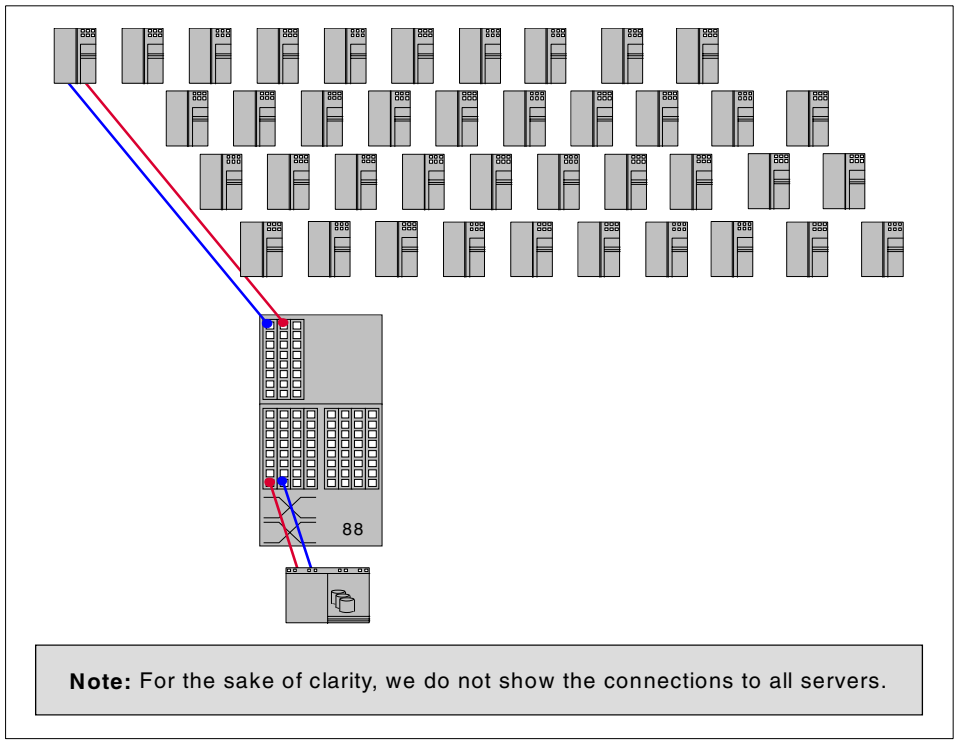


Figure 7-3 The 40 server, single Director solution

There is a summary of the ports and their usage in Table 7-3.

Table 7-3 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 FC/9000 over 64 ports required some down time (INRANGE 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 7-4.

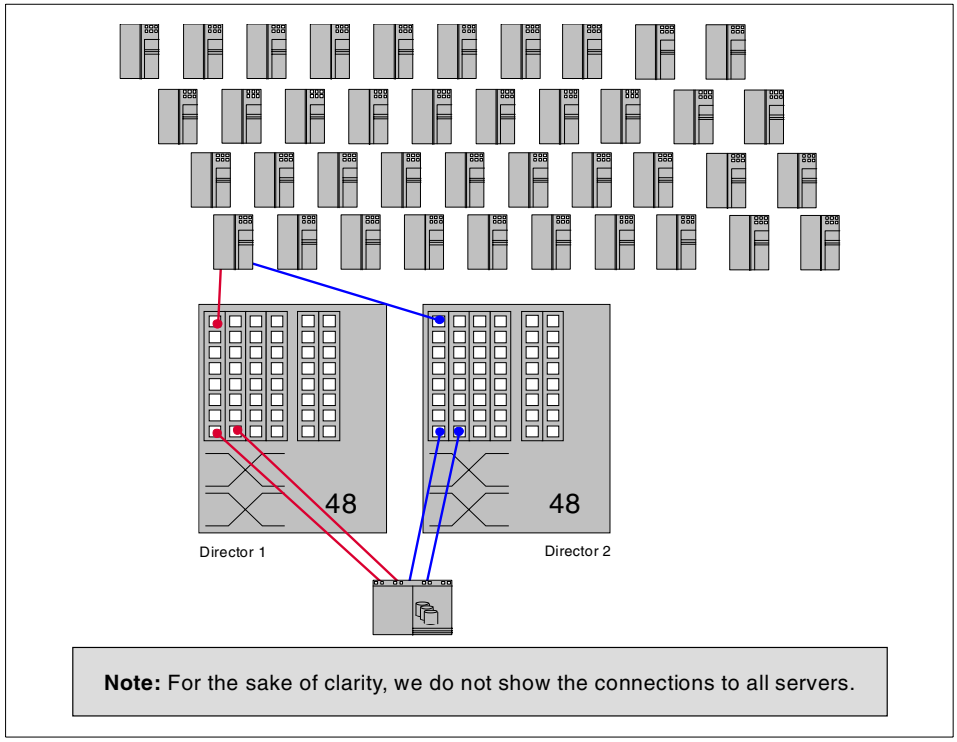


Figure 7-4 The 40-server, dual Director solution

There is a summary of the ports and their usage in Table 7-4.

Table 7-4 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 7-2 to the dual fabric solution in Figure 7-4 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.

7.1.2 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 is no need to go above 1 Gb/s at any point in the fabric, and there are no concerns over latency in the design.

7.1.3 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.

7.1.4 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.

7.1.5 Distance

In this case, there are no requirements for long distances. Shortwave lasers and multimode cable will allow 500m which is ample for this scenario.

7.1.6 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 Director 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.

7.1.7 "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 GBIC.
- 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.

7.1.8 Manageability and management software

The management of the SAN solution will be carried out using IN-VSN which was described in “Management software IN-Vision Enterprise Manager” on page 200. 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 7-5.

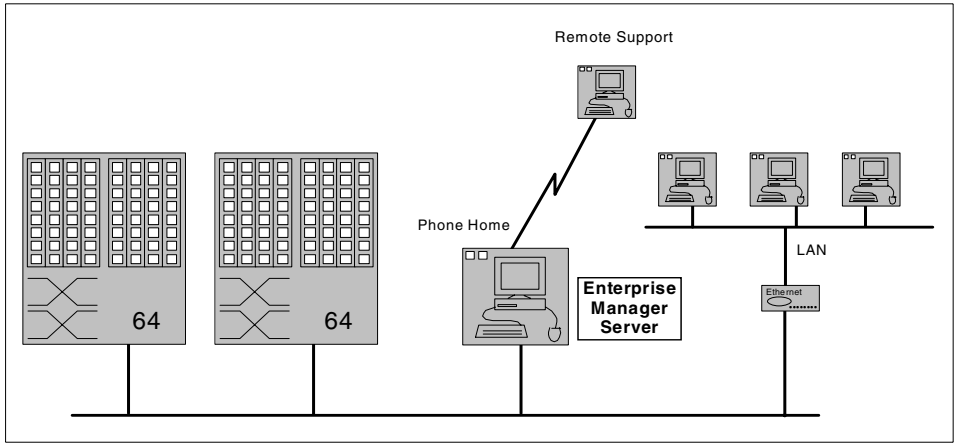


Figure 7-5 Management of an INRANGE SAN solution

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

7.2 Case Study 2: Company Two

If we consider the company and its requirements as detailed in 5.2 “Case Study 2: Company Two” on page 291, we will propose the following solution.

7.2.1 Design

In considering the requirements as detailed in 5.2.5 “Analysis (ports and throughput)” on page 294, we selected a solution based around the INRANGE FC/9000.

We have decided to implement all ISLs using 2 Gb/s technology:

- To provide ample bandwidth for disaster recovery scenarios

- To provide headroom for future expansion needs
- It makes this complex SAN somewhat less confusing.

This technology is not yet available. It should be possible to implement the solution using 2 Gb/s within the one year time frame specified by the requirements. If the SAN must be implemented before the 2 Gb/s technology comes to market, then this can be done by using 1 Gb/s components and links. There would need to be more links than we plan for using 2 Gb/s. This is discussed further in “Implementation using 1 Gb/s” on page 376.

The present port requirements for Company Two’s SAN ports are tabulated in Table 7-5.

Table 7-5 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				

To provide a solution for this company, we to need carefully consider the situation at the Getwell site. We have a need for 128 ports. This is an interesting number. The INRANGE FC/9000 currently offers 128 ports with a pre-announcement of support of up to 256 ports per Director. To advance beyond the 128 level requires a second cabinet.

Another way to achieve the 128 port count is to use two 64 port chassis. 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. The two at Getwell are capable of providing more than 64 ports each by having two chassis connected. This allows the SAN to expand in line with the customers predictions for the next few years without having to carry out any intrusive upgrades.

Getwell site

The design for Getwell is shown in Figure 7-6.

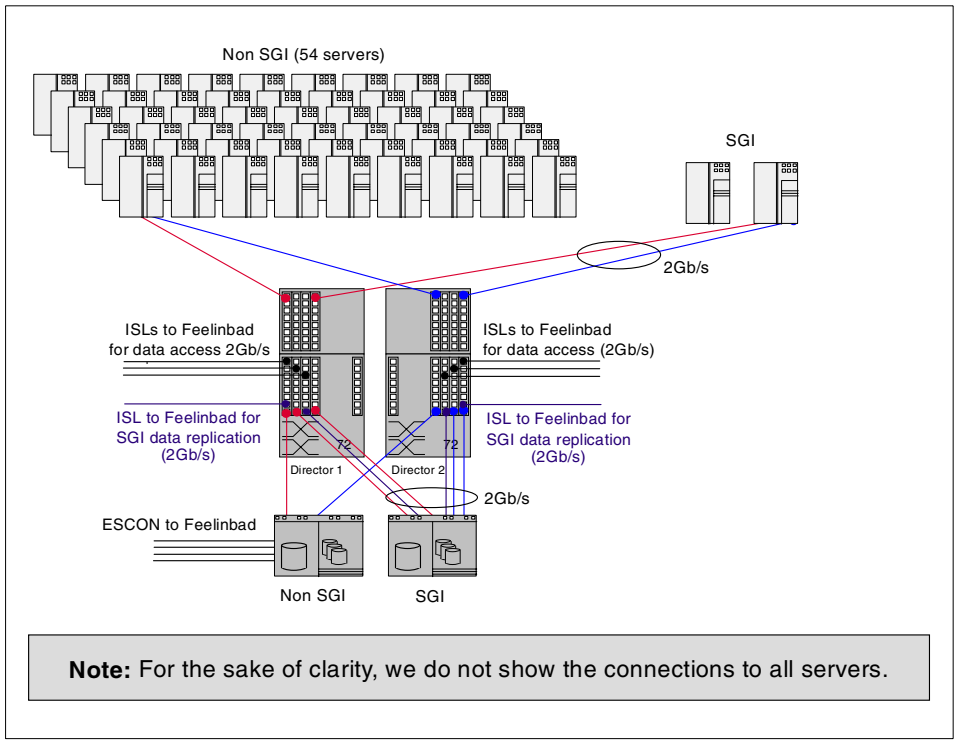


Figure 7-6 The Getwell configuration

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
- Sixteen free SAN ports
- Four ESCON 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 7-6.

Table 7-6 Port usage at Getwell

Director	Servers	Storage	ISLs	Spare
Director 1	56	4	4	8
Director 2	56	4	4	8
Total	112	8	8	16

Feelinbad site

The design for Feelinbad is shown in Figure 7-7.

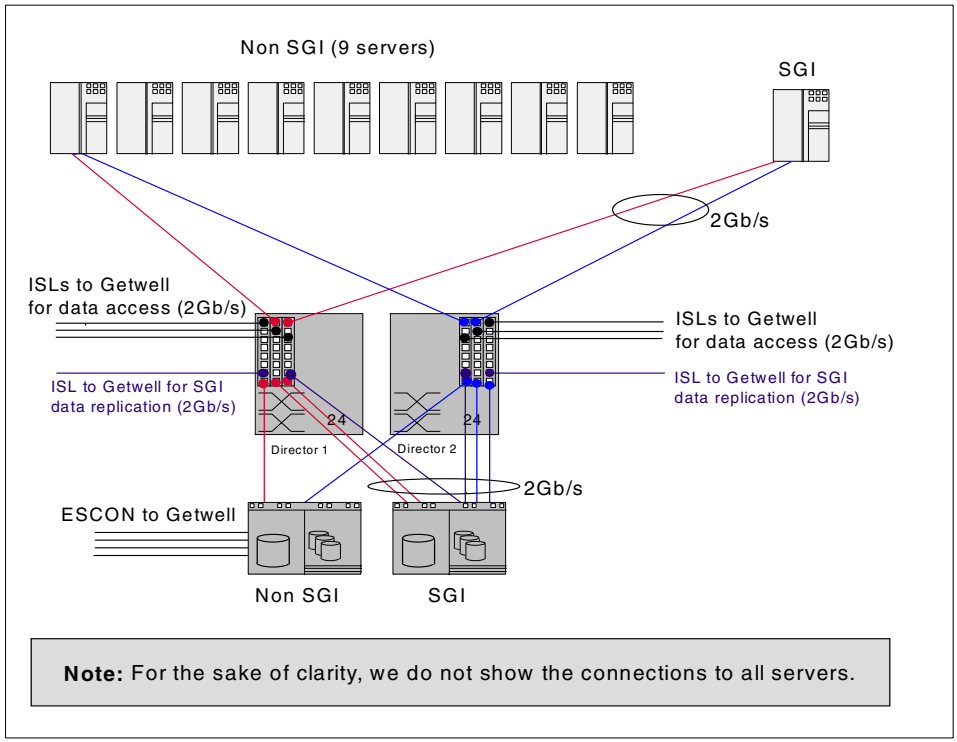


Figure 7-7 The Feelinbad configuration

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
- Twelve free SAN ports
- Four ESCON 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 7-7.

Table 7-7 Port usage at Feelinbad

Director	Servers	Storage	ISLs	Spare
Director 1	10	4	4	6
Director 2	10	4	4	6
Total	20	8	8	12

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. The INRANGE FC/9000 allows us to upgrade firmware non disruptively.
- Possible growth without impact on the production. The INRANGE FC/9000 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

In the following sections, we outline some aspects of the design.

Implementation using 1 Gb/s

In the event that the SAN needs to be implemented before 2 Gb/s technology becomes available, some changes will need to be made. There will need to be more ISLs to cover the bandwidth requirements, and also there will need to be more links into the SGI storage for the same reasons. It might be assumed that because the bandwidth of the links is half that we would need twice as many. In fact, the designs have been created allowing for some resilience by redundancy. As this is N+1 redundancy rather than N+N redundancy, we do not need twice as many links.

In fact the port count with a 1 Gb/s design can be seen below in Table 7-8 and Table 7-9 1 Gb/s port usage at Getwell.

Table 7-8 1Gb/s port usage at Getwell

Director	Servers	Storage	ISLs	Spare
Director 1	58	6	6	2
Director 2	58	6	6	2

Table 7-9 1Gb/s b/s port usage at Feelinbad

Director	Servers	Storage	ISLs	Spare
Director 1	11	6	6	1
Director 2	11	6	6	1

As we can see, the whole project can be carried out using 1 Gb/s devices and links, without needing to increase the port count.

7.2.2 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.

In our example, the strong recommendation is that all ISLs are 2 Gb/s. There is a requirement for some of the ISLs and all the SGI fibers to be running at 2 Gb/s and in fact it might be considered sensible to implement the entire solution using 2 Gb/s technology. This would make the entire topology more uniform and easier to understand, document and manage, as shown in Figure 7-6 on page 374, and Figure 7-7 on page 375. There will be a slight increase in the additional outlay, but this is justified by the simpler solution and the additional bandwidth available for growth.

7.2.3 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.

7.2.4 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.

7.2.5 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 GBICs as the servers are within a radius of 500 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 Longwave GBICs which will be connected over DWDM products. With the use of a DWDM products we will solve the problem of degradation of laser signal. The INRANGE FC/9000 also allocates enough BB_Credit to all ports to cope easily with this distance. You can read more about buffers in 2.12.9 “Buffers and credits” on page 114.

7.2.6 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.

7.2.7 “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 GBIC.
- 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.

7.2.8 Manageability and management software

The management techniques are the same as in Case Study 1, and we described them in 7.1.8 “Manageability and management software” on page 372.

7.3 Case Study 3: Company Three

If we consider the company and its requirements as detailed in 5.3 “Case Study 3: Company Three” on page 299, we will propose the following solution.

7.3.1 Design

Considering the requirements as detailed in 5.3.5 “Analysis (ports and throughput)” on page 301, we selected a solution based around the INRANGE FC/9000. The smallest configuration available via IBM is 24 ports.

The design can be seen in Figure 7-8.

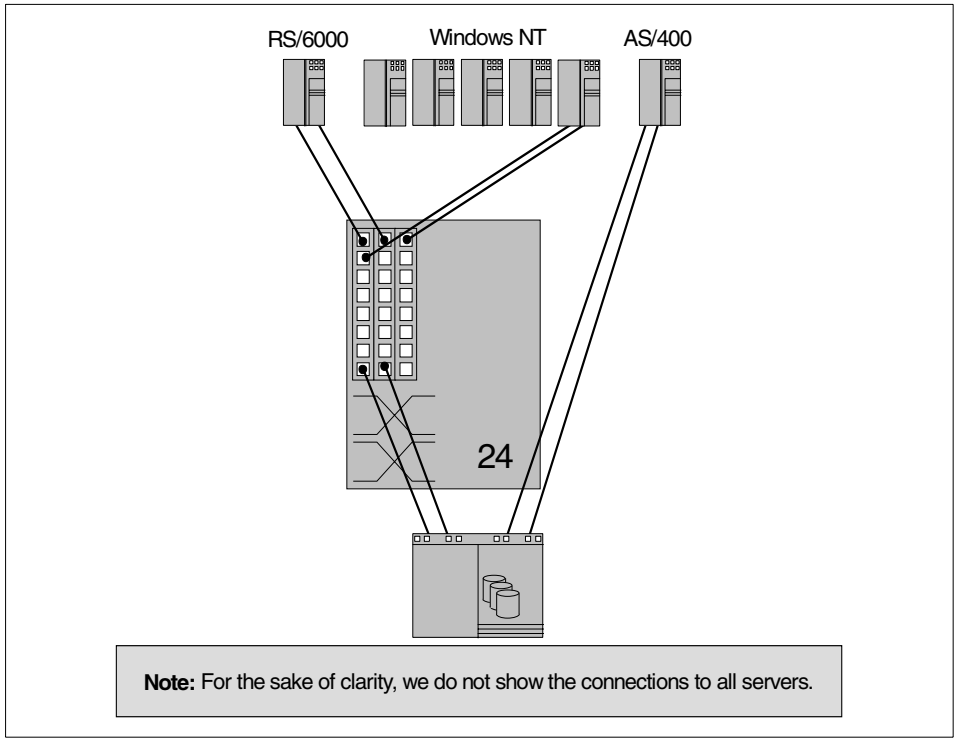


Figure 7-8 The proposed design for Company Three

Note: The design does not connect the AS/400 to the SAN.

It will be noticed that the AS/400 is not connected to the SAN at this stage. This is because it is not certified to work at present. It is not unreasonable to expect this to be certified at some time in the future and at that point, it will be possible to simply connect the AS/400 into the SAN using spare ports on the FC/9000.

There is a summary of the ports and their usage in Table 7-10.

Table 7-10 Port usage at Company Three

Director	Servers	Storage	ISLs	Spare
Director 1	12	2	0	10

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. The FC/9000 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.

7.3.2 Performance

In the initial design we have a 6:1 fan-out ratio for all servers except for AS/400. This means that six 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.

7.3.3 Availability

The 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.

7.3.4 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**

  In this case we will not implement any zoning.

7.3.5 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 GBICs.

7.3.6 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.

7.3.7 “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 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: The FC/9000 has redundant power supplies.
- Director port: If one of the ports fails, you may replace it using a hot-pluggable GBIC, 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.

7.3.8 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 372.

7.4 Case Study 4: Company Four

If we consider the company and its requirements as detailed in 5.4 “Case Study 4: Company Four” on page 303, we will propose the following solution.

7.4.1 Design

If we consider the requirements as detailed in 5.4.5 “Analysis (ports and throughput)” on page 306, we selected a solution based around the INRANGE FC/9000. A feature of the INRANGE FC/9000 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 to accommodate the storage and all the servers into a single SAN solution.

The proposed solution for the East site is shown in Figure 7-9.

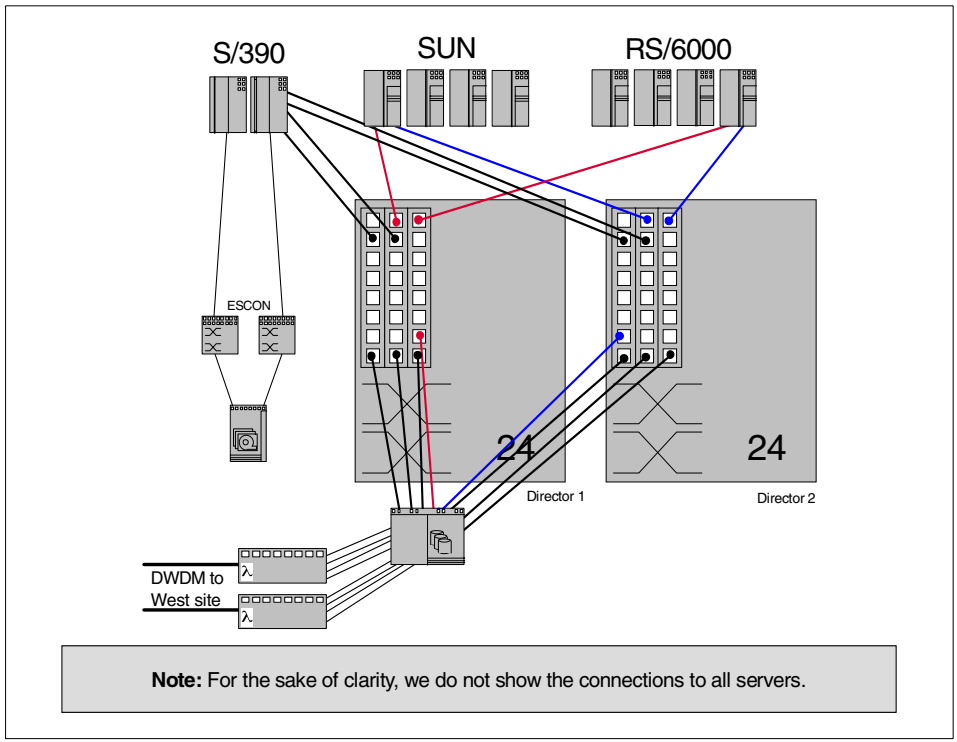


Figure 7-9 The proposed solution for the East site

There is a summary of the ports and their usage in Table 7-11.

Table 7-11 Port usage at the East site

Director	Servers	Storage	ISLs	Spare
Director 1	12	4	0	8
Director 2	12	4	0	8
Total	24	8	0	16

The proposal is to use four FICON connections, from the S390s, 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 7-10.

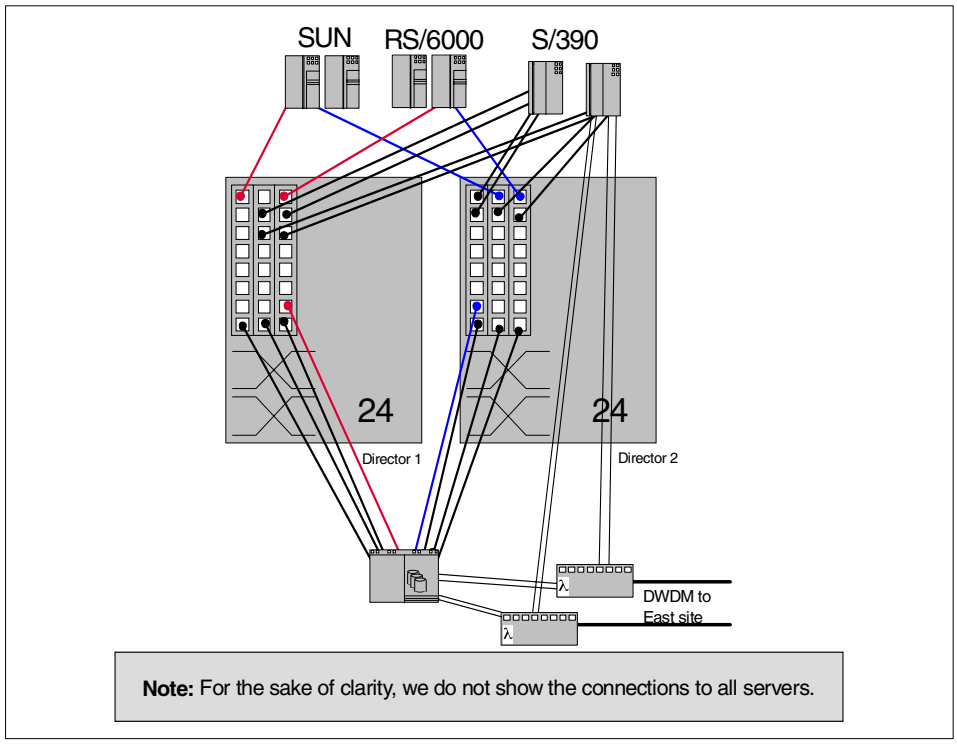


Figure 7-10 The proposed solution for the West site

There is a summary of the ports and their usage in Table 7-12.

Table 7-12 Port usage at the West Site

Director	Servers	Storage	ISLs	Spare
Director 1	12	4	0	8
Director 2	12	4	0	8
Total	24	8	0	16

The FICON connections from the S/390 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 7-11.

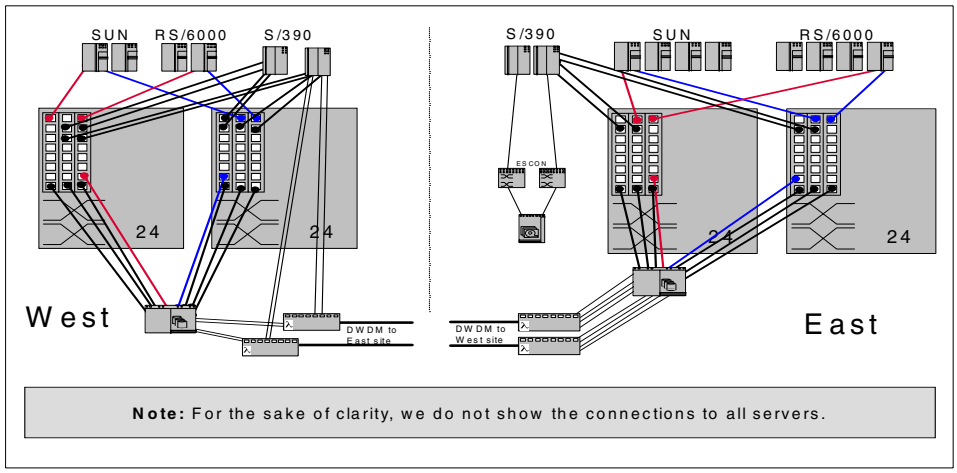


Figure 7-11 The combined solution

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 INRANGE FC/9000 allows us to upgrade Firmware non disruptively.
- Possible growth without impact on the production. The INRANGE FC/9000 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.

7.4.2 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.

7.4.3 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.

7.4.4 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.

7.4.5 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 GBICs.

There are no ISLs from site-to-site. The DWDM connections between the two sites will utilize the fiber optic connections.

7.4.6 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.

7.4.7 “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 GBIC, 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.

7.4.8 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 372.

7.5 Case Study 5: Company Five

If we consider the company and its requirements as detailed in 5.5 “Case Study 5: Company Five” on page 308, we will propose the following solution.

7.5.1 Design

Considering the requirements as detailed in 5.5.5 “Analysis (ports and throughput)” on page 311, we selected a solution based around the INRANGE FC/9000. We are using a 24-port director as this is the smallest configuration available via IBM.

The proposed design is shown in Figure 7-12.

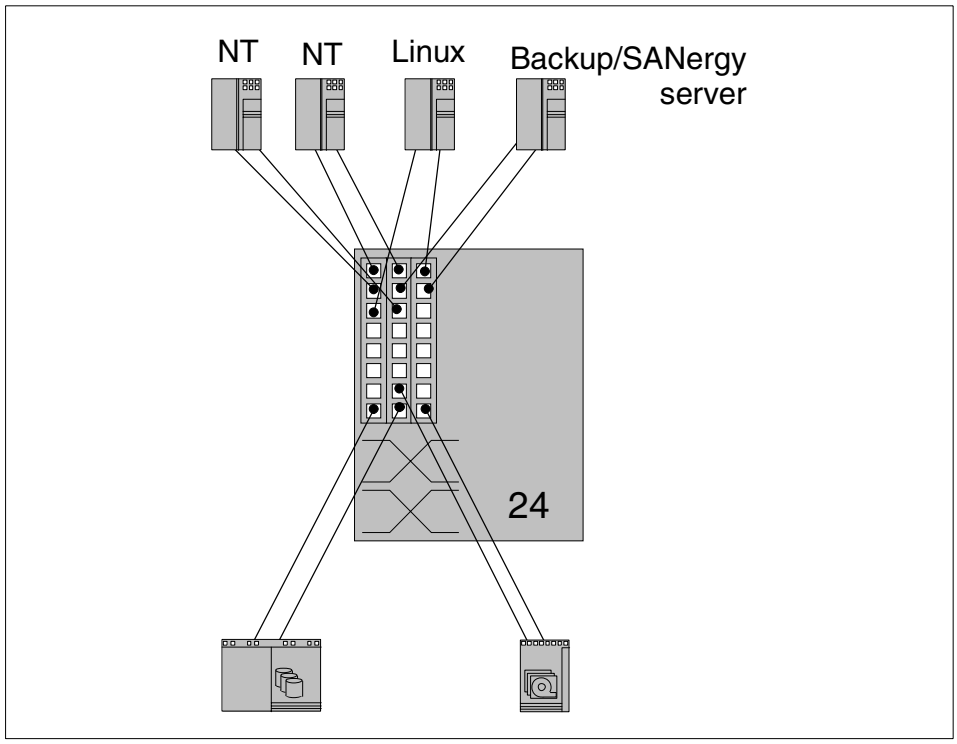


Figure 7-12 The proposed design for Company Five

There is a summary of the ports and their usage in Table 7-13.

Table 7-13 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 used supports FC-AL connectivity, we can connect tape devices directly to it.

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.

7.5.2 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.

7.5.3 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.

7.5.4 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.

7.5.5 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 GBICs.

7.5.6 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.

7.5.7 "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 GBIC, 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.

7.5.8 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 372.

7.6 Case Study 6: Company Six

If we consider the company and its requirements as detailed in 5.6 “Case Study 6: Company Six” on page 313, we will propose the following solution.

7.6.1 Design

Considering the requirements as detailed in 5.6.5 “Analysis (ports and throughput)” on page 315, we selected a solution based around the INRANGE FC/9000.

You can see the proposed design for Primary site in Figure 7-13.

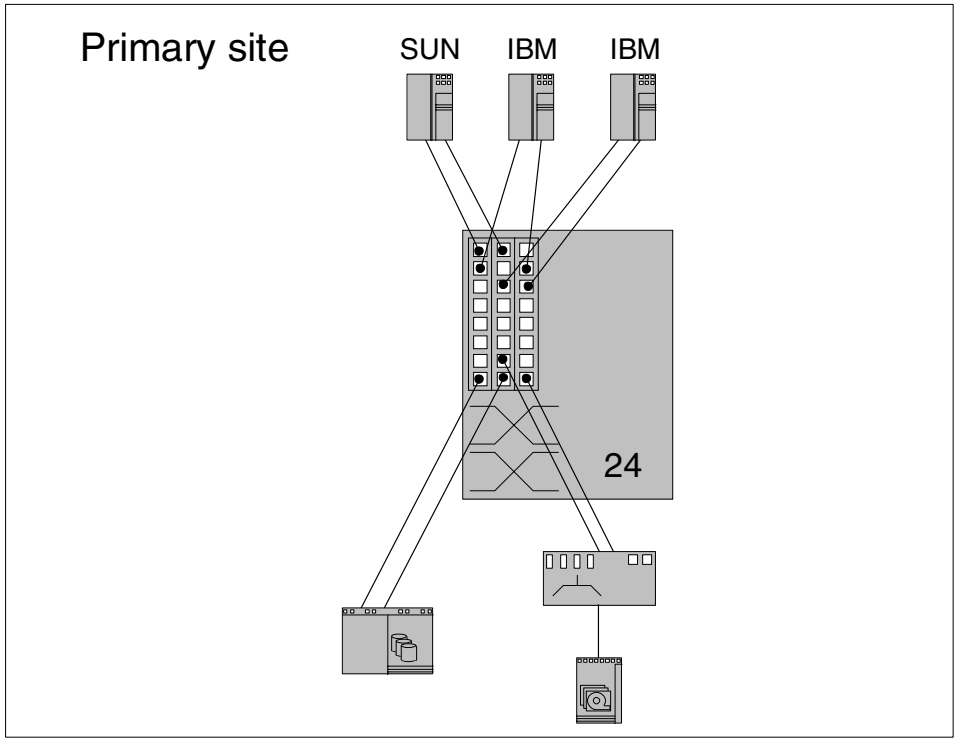


Figure 7-13 The proposed design for the Primary site

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 7-14.

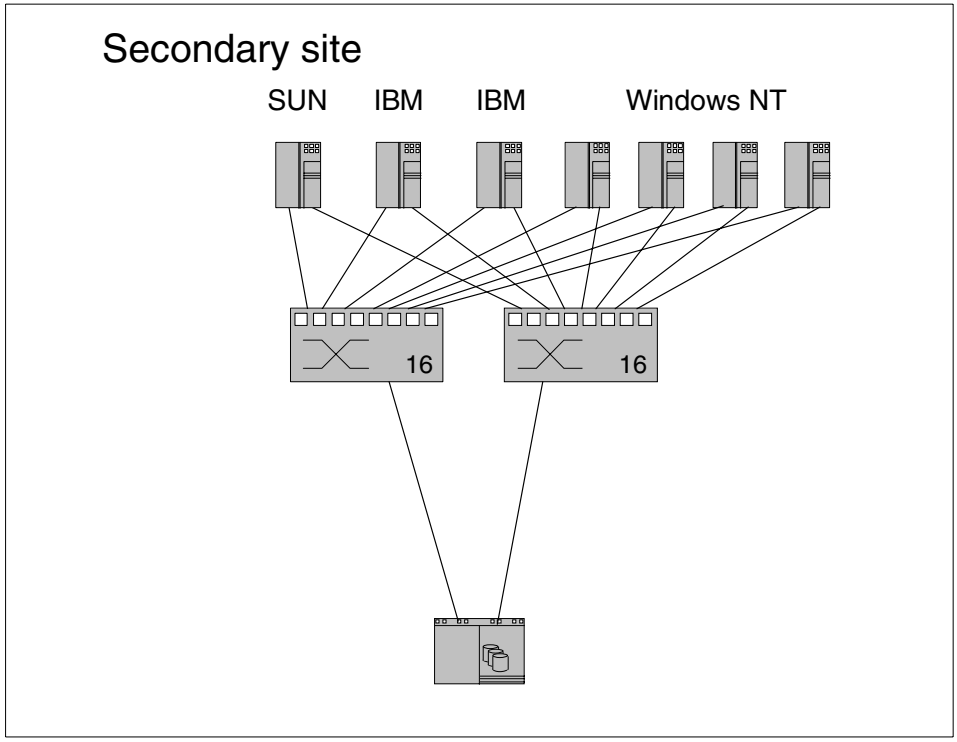


Figure 7-14 The proposed design for the Secondary site

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 7-15.

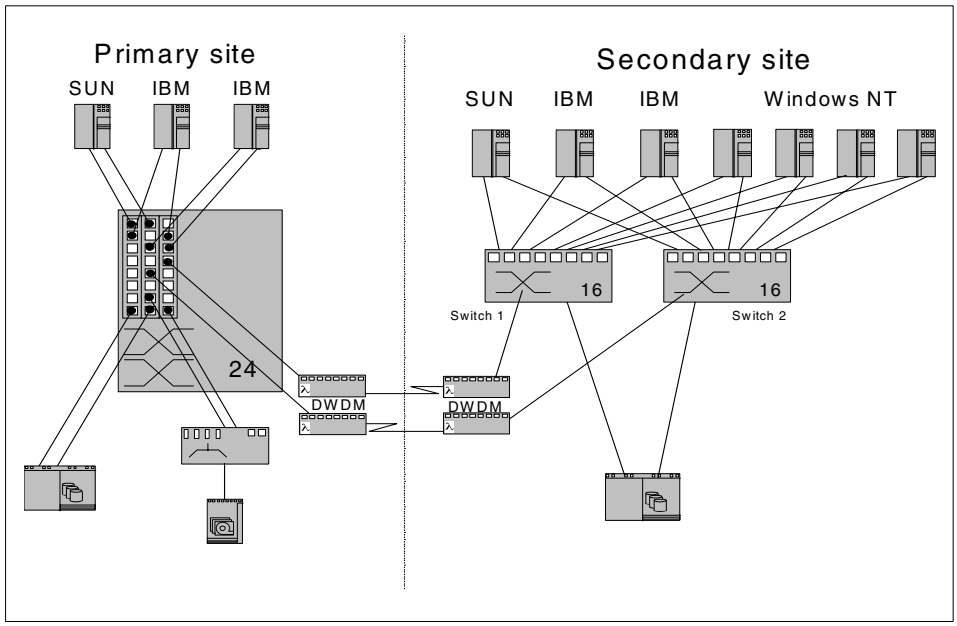


Figure 7-15 The complete solution

There is a summary of the ports and their usage in Table 7-14 and Table 7-15.

Table 7-14 Port usage at the Primary site

Director	Servers	Storage	ISLs	Spare
Director 1	6	4	2	12

Table 7-15 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.

7.6.2 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.

7.6.3 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.

7.6.4 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.

7.6.5 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 GBICs.

For connections from site-to-site we will use Longwave GBICs, 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. We do not need to carry out any such actions on the Director as the FC/9000 assigns 64 BB_Credits to all ports. You can read more about buffers in 2.12.9 “Buffers and credits” on page 114.

7.6.6 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.

7.6.7 “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 GBIC, 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.

7.6.8 Manageability and management software

The management techniques are the same as in Case Study 1 and we described them in 7.1.8 “Manageability and management software” on page 372.

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.

8

McDATA solutions

In this chapter, we will show solutions to the case studies based on the products in the IBM McDATA portfolio.

8.1 Case Study 1: Company One

If we consider the company and its requirements as detailed in 5.1, “Case Study 1: Company One” on page 288, we will propose the following solution.

8.1.1 Design using Directors

Considering the requirements as detailed in 5.1.5, “Analysis (ports and throughput)” on page 289, we have proposed in this section a Director-based solution for our design.

We show the proposed design in Figure 8-1.

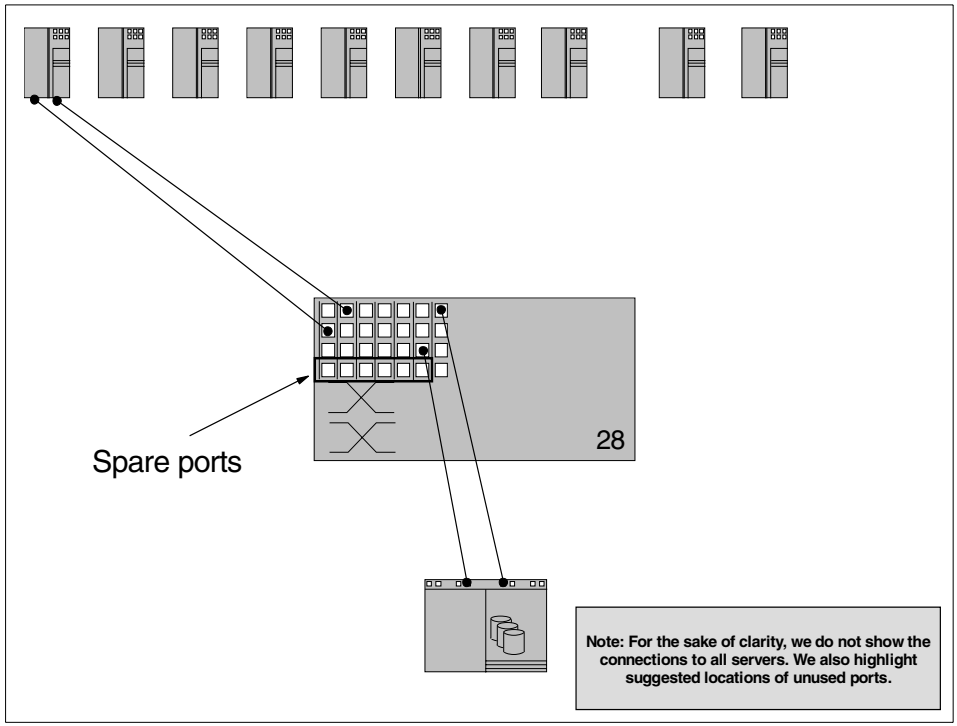


Figure 8-1 Core SAN design using a McDATA ED-6064 Director

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- 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 upgrade, 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 11 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.

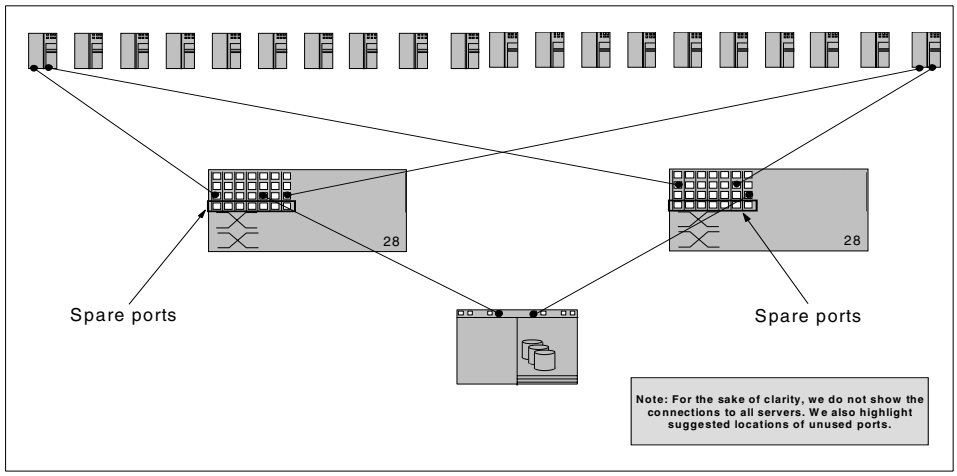


Figure 8-2 Fully redundant McDATA ED-6064 Director solution

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 8-3.

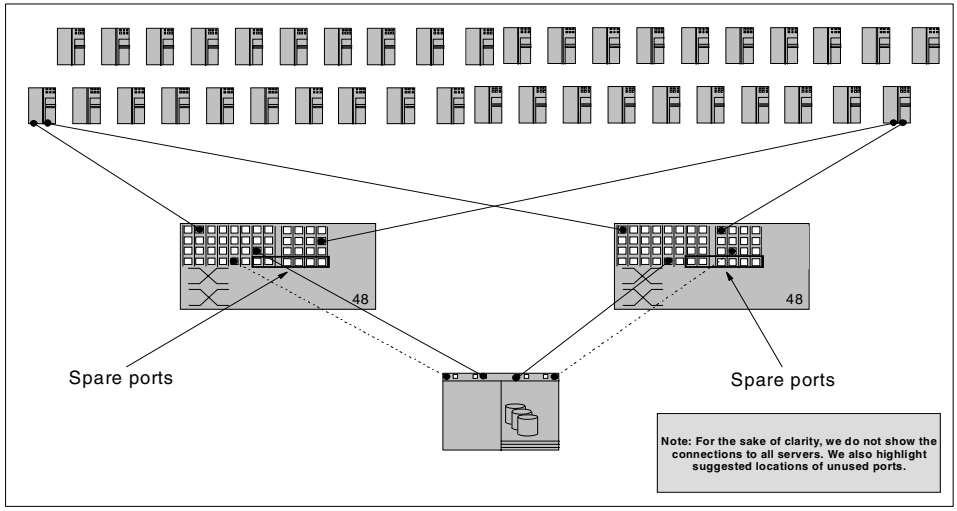


Figure 8-3 McDATA ED-6064 solution with all potential servers

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 8-1 we show Year 1 requirements.

Table 8-1 Year 1 ports: 10 servers

Storage	Server	ISL	Spare
2	20	0	6

In Table 8-2 we show Year 2 requirements.

Table 8-2 Year 2 ports: 20 servers

Storage	Server	ISL	Spare
2	40	0	14

In Table 8-3 we show Year 3 requirements.

Table 8-3 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.

8.1.2 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.

8.1.3 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.

8.1.4 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). 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 in separately, then into zones for performance reasons. This can be implemented 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.

8.1.5 Distance

In all the connections, we will use shortwave GBICs as the servers are within a radius of 500 m.

8.1.6 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.

8.1.7 “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 GBIC.
- 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.

8.1.8 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 8-4.

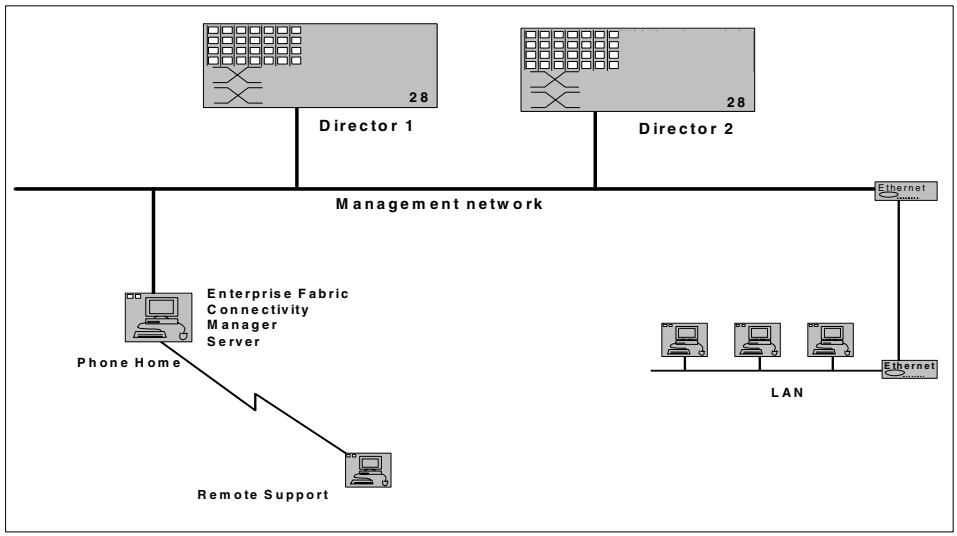


Figure 8-4 Management network

Note: This network can be part of your existing network infrastructure.

Full functional management is available through EFC Manager software.

8.1.9 Design using switches

Considering the requirements as detailed in 5.1.5, “Analysis (ports and throughput)” on page 289, we have proposed in this section a switch based solution for our design.

We show the proposed design in Figure 8-5.

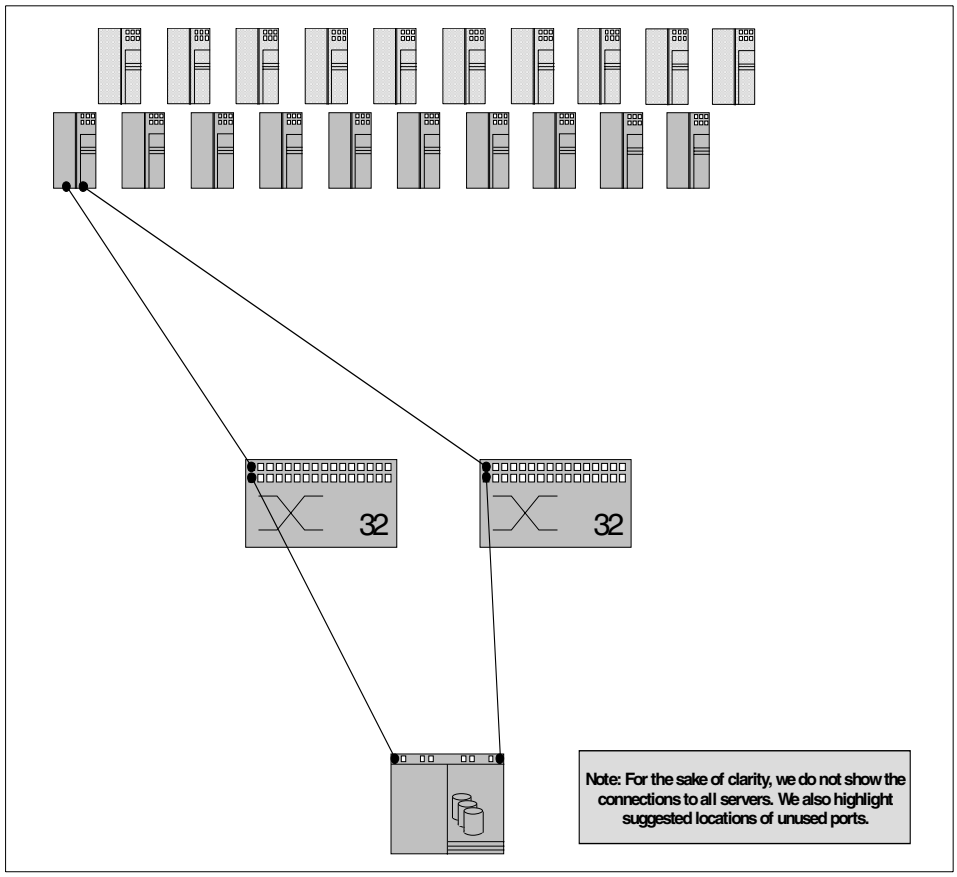


Figure 8-5 Initial design using McDATA ES-3032 Switches

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).
- 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.

With this design, we have 6 SAN ports free for future expansion. This means that you could potentially connect an additional three 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 server 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 primary connections for the new servers. The new servers will need a total of 20 secondary connections across the original switches which can be achieved as we are only using 21 of the 32-ports per switch (that is (32-21)*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 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 8-6.

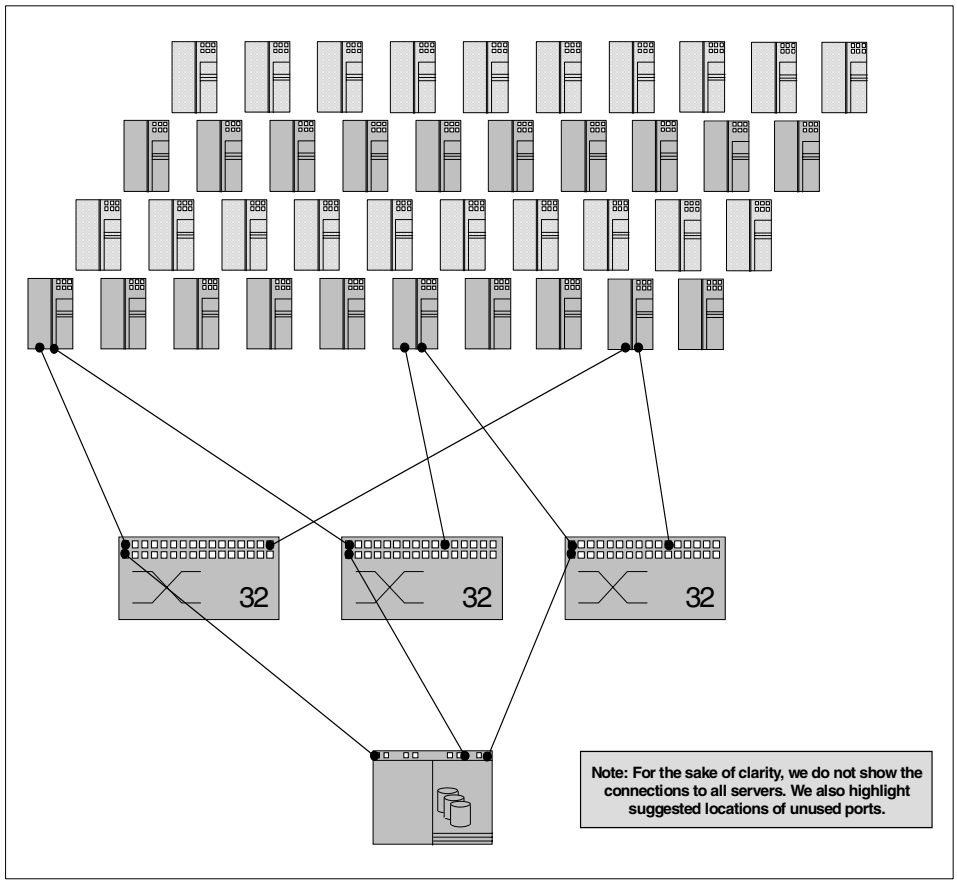


Figure 8-6 Final design to accommodate all potential servers

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 8-4 we show Year 1 ports.

Table 8-4 Year 1 ports: 10 servers

Storage	Server	ISL	Spare
2	20	0	42

In Table 8-5 we show Year 2 ports.

Table 8-5 Year 2 ports: 20 servers

Storage	Server	ISL	Spare
2	40	0	22

In Table 8-6 we show Year 3 ports.

Table 8-6 Year 3 ports: 40 servers

Storage	Server	ISL	Spare
4	80	0	12

We also took these aspects into account.

8.1.10 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.

8.1.11 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 ES-3032 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 be temporarily single points of failure for six servers as the secondary link is migrated to the new switch.

8.1.12 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). Because we have only one storage port for all servers there is no need to implement any zoning.

8.1.13 Distance

In all the connections, we 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 extending the fabric.

8.1.14 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.

8.1.15 “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 ES-3032 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.

8.1.16 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 8-7.

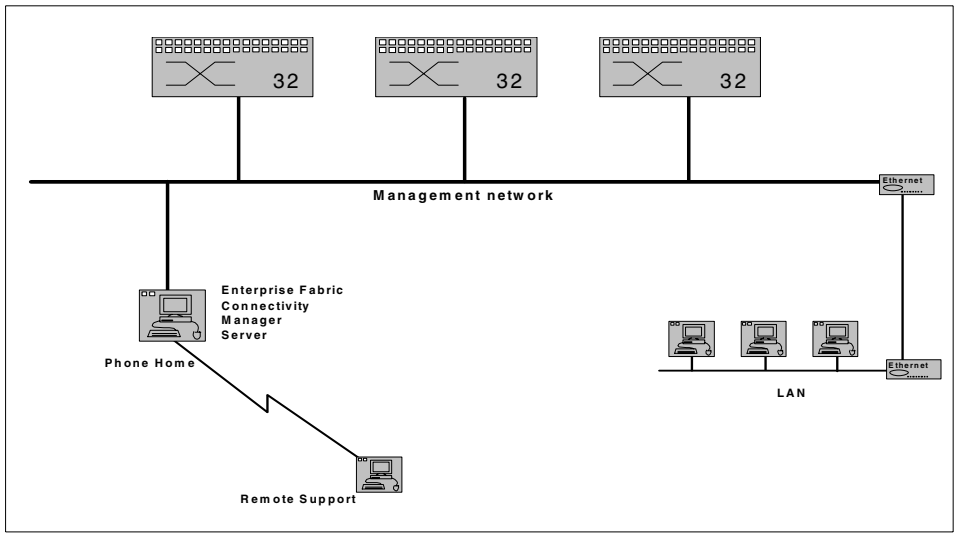


Figure 8-7 Management network

Note: This network can be part of your existing network infrastructure.

Full functional management is available through EFC Manager software.

8.2 Case Study 2: Company Two

If we consider the company and its requirements, as detailed in 5.2, “Case Study 2: Company Two” on page 291, we will propose the following solution.

8.2.1 Design

Considering the requirements as detailed in 5.2.5, “Analysis (ports and throughput)” on page 294, 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 8-8.

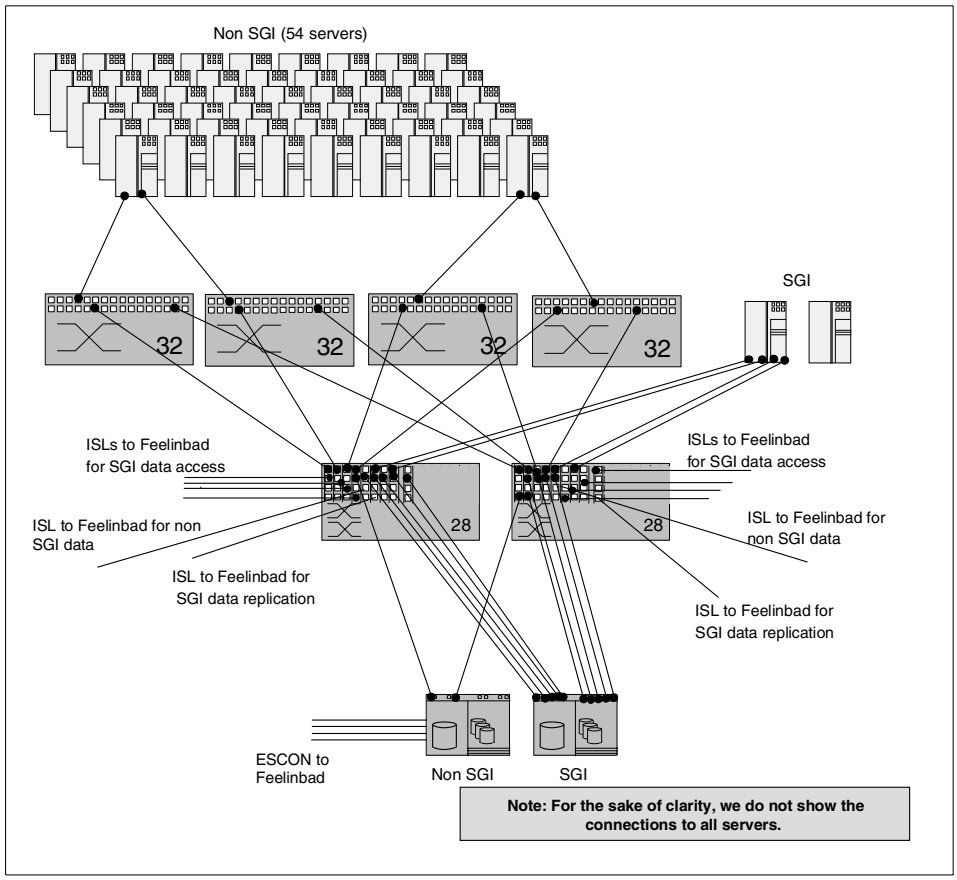


Figure 8-8 Getwell SAN design using McDATA ED-6064 Director and ES-3032 switches

To summarize the port allocation:

- One hundred eight 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 ESCON 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 8-7 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 8-9

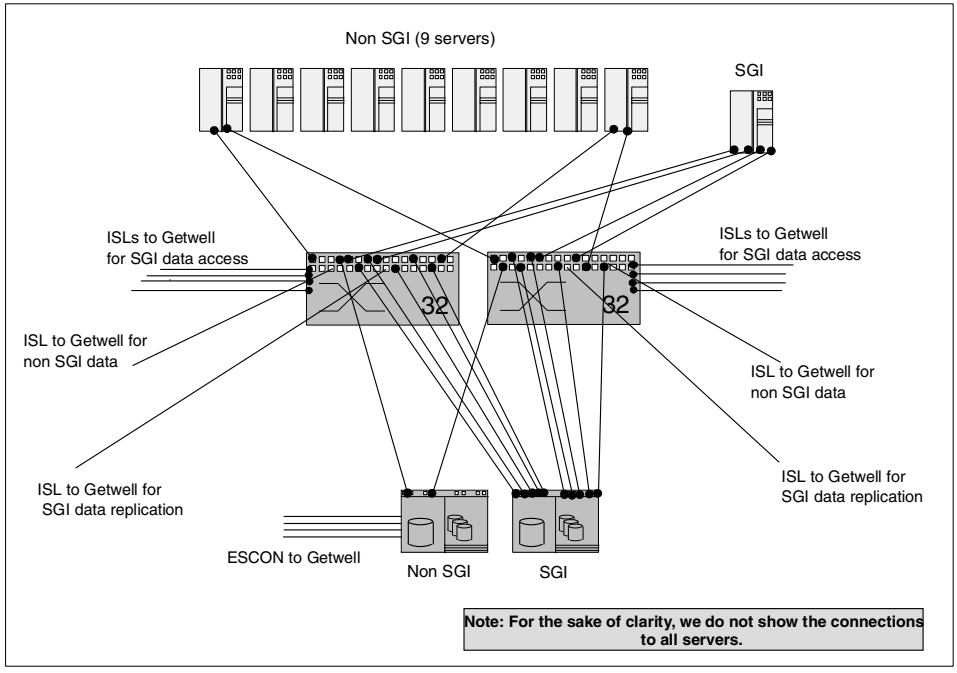


Figure 8-9 Feelinbad SAN design using McDATA ES-3032 switches

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 ESCON connections from the 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 46 Fibre Channel connections.

Table 8-8 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.

8.2.2 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.

In our example, the whole implementation can be done using 1 Gb/s technology. Port consolidation may be possible with 2 Gb/s technology when supported by IBM. This will require port replacement and a firmware upgrade.

8.2.3 Availability

The McDATA ED-6064 Director solution provides 99.999% availability. This number will 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.

The McDATA ES-3032 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.

8.2.4 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 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. This will be implemented using software zoning.

  Note: There are no data integrity issues if you do not implement zoning. In our example, we only use zoning for performance reasons.

8.2.5 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 GBICs 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. E_Port buffer credits can be set to 60 (maximum) per port on both switches and Directors used for long distance ISLs. We go into more detail about buffers in 2.12.9, “Buffers and credits” on page 114.

8.2.6 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.

8.2.7 “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 GBIC, 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 GBIC.
- 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.

8.2.8 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 8-10.

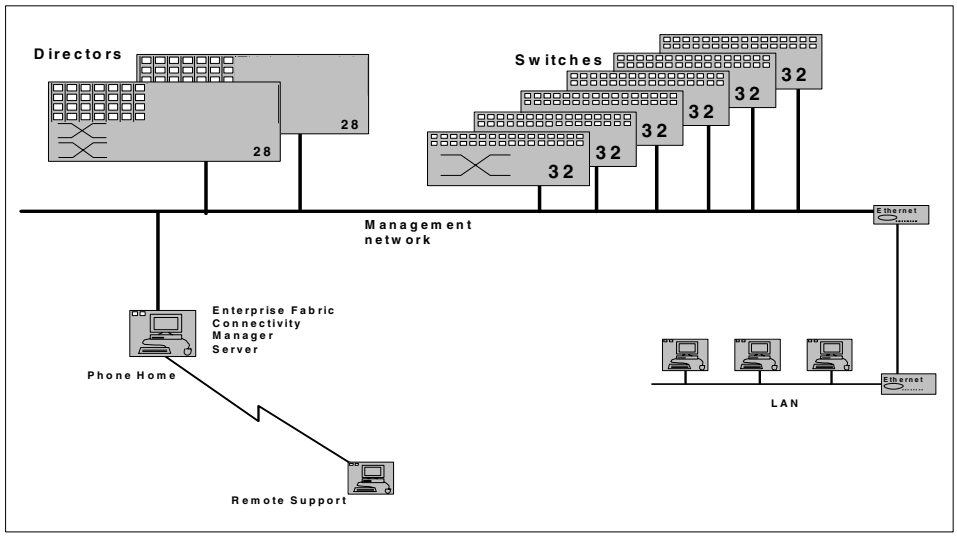


Figure 8-10 Management network

Note: This network can be part of your existing network infrastructure.

Full functional management is available through EFC manager software.

8.3 Case Study 3: Company Three

If we consider the company and its requirements, as detailed in 5.3, “Case Study 3: Company Three” on page 299, we will propose the following solution.

8.3.1 Design

Considering the requirements as detailed in 5.3.5, “Analysis (ports and throughput)” on page 301, we have proposed in this section a switch based solution for our design.

We show our proposed design in Figure 8-11.

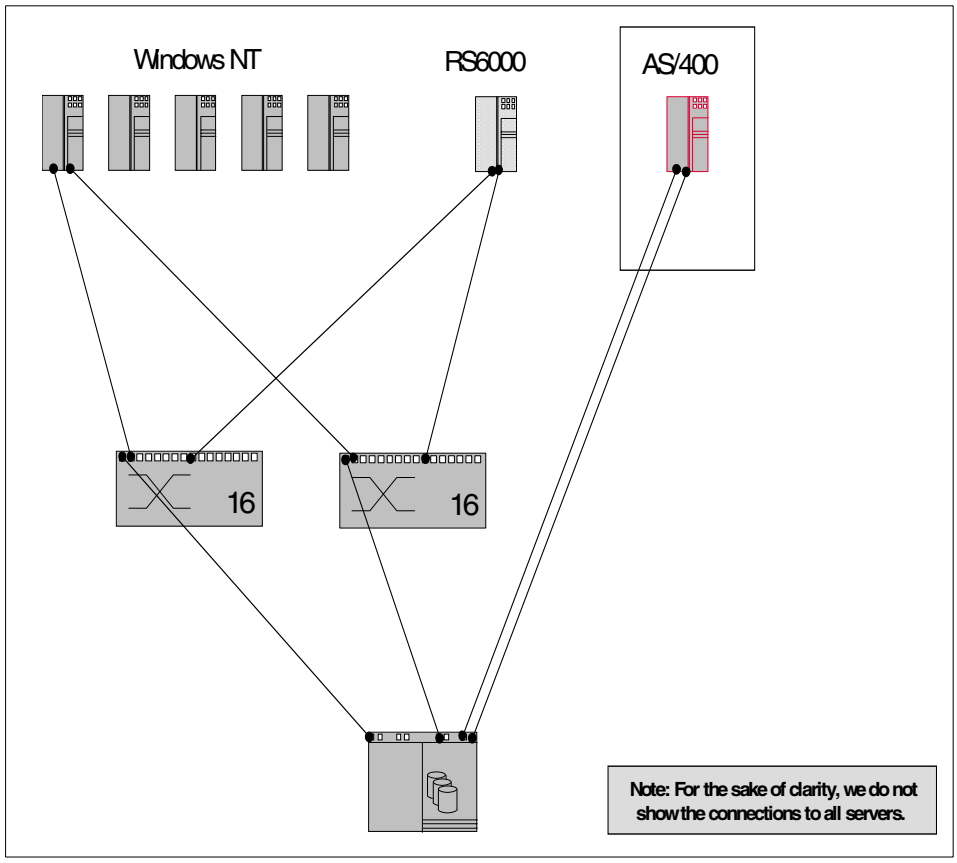


Figure 8-11 Initial design using McDATA ES-3016 Switches

It is important to note that the McDATA solutions do not support AS/400 as of the writing of this book. The AS/400 requires private loop functions that cannot be performed using the McDATA ES-1000 loop switch.

Note: The AS/400 platform switched fabric support is expected in the future, and we will introduce this system to the SAN when support is available. For now, we will leave the AS/400 out of the SAN configuration, but will directly connect it to the consolidated storage device using FC.

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-SDD) on each of the servers. 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 8-9 we summarize the port usage.

Table 8-9 Port usage

Storage	Server	ISL	Spare
2	12	0	18

In the following sections, we outline some of the aspects of this solution.

8.3.2 Performance

In the design we have a fan-out ratio of 6:1 for all servers. Latency through the switches is negligible (~2 micro seconds).

8.3.3 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 ES-3016 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.

8.3.4 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**

  As we have only one storage port per switch for all servers, there is no need to implement any zoning.

8.3.5 Distance

In all the connections we will use shortwave optics as the servers are within a radius of 500 m. It is possible to move some server further away by using longwave optics for ISLs.

8.3.6 Scalability

As you can see from the designs, we accommodated three years growth. The growth can be achieved without any interruptions in the production environment.

8.3.7 “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 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 ES-3016 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.

8.3.8 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 8-4.

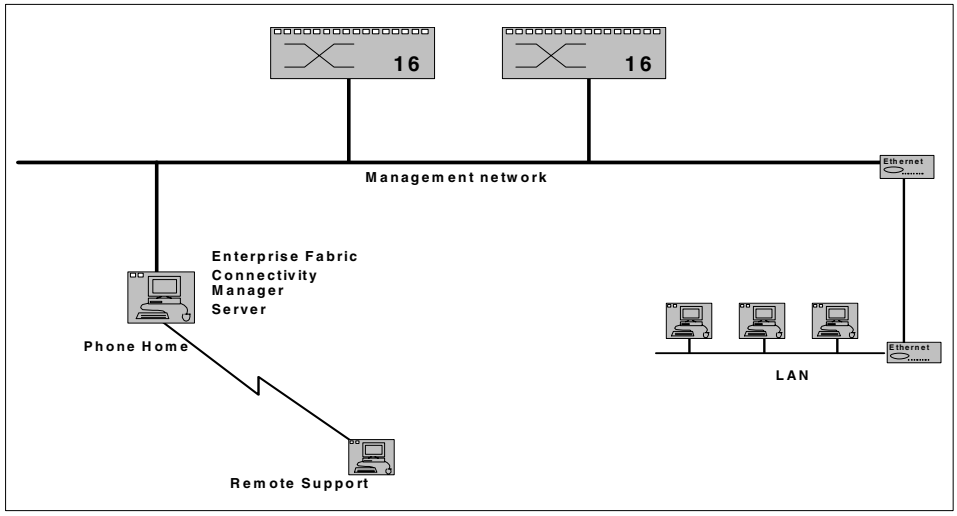


Figure 8-12 Management network

Note: This network can be part of your existing network infrastructure.

Full functional management is available through EFC manager software.

8.4 Case Study 4 - Company Four

If we consider the company and its requirements as detailed in 5.4, “Case Study 4: Company Four” on page 303, we will propose the following solution.

8.4.1 Design

Considering the requirements as detailed in 5.4.5, “Analysis (ports and throughput)” on page 306, we have proposed in this section a switch based solution for our design.

Open systems

Both the open systems and OS/390 platforms will share the same storage device. The following section refers to the open system SAN infrastructure and we will discuss the OS/390 platform separately.

We show the proposed open systems design in Figure 8-13.

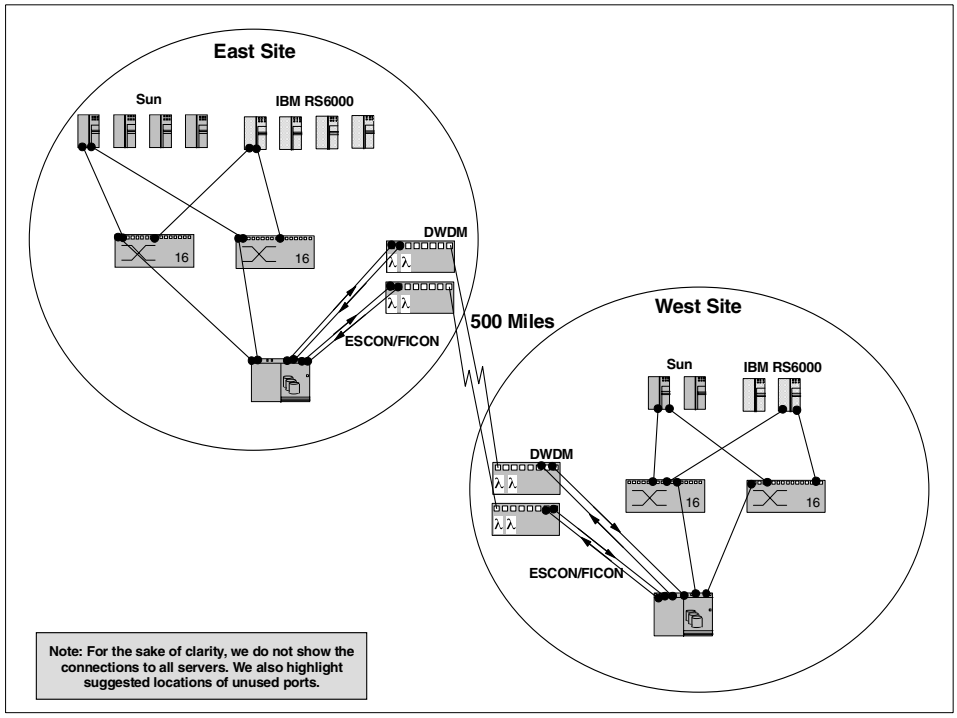


Figure 8-13 Initial design using McDATA ES-3016 Switches for open systems

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. 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 FICON. 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 8-10 we show the ports at the East site.

Table 8-10 Ports: East site

Storage	Server	ISL	Spare
2	16	0	14

In Table 8-11 we show the ports at the East site.

Table 8-11 Ports: West Site

Storage	Server	ISL	Spare
2	8	0	22

8.4.2 Performance

In the design, we have a fan-out ratio of 6:1 for all servers in the East site. The fan-out ratio in the West site is 4:1. Latency through the switches is negligible (~2 micro seconds).

8.4.3 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 ES-3016 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.

8.4.4 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**

  As we have only one storage port per switch for all servers, there is no need to implement any zoning.

8.4.5 Distance

In all the connections, we will use shortwave GBICs as the servers are within a radius of 500 m. It is possible to move some servers further away by using longwave GBICs for ISLs.

8.4.6 Scalability

The design accommodates three years growth. The growth can be achieved without any interruptions in the production environment.

8.4.7 “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 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 ES-3016 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.

- DWDM device: Should the DWDM device fail, the alternate one will be used.

8.4.8 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 8-14.

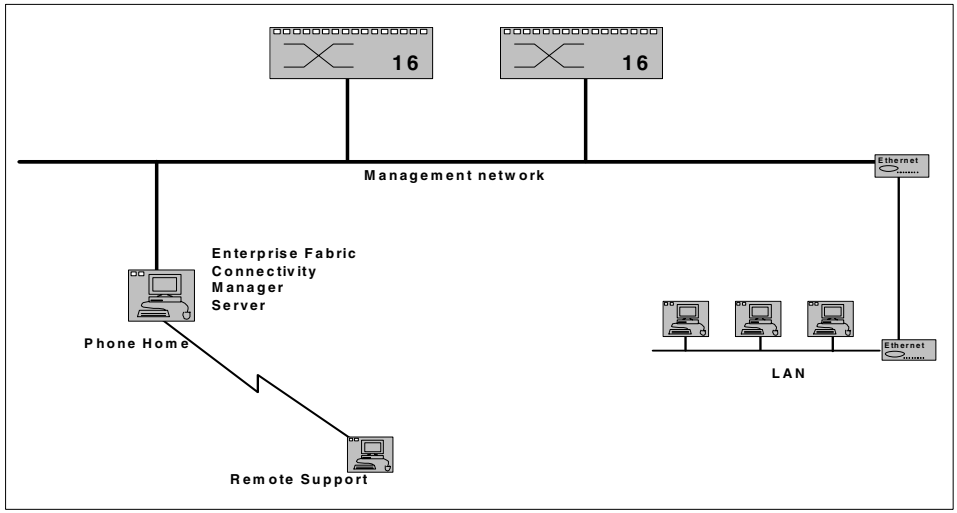


Figure 8-14 Management network

Note: This network can be part of your existing network infrastructure.

Full functional management is available through EFC Manager software.

OS/390

The design of the OS/390 systems is depicted in Figure 8-15.

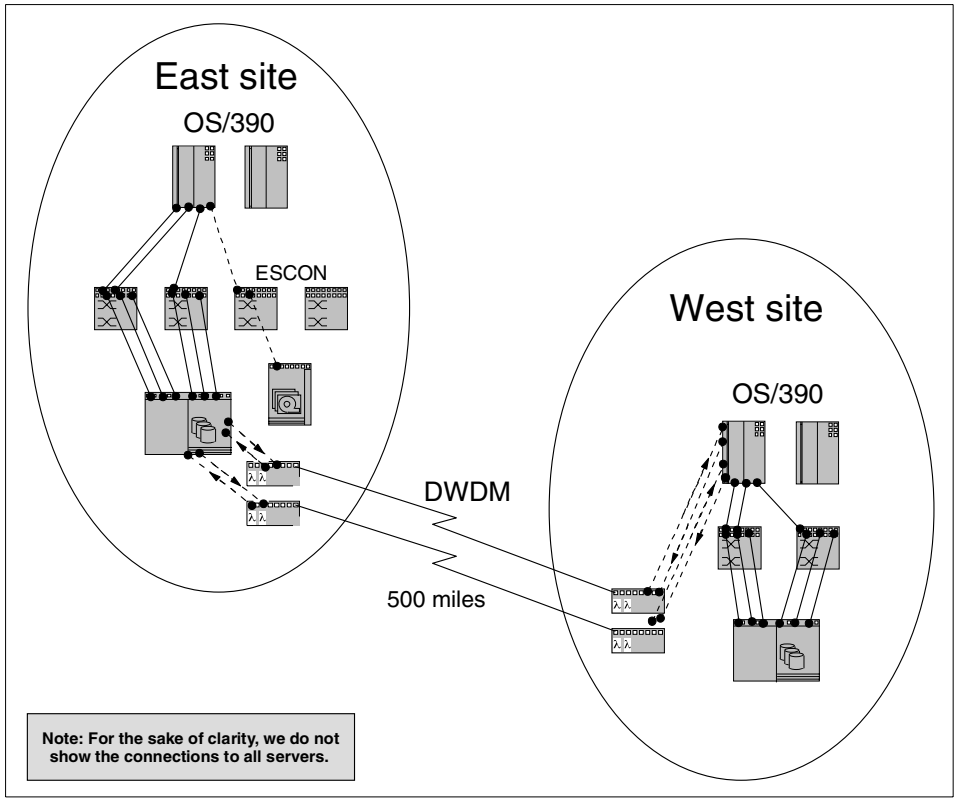


Figure 8-15 OS/390 design

8.5 Case Study 5 - Company Five

If we consider the company and its requirements as detailed in 5.5, “Case Study 5: Company Five” on page 308, we will propose the following solution.

8.5.1 Design

Considering the requirements as detailed in 5.5.5, “Analysis (ports and throughput)” on page 311, we selected a solution based around the McDATA ES-3016 and McDATA ES-1000.

The proposed design is shown in Figure 8-16.

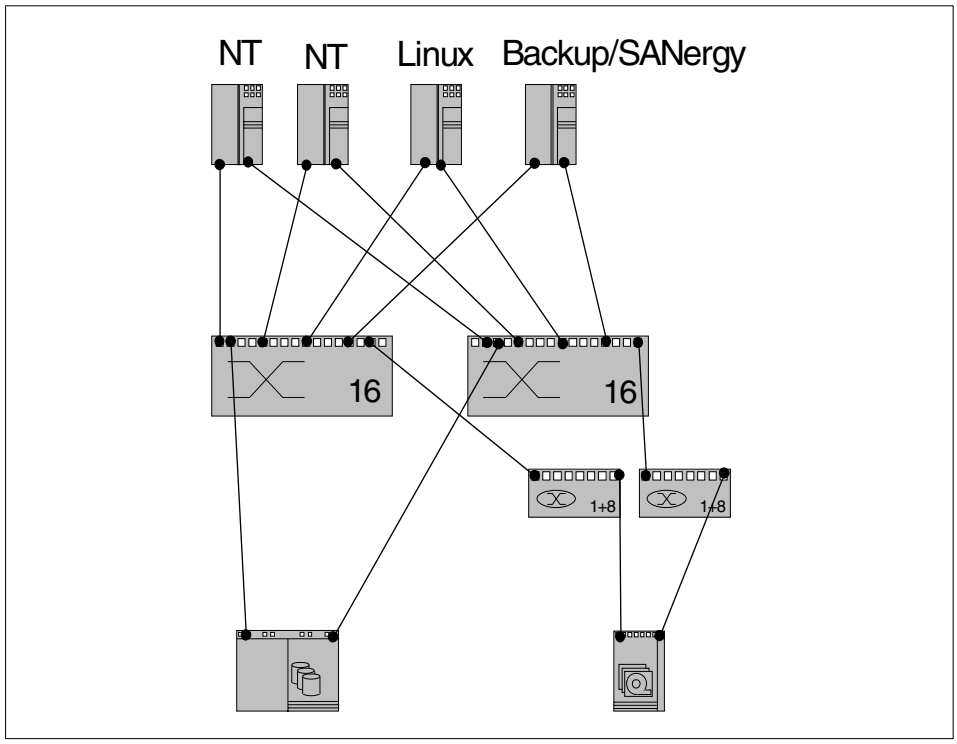


Figure 8-16 The proposed design for Company Five

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. With this design we are partially fulfilling the requirements:

- Redundant SAN with two switches and two loop switches. All servers are dual connected to the switches. Each tape drive is connected to a separate loop switch.
- 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 8-12 we show the initial ports used.

Table 8-12 Initial ports

Storage	Server	ISL	Spare
4	8	0	20

Note: No ISLs have been counted even though we have connections to the ES-1000 Loops switch devices. As these devices act as a bridge to FC-AL, we have just increased the storage port count.

In the following sections, we outline some of the aspects of this solution.

8.5.2 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.

As the McDATA switch does not support FC-AL, which is required for accessing the tape drives, we have introduced McDATA ES-1000 loop switches. This will mean that all tape traffic will experience a hop within the SAN, but latency for this should be under 3 microseconds.

8.5.3 Availability

All servers and the storage have multiple paths giving resilience through redundancy.

The McDATA ES-3016 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.

8.5.4 Security

When implementing the devices, we need to address some security issues:

- **Switch and loop switch security**
  - Change the default passwords.
  - Either put the switches and loop switches into an existing separate management network if one is already in place for other functions, or implement a new network.
- **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.

8.5.5 Distance

All devices (servers, switches, and storage) will use shortwave optics and cables as they are within a radius of 500 m.

8.5.6 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 ES-3016 can accommodate another 10 servers. If we attach additional tape drives connections using the ES-1000s, this will not require additional ports on the ES-3016.

8.5.7 “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: 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 Port: If one of the ports fails, you may replace it using a hot-pluggable optic, without outage or performance problem.
- Loop Switch Port: If one of the ports fails, you may replace using a hot-pluggable optic, without outage or performance problem.
- Switch: If a switch fails the alternate switch will accommodate the workload.
- Loop Switch: If a loop switch fails the alternate loop switch will allow access to another tape drive.

8.5.8 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 8-17.

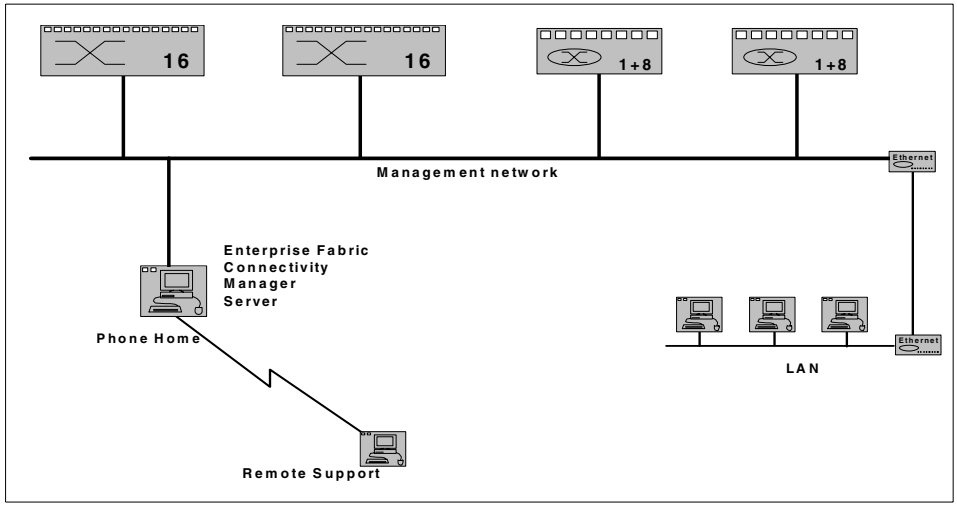


Figure 8-17 Management network

Note: This network can be part of your existing network infrastructure.

Full functional management is available through EFC Manager software.

8.6 Case Study 6: Company Six

If we consider the company and its requirements, as detailed in 5.6, “Case Study 6: Company Six” on page 313, we will propose the following solution.

8.6.1 Design

Considering the requirements as detailed in 5.6.5, “Analysis (ports and throughput)” on page 315, we selected a solution based around the McDATA ES-3016 and the IBM SAN Data Gateway SCSI Tape Router.

We show the proposed design for Primary site in Figure 8-18.

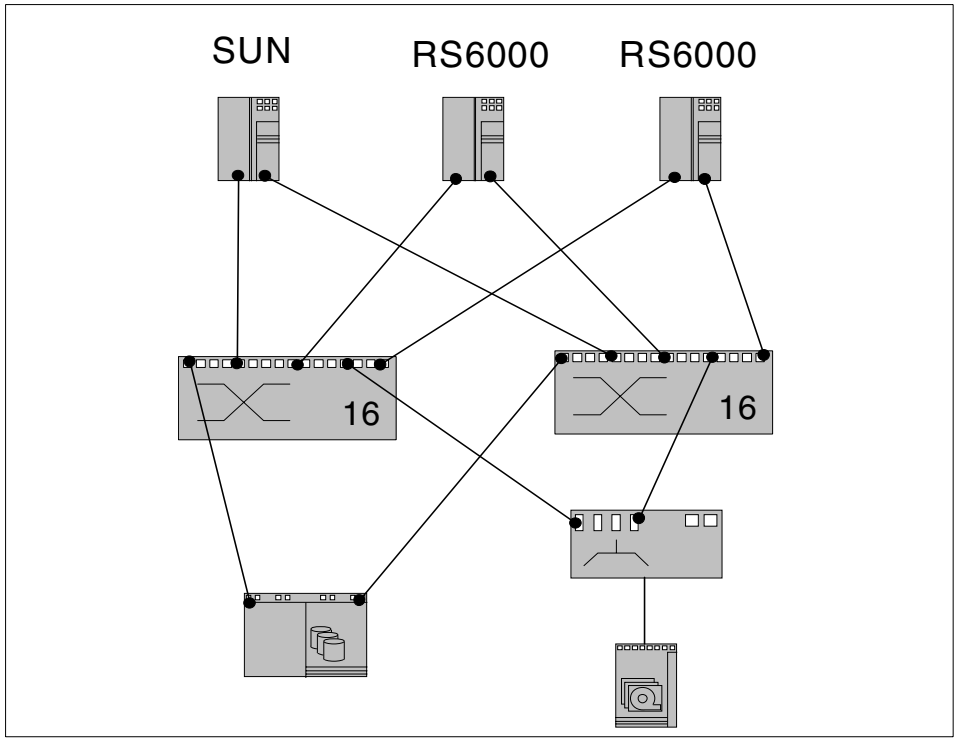


Figure 8-18 The proposed design for the Primary site

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 8-19.

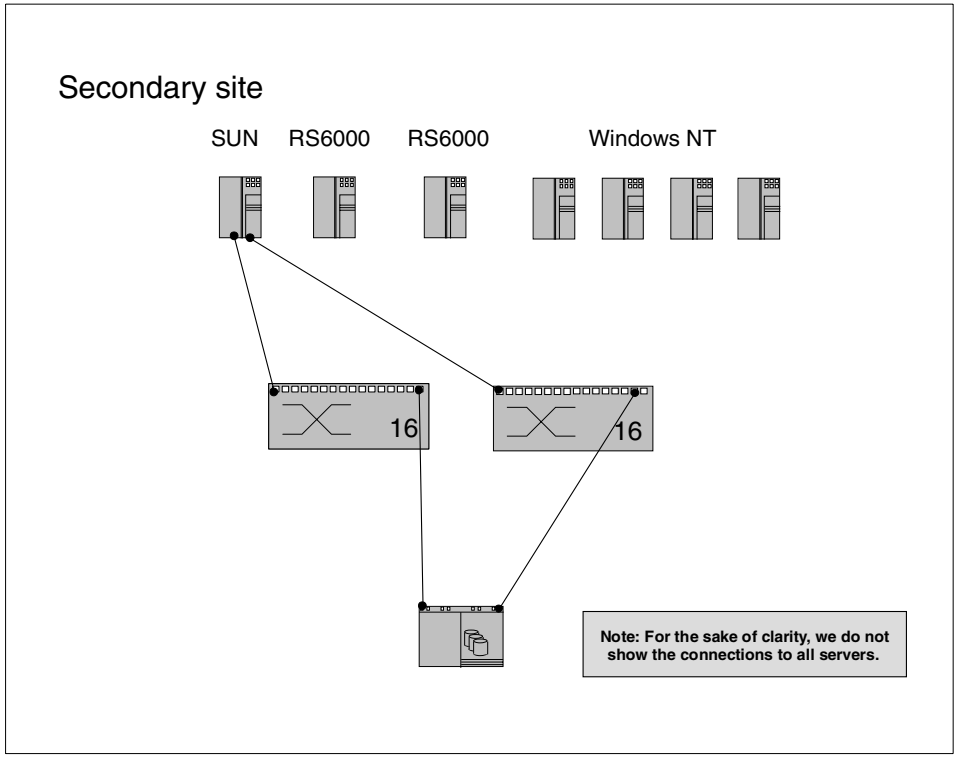


Figure 8-19 The proposed design for the Secondary site

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 8-20.

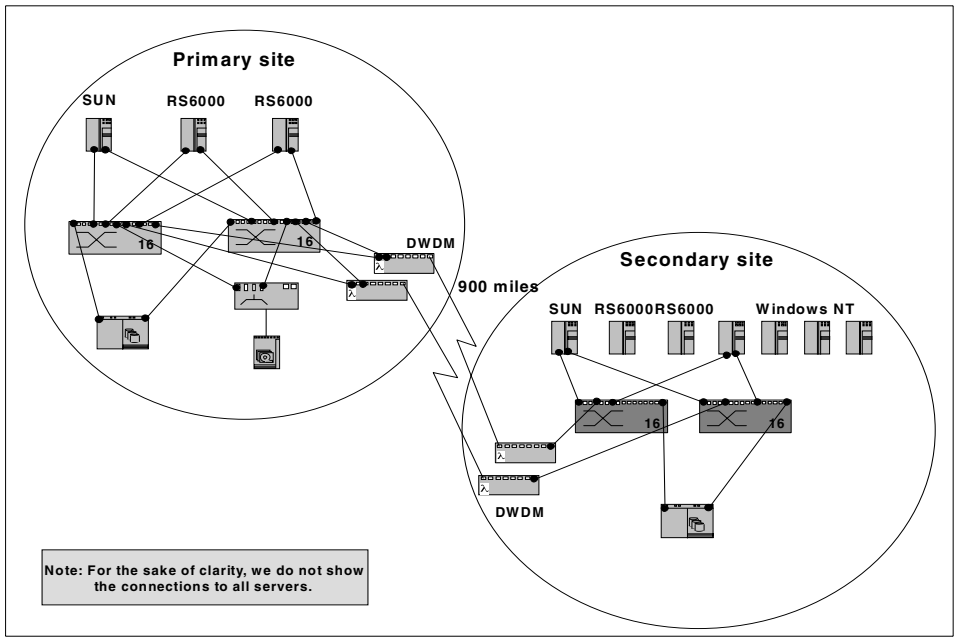


Figure 8-20 The complete solution

Note: To the best of our knowledge, this solution (with the McDATA ES-3016 and the Compaq) has not been tested, and is therefore, not certified.

With such a design, we are fulfilling these 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.
- We are re-using 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.

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 8-13 we show the ports at the Local site.

Table 8-13 Ports at Local site

Storage	Server	ISL	Spare
4	6	2	20

Note: We have not counted the connections to the FC/SCSI bridge as ISLs, they are accounted for in the storage port count.

There are ample ports left at the Remote site, the final port allocation for the remote site is represented in Table 8-14.

Table 8-14 Ports at Remote site

Storage	Server	ISL	Spare
2	14	2	14

In the following sections, we will outline some aspects of the design.

8.6.2 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.

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.

8.6.3 Availability

All servers and the storage have multiple paths giving resilience through redundancy.

The McDATA ES-3016 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 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.

8.6.4 Security

When implementing the switches, we need to address some security issues:

- **Switch 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 only use zoning for performance reasons.

8.6.5 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 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.

We do need to adjust the McDATA ES-3016 switch to assign 60 BB_Credits (which is the maximum) to the required ports.

We explain buffering in 2.12.9, “Buffers and credits” on page 114.

8.6.6 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 switches.

8.6.7 “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: The ES-3016 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.

8.6.8 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 8-4.

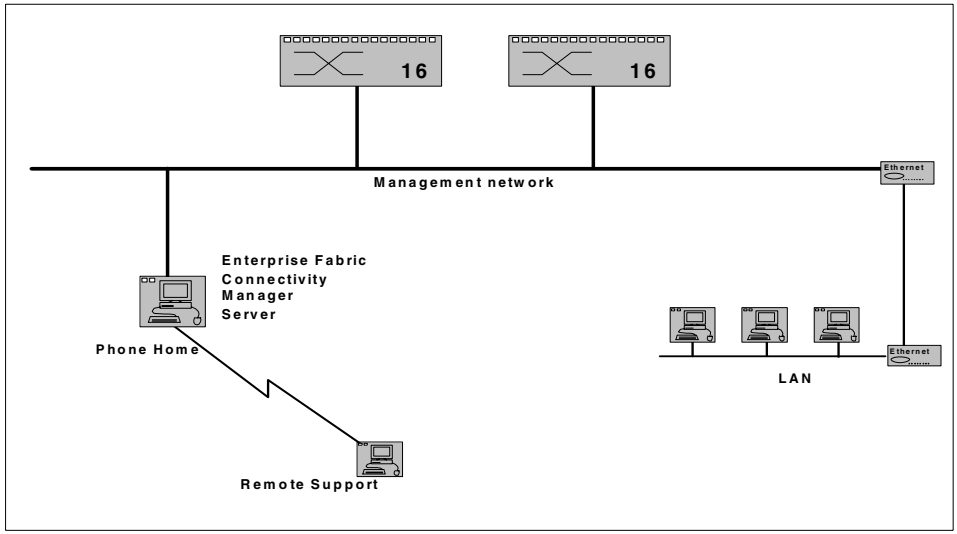


Figure 8-21 Management network

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.

9

IBM TotalStorage SAN Switch 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.

9.1 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: edge and core
- Ports
- Space
- Cables
- Bandwidth

9.1.1 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.

9.1.2 How to avoid downtime

We recommend that you use dual connections from each server into two different switches/core switches. 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, core switches 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 GBICs and cables. These components are the most susceptible to failure. Use dust caps on these components whenever possible.

9.1.3 Adding a switch

When adding a switch or core switch 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 core switch.

Backup your configuration before you perform any changes. You should also backup all your other definitions: zones, QuickLoops, alias definitions, configurations, and static routes.

Before adding a new switch to your SAN ensure that you have performed these actions on the *new* switch:

- 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 then check with the manufacturer to ensure that it is compatible, and that it can interoperate with your existing equipment.

9.1.4 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. Refer to the guidelines mentioned in 2.2.2 “Topologies” on page 58, for optimal use.

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.

9.1.5 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 9.5 “Tools” on page 446.

9.2 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 9-1.

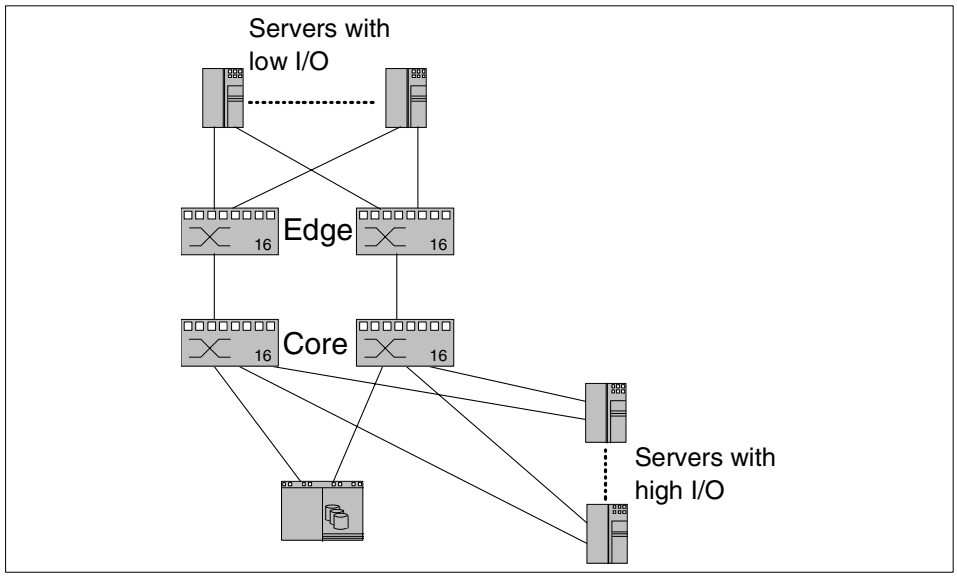


Figure 9-1 Connecting high I/O servers to the core switches

- 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.

9.3 Port placement

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 if possible.
- Group the devices which communicate a lot to each other on the same ASIC and blade inside the switch.
- Separate two connections from the same device (server or storage) to two switches 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 (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.

9.4 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.

9.5 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.

The IBM TotalStorage SAN Switch also supports setting up log forwarding to syslog daemon.

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 9.6 “Documentation” on page 448.

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.

IBM TotalStorage SAN Switches do not support dial out or dial in functions. To provide some kind of alarm for problems in the fabric the best practice would be to use SNMP alerts to be routed to the central management server like Tivoli Enterprise Console or IBM Director. This tool allows you setup notification to alert support personnel. To perform remote operations on the switches, you should setup a dial in server, which will be connected to the same *network* as the switches. With such a setup, remote operations will be possible on the switches.

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.

9.6 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

9.7 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.

The data which should be backed up at a minimum includes:

- Alias definitions
- Zone definitions (hard/soft)
- Static routes if implemented
- Trunking definitions if implemented
- Frame filtering definitions if implemented
- User IDs/passwords
- IPs (Ethernet/Fibre channel)
- Serial port setup

9.8 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. Refer to 2.7 “Fabric Shortest Path First” on page 78, for more information on how ISLs can affect your SAN.

9.9 Zoning

Use zones to:

- Separate one platform from another
- Separate high bandwidth devices (ports) from low bandwidth devices

Hard zones are based on the port number on the switch. If there are multiple switches, the ports are identified by the switch domain ID and the port number on the switch. You would select a hard zone for:

- Ease of management: If there is a SAN port failure in the server or a storage device, you do not need to change the zoning.

  Attention: If the port on the switch itself fails, and you are using hard zoning, unless you have a spare port in the hard zone, then reconfiguration of the hard zone may become necessary to include the port you substitute.

- If you do not care about security from the standpoint of which server can logically access the SAN resources. If someone has access to the switch they can plug-in with another server into the fabric and gain access to the resources.

- Hard zones are done at the level of ASIC.
- It is easy to replicate the zone information to the same type of fabric in the SAN. For example, if you are building a redundant SAN, you would define the zone just once and then replicate it to the other part of the SAN.

Note: All ISLs are excluded from zoning.

Soft zones are based on the WWN of the devices. You would select soft zoning for:

- Higher security from a *logical* view. Only predefined WWNs can participate in the zone.
- More complex management. In a SAN port fails in the server or storage device you need to update the zone information with the new WWN.
- You cannot separate traffic across different ISLs.
- Soft zones are done at the level of the Simple Name Server (SNS). This means that when the server queries the SNS, it will only be returned the devices in that zone. If the device is directly accessing other devices, bypassing SNS, you should use hard zones.

Note: Zones should always be used in combination with LUN masking on the storage devices.

- It is highly recommended that you make backup copies of your zone definitions.

Note: When you are adding a new switch, its configuration must be cleared or you will get a conflict in zoning.

- For ease of management, we recommend the use of aliases in zones. So, in case of replacing the port in the server or storage device, you just change the alias and not all associated zones.
- If you use soft zones, ISLs are shared across all zones.
- When you make any changes to the zones, document them for the future.

Note: Whenever you have problems accessing devices across the SAN, you should check your zone definitions first.

9.10 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. When the fabric is first configured the first switch in the fabric will become the principal switch 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 you are in the same situation as if the switch 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.

When you power on the switches, wait for indication that the power on sequence has successfully finished. For example, after the power on sequence is finished you will see green light on the ports.

9.11 Going from 1 Gb/s to 2 Gb/s

The F16 is 2 Gb/s capable, but can work at 1 Gb/s speed also. So, if you are setting up a new SAN you should consider implementing only 2 Gb/s technology. If you already have 1 Gb/s SAN component, you can easily replace them with 2 Gb/s models.

10

INRANGE best practices

In this chapter we consider that the SAN may already be implemented and offer some recommendations about the best ways to manage it.

10.1 Scaling

During our case study examples we have planned for additional growth. There comes a time though when a SAN needs to be extended beyond the limits of its initial design.

We recommend that you perform an inventory of all HBAs, switches, storage and ports to determine, exactly, the current situation. Then apply what is known with regards to anticipated growth and new projects to determine the requirements for the next 18 months. When this information has been gathered, we recommend that a review should take place: Is the current SAN topology the most efficient, or is a complete redesign called for?

Ensure that you plan for growth in the following areas:

- Switches/Directors
- Ports
- Space
- Cables
- Bandwidth

10.1.1 How to scale easily

We recommend that redundant fabrics should be used, wherever possible, to remove single points-of-failure and to allow for devices to be taken down for service when appropriate.

We recommend that there is a clear definition of who is responsible for performing what actions when it comes to changing the SAN fabric environment. There should also be communication with all parties concerned as to when a change is going to occur. Also, document the activities within the change management process. This will be a record that can be referred back to at a later date. We recommend that clear and standard procedures should be established for each person who must perform functions that will change the fabric.

In short, produce a thorough project and process change plan.

10.1.2 How to avoid downtime

We recommend dual connections from each server into two different 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 storage devices, providing minimal bandwidth in the case of upgrades.

Ensure that all switches, Directors and storage devices have redundant power supplies and these power supplies have different power feeds. All INRANGE products have dual power supplies as standard, but other vendors may not.

We recommend that there should be a supply of spare GBICs (optics) and cables available on site. These components are the most susceptible to failure. Use dust caps on these components whenever they are being stored.

10.1.3 Adding a switch/Director

When adding a switch or Director to the SAN fabric, it is imperative to plan for its addition. Know what the end result will look like.

Consider using SAN certified professionals to perform the installation and integration. From their experience, they can establish and manage the project plan and may be less prone to causing an unplanned outage.

The process also needs considerable planning. Changes should be performed one port at a time, wherever possible, to minimize the duration of single points-of-failure.

When moving cables, one cable should be moved at a time. By doing this, we can be sure that only redundant cables are being taken out of service. That is, at no point will both connections to the same node be removed.

The configuration should be backed up before you perform any changes.

Before a new switch or Director is added to the SAN, these actions must be performed:

- 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 a different model of switch or Director is being added then it is essential to check with the manufacturer to ensure that it is compatible and that it can interoperate with the existing equipment.

10.1.4 Adding ISLs

The most important aspect to consider when adding ISLs is whether or not they are being used cost effectively. Refer to the guidelines mentioned in 2.2.2 “Topologies” on page 58, for optimal use.

We recommend that the use of new SAN functions should be considered. For example, *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. For a more detailed discussion on trunking see 2.9 “Trunking” on page 97.

10.1.5 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 10.5 “Tools” on page 458.

10.2 Know your workloads

Before designing the SAN be aware of the following:

- Collect the I/O bandwidth required for all applications.
- Design the SAN for the peak, not the average traffic.
- Group the servers with high demand directly with storage devices and separate them from other workloads, especially when there are ISLs. In a

core-edge design, this would imply connecting servers with high demand to the core switches.

We show an example of such a setup in Figure 10-1.

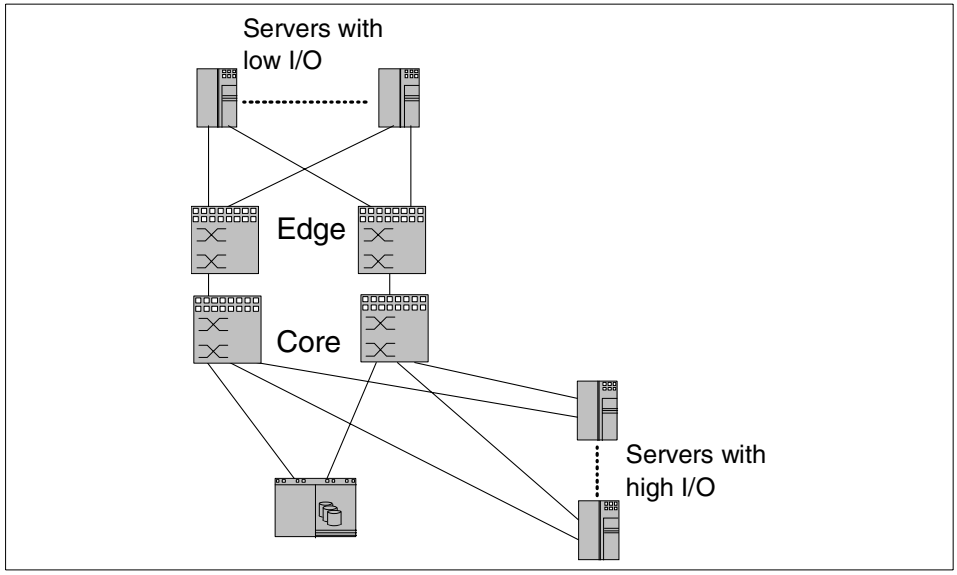


Figure 10-1 Connecting high I/O servers to the core switches

- Constantly monitor the utilization to see if the real world matches the design expectations.
- Increase or decrease the number of servers ports or ISLs if this will solve performance problems.

10.3 Port placement

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.
  - INRANGE have separate buffer pools for each port.
  - All ports on the INRANGE 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 INRANGE 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 10-2.

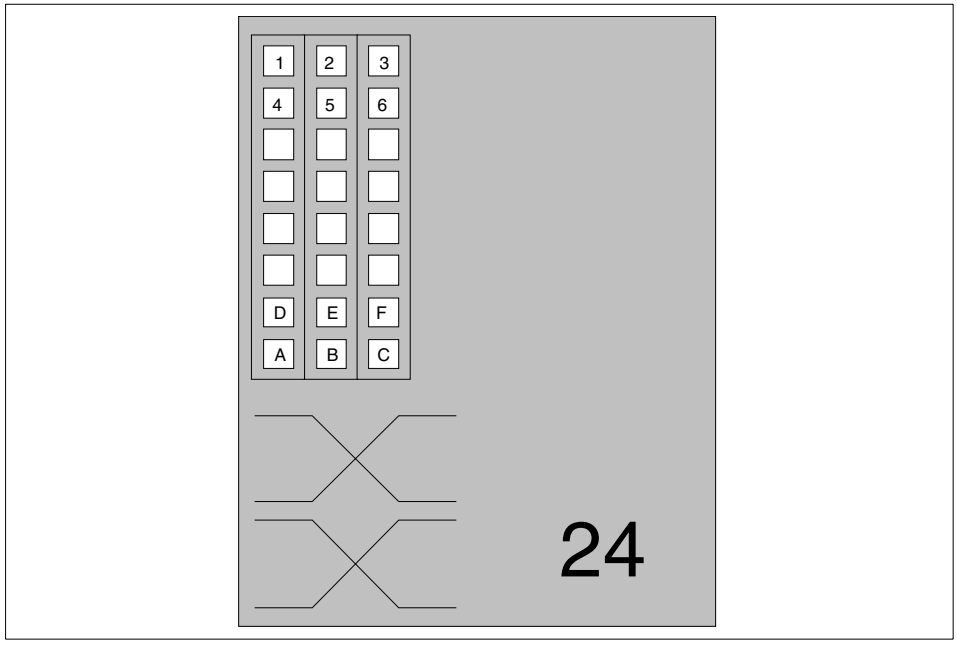


Figure 10-2 The INRANGE port population guidelines

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 INRANGE 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.

10.4 WWNs

It is important to keep track of the WWNs used in the SAN. For example, creating *and maintaining* 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 aliases should be assigned to the WWN used. Later on when introducing zoning, those aliases can be used instead of the WWNs. Some devices (e.g. IBM ESS) have the 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.

10.5 Tools

Typical features that the GUI tools from all manufacturers provide 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 complement 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 from all the switch and Director manufacturers. They tend to include standard Fibre Channel Alliance MIBs support along with their custom private MIB.

Scripting can be interpreted 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, and this is covered further in 10.6 “Documentation” on page 461.

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 form of logs. There may be more than one log on each switch/Director 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.

INRANGE Directors provide a *Phone Home* facility. This will allow the Directors to contact an INRANGE service facility and report problems. An example being the failure of a redundant power supply. This means that the unit can be brought back up to a fully resilient configuration without delay. If the security policy at the site allows for this facility, then it is strongly recommended that it is implemented.

Additionally, INRANGE’s management system, IN-VSN Enterprise Manager, allows the administrator the ability to provide alarms for problems in the fabric. The best practice would be to allow SNMP alerts to be routed to a central management server like Tivoli Enterprise Console or IBM Director. This tool allows notification to support personnel to be set up. To perform remote operations on the switches there should be a dial in server, which will be connected to the same network as the Directors.

Beside using the built in features and management tools from the vendors there are also other tools which give 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/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.

10.6 Documentation

The following information and documentation should be kept to hand for reference:

- Configuration diagrams
- Logical configuration
- Physical setup of the SAN including the cabling infrastructure
- Phone numbers of the devices which support dial-in. This will prevent unnecessary delays when a support organization needs access for maintenance or fault finding
- Levels of firmware that are loaded in the components used in the SAN
- Serial numbers in case the components have to be replaced or perhaps, simply identified by it
- Phone number of the support organization that has to be called in case of problems and the procedures for reporting problems
- Access information, IP addresses, user IDs, passwords, keys for people who perform any kind of operations in the SAN
- Change management should be implemented, so that all the changes (for example, reallocation or re-cabling) are documented and accessible for someone who was not involved in the original change process
- All the documentation from the supplier of the equipment has to be accessible if needed

10.7 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 easy reused by setup tools in case the SAN components are replaced.

Attention: A disaster recovery plan should be designed, documented, and tested.

It is also recommended that you backup the configurations after each change, and that you keep the history of changes so you can easy recover from mistakes.

The data which should be backed up should include at a minimum:

- Alias definitions
- Zone definitions (hard/soft)
- Static routes if implemented
- Trunking definitions if implemented
- Frame filtering definitions if implemented
- User IDs/passwords
- IP addresses (Ethernet/Fibre Channel)
- Serial port setup

10.8 Common practice faults

Here are some of the most common mistakes made in SAN setup and management:

- Bad cabling practice: Try to mark all the cables and have a clear picture of how the devices are 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
- Not clearing the configuration of a new or additional switch/Director. Not doing so can cause the entire SAN to fail.
- When zoning, forgetting to propagate zones to all switches/Director 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 general zone. Every SAN port must belong to at least one zone after zones are introduced. This step is often forgotten.
- If zones overlap, check them twice before activating them.
- Implementing temporary zones and then not reconfiguring back to the status quo. For example, if there is a special zone configuration for backups, do not forget to change them back to normal operations when they are not needed anymore.
- Making changes based on assumptions. For example, 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 even worse.

Refer to 2.7 “Fabric Shortest Path First” on page 78, for more information on how ISLs can affect your SAN.

10.9 Zoning

Use zones to:

- Separate one platform from another.
- Separate high bandwidth devices (ports) from low bandwidth devices.

Hard zones are based on the port number on the switch. If there are multiple switches, the ports are identified by the switch domain ID and the port number on the switch. Hard zoning should be selected for:

- Ease of management. In the event that a SAN port fails in a server or storage device, there is no need to change the zoning.

  Attention: If the port on the Director itself fails, and you are using hard zoning, unless you have a spare port in the hard zone, then reconfiguration of the hard zone may become necessary to include the port you substitute.

- To separate traffic across ISLs.
- If you do not care about security from the standpoint of which server can logically access the SAN resources. If someone has access to the switch/Director they can plug another server into the fabric and get access to the resources.
- The implementation of hard zones varies slightly form manufacturer to manufacturer.
- In INRANGE equipment:
  - It effectively partitions the Director.
  - Hard zones have multiples of four ports within them.
- It is easy to replicate the zone information to the same type of fabric in the SAN. For example, if you are building a redundant SAN, you would define the zone just once and then replicate it to the other part of the SAN.

Soft zones are based on the WWN of the devices. Soft zoning should be selected for:

- Higher security from a *logical* view. Only predefined WWNs can participate in the zone.
- More complex management. In the event that a SAN port fails in a server or storage device, you will need to update the zone information with new WWN.

- You cannot separate traffic across different ISLs.
- Soft zones are done on the level of Simple Name Server (SNS). This means that when a server queries SNS it will only get the devices in that zone. If the device is directly accessing other devices bypassing SNS you should use hard zones.

  Note: Zones should be always used in combination with LUN masking on the storage devices.

- It is recommended that you make backup copies of your zone definitions.

  Note: When adding a new switch or Director, its configuration must be cleared or there may be a conflict in zoning.

- For ease of management, the use of aliases in zones is recommended. So, in the case of replacing the port in the server or storage device, you just change the alias and not all the associated zones.
- If soft zones are used, ISLs are shared across all zones.
- When any changes are made to the zones, they must be documented for future reference.

Note: Whenever there are troubles accessing devices across a SAN, you should check the zone definition first.

10.10 Powering a SAN (up or down)

After the SAN is setup and all the configurations are defined, the domain IDs for each Director should be determined and documented. When the fabric is first configured the first Director in the fabric will became the principal with domain ID 1. All the others will become subordinates and they will receive their domain ID from the principal switch.

Note: When you power on the first switch in the fabric you should clear the configuration and get the default domain ID 1.

When you power down the switch/Director you are in the same situation as if the switch/Director has failed. If the domain ID used by the switch is the same as the original, then there is no need to worry about the power on and off procedures. But if you have replaced the switch with a new one, it should have its domain ID set to the same value as the one it is replacing.

Note: If the domain ID of the switch changes you should change all your hard zone definitions.

If the domain ID of the switches are setup manually the order of powering the Directors on does not matter. But, it is still recommended that you perform the power on sequence in the same order every time if it is possible.

When the principal switch/Director is powered down, the switch with the lowest WWN in the fabric will become the principal.

When you power on the switches, wait for an indication that the power on sequence has finished, before powering on devices which might try to log into it.

10.11 Going from 1 Gb/s to 2 Gb/s

The backplane in the INRANGE FC/9000 is 2 Gb/s compliant, that is, a Director could be fully populated with FIO blades with all ports running at 2 Gb/s without the backplane or ASICs blocking. There will need to be a careful plan, however, when incorporating 2 Gb/s technology in a Director, which was originally configured for 1 Gb/s:

First, the FIO blades will need to be replaced with blades that support 2 Gb/s ports. The 1Gb/s ports use GBICs, but the new 2 Gb/s ready blades use LC format connections into SFF transceivers. It will be possible for 1 Gb/s and 2 Gb/s blades to co-exist.

In order to use the existing cabling then one of the following will be needed:

- SC to LC converter cable
- Remove and splice LC connectors onto the cable ends.
- Run a new cable from any existing patch panel.

In order to support the 2 Gb/s blades, the FSWs will need to be upgraded too. In order to perform this, there will be some down time required. This will be of the order of minutes, and can be planned. If there are two fabrics, then one fabric can be upgraded and brought back up, and then the second fabric can be upgraded.

11

McDATA best practices

In this chapter 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.

11.1 Scaling

During our case study examples, we have planned for additional growth. When your SAN was implemented there was a strong possibility that it was not planned for 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

11.1.1 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 fabric of your SAN first without disturbing the other fabric. After you have performed the fabric upgrade, and traffic is restored, you can grow the second fabric 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.

11.1.2 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 fabric of the SAN. This will give you enough bandwidth should you have to disable one of the fabrics for upgrades and maintenance. However, bear in mind that McDATA switches and Directors can be upgraded concurrently.

McDATA Directors and switches have redundant power supplies. Ensure that all devices (including storage) have their alternate power supplies connected to different power feeds.

We recommend that you have a supply of spare GBICs and cables. These components are the most susceptible to failure. Use dust caps on these components whenever possible.

11.1.3 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 wherever 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 your configuration before you perform any changes. Brocade fabric OS offers you the ability to backup the configuration to the remote server. You should also backup all your other definitions: zones, quickloops, aliases definitions, configurations, static routes.

Before adding a new switch or Director to your SAN, ensure that you have performed these actions on the new switch or Director:

- Clear configuration, so you do not get conflicts.
- Firmware checks (compatibility) and hardware: Try to use the same firmware as on other existing switches if possible.

If you are adding a different model of switch or Director, then check with the manufacturer to ensure that it is compatible and that it can interoperate with your existing equipment.

11.1.4 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. Refer to the guidelines mentioned in 2.2.2 “Topologies” on page 58, for optimal use.

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.

11.1.5 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 11.5 “Tools” on page 471.

11.2 Know your workloads

Before designing your SAN, be aware of the following:

- Collect the I/O bandwidth required for your applications.
- Design your SAN for peak workloads, not for average traffic.
- Group the servers with high demand directly to the core switch devices and separate them from other workloads, especially when you have ISLs. In the core-edge design, we show an example of this in Figure 11-1.

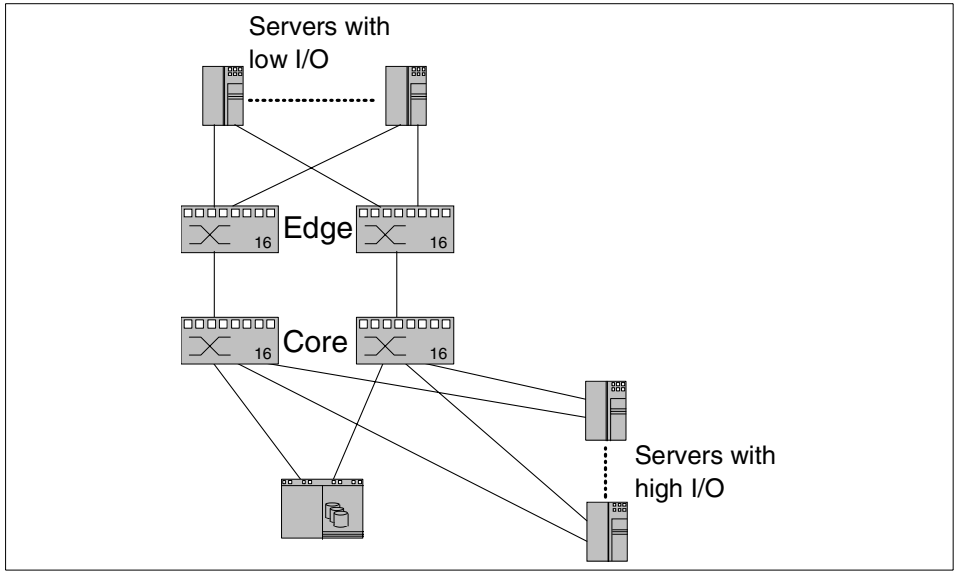


Figure 11-1 Connecting high I/O servers to the core switches

- Constantly monitor utilization to see if the real world utilization matches your design expectations.
- Adjust the fabric connections based on your experiences. You may have to increase the number of server ports or ISLs, if this will solve your performance problems, or you may be able to decrease connections and free up ports for future use.

11.3 Port placement

When connecting devices to the SAN, it is important to follow some rules:

- Put the devices (servers and storage) 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.
- Separate two connections from the same device (server or storage) to two switches/Directors if possible. Build redundant fabrics whenever it is possible.
- Try to avoid 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. For example, when connecting storage devices to several switches use Port 0 on the switch. If more than one storage device is attached, then chose another standard port to use that is not on the same ASIC. If the ASIC does fail, then you will still have access to at least one of the storage devices through this switch.

11.4 WWNs

You should keep track of the WWNs used in the SAN. For example, create and maintain 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 nicknames (also known as aliases) to the WWN used. Later on when you introduce zoning, you can use those aliases instead of the WWNs. Some of the devices (e.g. IBM ESS) have the 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.

11.5 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
- Status indicators (usually graphical representation of LEDs)
- Performance monitors (usually capturing instantaneous data but storing high and low 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 complement it 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 from all the switch and Director manufacturers. They tend to include standard Fibre Channel Alliance MIBs support along with their custom private MIB.

Scripting can be interpreted in a number of ways. If we consider scripting programs then McDATA offers telnet functions that can be scripted to execute commands and capture responses. Scripting can also imply having hardcopy procedures to follow, this is covered further in 11.6 “Documentation” on page 474.

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/Director 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/Director devices store forms of logs which can be accessed through the EFC manager console. There may be more than one log on each switch/Director and these may become consolidated at the management server. Logs tend to be proprietary in format and will be separated in the McDATA solutions into:

- Audit Log: Provides a timestamped record of all configuration changes and where they were made.
- Event Log: Significant events are recorded here, such as hardware failures.
- Hardware Log: Stores information regarding the removal or insertion of FRUs.
- Link Incident Log: Stores conditions that have occurred on a fiber optic link.
- Threshold Alert Log: Stores timestamped notifications of threshold alerts. Changes and initial threshold settings are also written to the log.

All McDATA switches and Directors support phone home features with the use of the EFC Management server.

Additional maintenance data can be collected to assist in trouble shooting. This can be found under maintenance options from the EFC Management Server.

To assist in troubleshooting, when the fabric is implemented, reset all counters. An indication that a problem may be being encountered could be recognized if the error counters are rising rapidly.

To provide alarms for problems in the fabric the best practice would be to use SNMP alerts to be routed to a central management server like Tivoli Enterprise Console or IBM Director. These tools allow you to setup the notification to alert support personnel. To perform remote operations on the switches, you should set up a dial-in server, which will be connected to the same network as the switches. With such a setup, remote operations will be possible on the switches.

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 winning 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/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.

11.6 Documentation

You should keep the following information or documentation ready at your disposal:

- Configuration diagrams
- Logical configurations
- Physical setup of the SAN including the cabling infrastructure
- Phone numbers of devices which support a dial-in function, so this can be used when intervention is needed from the support organization
- Firmware levels of the SAN components
- Serial numbers, in case components have to be replaced or identified by them
- Phone number of the support staff to be called in case of problems and the procedures for reporting problems
- Access information, IP addresses, user IDs, passwords, keys for the people who perform any kind of operations in the SAN
- Change management should be used for all changes, so that all the changes (reallocation, re-cabling) are documented and accessible for someone who was not involved in the original change process
- All the documentation from the supplier of the equipment has to be accessible (in hardcopy also) if needed

11.7 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. They should also be saved in a form which can be easily reused by the setup tools in case the components are replaced.

It is also recommended that you backup the configurations after each change and that you keep a history of changes so you can easily recover from mistakes.

The McDATA solution offers two methods for backup. The first is to backup the NV-RAM from the CTP card to the EFC Server. This backup stores all data pertinent to that particular switch or Director.The other method is to use the Zip drive that is attached to the EFC Server for just such purposes, this backs up the entire EFC Manager data directory.

The data in the data Directory includes:

- Nicknames
- Call home settings
- Zone definitions/ zone set definitions
- Trunking definitions if implemented
- Frame filtering definitions if implemented
- User IDs/passwords/session options/e-mail recipients/SNMP recipients
- IPs (Ethernet/Fibre Channel)
- Serial port setup
- Each switch/Director configuration

> **Note:** To restore a configuration to a switch or Director, the switch or Director must be set to offline status through the EFC Manager Hardware view.

11.8 Common practice faults

Here are some of the most common mistakes made in SAN setup and management:

- Bad cabling practice: Try to label all your cables and have a clear picture of how the devices are 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/Director you should clear its configuration
- When you create the first zone, this does not imply that all the other devices outside this zone are now in the default zone. However, the default zone needs to be activated separately or included in the same zone set that is active with the zone you created. Every SAN port must belong to at least one zone after zones are introduced, even if it is the default zone
- If you overlap zones, check them twice before you activate them
- If you change zones based on policy. For example, 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 ISLs 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 could become even worse. Refer to 2.7 “Fabric Shortest Path First” on page 78, for more information on how ISLs can affect your SAN
- When you add a switch or Director ensure that you have previously assigned the master Director or switch as the principal. Otherwise, there is a chance that the new switch or Director assumes this responsibility, if it has a higher priority or lower WWN

11.9 Zoning

Soft zones are available from McDATA products and these are based on the WWN of the devices. You would select soft zoning for:

- Higher security from a *logical* view. Only predefined WWNs can participate in the zone.
- More complex management. In a case that the SAN port fails in the server or the storage device you need to update the zone information with the new WWN.
- You cannot separate traffic across different ISLs.
- Soft zones are done at the level of the Simple Name Server (SNS). This means that when the server queries the SNS it will only get the devices in that zone.

Note: Zones should always be used in combination with LUN masking on the storage devices.

- It is recommended that you make backup copies of your zone definitions. In the McDATA solutions, you would store the zones in a zone set and deactivate the old zone set and activate the new one. Each zone set can contain 1023 zones (not including the default zone) and there can be up to 64 zone sets stored per fabric.

Note: When you are adding a new switch or Director, it is recommended that you clear its zone definitions. Only if you are sure that the zone it contains is identical in name and member definitions and that the domain ID is different from the principal switch/Director should you connect them both without clearing. The fabrics will be segmented should the configuration not conform to the above rules.

- For ease of management it is recommended to use nicknames in zones. If you replace the port in the server or storage device you just update the nickname and not all the associated zones.
- ISLs are shared across all zones.

Note: Whenever you have trouble accessing devices across the SAN you should check your zone definitions first.

11.10 Powering a SAN (up or down)

After the SAN is setup and you have defined all your configurations you should collect the domain ID information for each switch or Director. When the fabric is first configured the switch in the fabric with the lowest WWN will became the principal switch with domain ID 1. The switch priority can also be changed through the EFC Management server to override the default lowest WWN selection process. All the others will become subordinate and they will request and receive their domain ID from the principal switch. Domain ID's can be manually set and can have values of 1 to 31.

When you power down the switch/Director you have the same situation as if the switch/Director has failed. However, perform an orderly shutdown by quiescing the ports and putting the switch or Director offline before powering off the device. This can be achieved through the hardware view of the EFC Management server. If the domain ID used by the switch is the same as the original, then you do not need to worry about power on and off procedures. But if you replaced the switch with a new one, you should manually setup the ID to be the same as the replaced switch.

If the domain ID of the switches are setup manually, the order of powering the switches on does not matter.

When the principal switch is powered down, the switch with the lowest WWN in the fabric will become the principal.

11.11 Going from 1 Gb/s to 2 Gb/s

The backplane in the McDATA 6064 is 2 Gb/s compliant, replacement of FRU optics (SFPs) and a firmware upgrade will be required to enable 2Gb/s throughput.

When implementing a new SAN or adding servers to an existing SAN, it is worthwhile considering purchasing 2 Gb/s HBAs, as these are typically autosensing and will not need replacing in the near future.

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.

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

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

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

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, 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.

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.

Channel A point-to-point link, the main task of which is to transport data from one point to another.

Channel I/O A form of I/O where request and response correlation is maintained through some form of source, destination and request identification.

CIFS Common Internet File System

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.

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.

Controller A component that attaches to the system topology through a channel semantic protocol that includes some form of request/response identification.

CRC Cyclic Redundancy Check - An error-correcting code used in Fibre Channel.

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.

dB Decibel - a ratio measurement distinguishing the percentage of signal attenuation between the input and output power. Attenuation (loss) is expressed as dB/km

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.

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.

E_Port Expansion Port - a port on a switch used to link multiple switches together into a Fibre Channel switch fabric.

ECL Emitter Coupled Logic - The type of transmitter used to drive copper media such as Twinax, Shielded Twisted Pair, or Coax.

Enterprise Network A geographically dispersed network under the auspices of one organization.

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.

ESCON Enterprise System Connection

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 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 Fiber 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 - AFibre Channel industry association that works to promote awareness and understanding of the Fibre Channel technology and its application and provides a means for implementors 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.

Fiber Optic Refers to the medium and the technology associated with the transmission of information along a glass or plastic wire or fiber.

Fibre Channel A technology for transmitting data between computer devices at a data rate of up to 4 Gb/s. It is especially suited for connecting computer servers to shared storage devices and for interconnecting storage controllers and drives.

FICON Fibre Connection - A next-generation I/O solution for IBM S/390 parallel enterprise server.

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

FSP Fibre Channel Service Protocol - The common FC-4 level protocol for all services, transparent to the fabric type or topology.

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 A mode of communications allowing either transmission or reception of frames at any point in time, but not both (other than link control frames which are always permitted).

Hardware The mechanical, magnetic and electronic components of a system, e.g., computers, telephone switches, terminals and the like.

HBA Host Bus Adapter

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

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.

In-Band Signaling This is signaling that is carried in the same channel as the information. Also referred to as inband.

Information Unit A unit of information defined by an FC-4 mapping. Information Units are transferred as a Fibre Channel Sequence.

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.

I/O input/output

IP Internet Protocol

IPI Intelligent Peripheral Interface

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.

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 See Local Area Network - 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.

Link A connection between two Fibre Channel ports consisting of a transmit fibre and a receive fibre.

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 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.

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.

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.

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) bits per second.

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.

NDMP Network Data Management Protocol

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.

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.

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 Disk A storage device that is written and read by laser light.

Optical Fiber A medium and the technology associated with the transmission of information as light pulses along a glass or plastic wire or 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.

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 outband.

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.

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.

Port The hardware entity within a node that performs data communications over the Fibre Channel.

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.

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.

Protocol A data transmission convention encompassing timing, control, formatting and data representation.

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.

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.

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.

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	Bus-Width (bits)	Max. DTR (MB/s)	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
Ultra 2 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

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

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.

ST Straight Tip connector. Also known as “Stick and Twist”.

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.

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.

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.

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

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.

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.

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

For information on ordering these publications, see “How to get IBM Redbooks” on page 496.

- *IBM SAN Survival Guide*, SG24-6143
- *IBM SAN Survival Guide Featuring the IBM 2109*, SG24-6127
- *IBM SAN Survival Guide Featuring the McDATA Portfolio*, SG24-6149
- *IBM SAN Survival Guide Featuring the INRANGE Portfolio*, SG24-6150
- *Designing an IBM Storage Area Network*, SG24-5758
- *Introduction to SAN Distance Solutions*, SG24-6408
- *Introducing Hosts to the SAN fabric*, SG24-6411
- *Implementing an Open IBM SAN,* SG24-6116
- *Implementing an Open IBM SAN Featuring the IBM 2109, 3534-1RU, 2103-H07*, SG24-6412
- *Implementing an Open IBM SAN Featuring the INRANGE Portfolio*, SG24-6413
- *Implementing an Open IBM SAN Featuring the McDATA Portfolio*, SG24-6414
- *Introduction to Storage Area Network, SAN*, 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
- *Storage Networking Virtualization: What’s it all about?*, SG24-6210
- *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:

- *ESS Web Interface User's Guide for ESS Specialist and ESS Copy Services*, SC26-7346
- *IBM Storage Area Network Data Gateway Installation and User's Guide*, SC26-7304
- *IBM Enterprise Storage Server Configuration Planner*, SC26-7353
- *IBM Enterprise Storage Server Quick Configuration Guide*, SC26-7354
- *IBM SAN Fibre Channel Managed Hub 3534 Service Guide,* SY27-7616
- *IBM Enterprise Storage Server Introduction and Planning Guide, 2105 Models E10, E20, F10 and F20*, GC26-7294
- *IBM Enterprise Storage Server User's Guide, 2105 Models E10, E20, F10 and F20*, SC26-7295
- *IBM Enterprise Storage Server Host Systems Attachment Guide, 2105 Models E10, E20, F10 and F20*, SC26-7296
- *IBM Enterprise Storage Server SCSI Command Reference, 2105 Models E10, E20, F10 and F20*, SC26-7297
- *IBM Enterprise Storage Server System/390 Command Reference, 2105 Models E10, E20, F10 and F20*, SC26-7298
- *IBM Storage Solutions Safety Notices*, GC26-7229
- *PCI Adapter Placement Reference, SA38-0583*
- *Translated External Devices/Safety Information*, SA26-7003
- *Electrical Safety for IBM Customer Engineers*, S229-8124
- *SLIC Router Installation and Users Guide*, 310-605759
- *SLIC Manager Installation and User Guide*, 310-605807

The IBM NUMA-Q publications which are also relevant as further information sources are available on the Web at:

- http://webdocs.numaq.ibm.com

The JNI publications which are also relevant as further information sources are available on the Web at:

- http://www.jni.com/Support/installguides.cfm

These McDATA publications are also relevant as further information sources:

- *ED-5000 Director Planning Manual*, 620-005000
- *Enterprise Fabric Connectivity Manager User Manual*, 620-005001
- *ED-5000 Director User Manual*, 620-005002
- *ED-5000 Director Service Manual*, 620-005004
- *ED-6064 Director Planning Manual*, 620-000106-100
- *ED-6064 Director User Manual*, 620-000107
- *ED-6064 Director Installation and Service Manual*, 620-000108
- *Enterprise Fabric Connectivity Manager User Manual*, 620-005001
- *FC-512 Fabricenter Equipment Cabinet Installation and Service Manual*, 620-000100
- *ES-3016 Switch Planning Manual*, 620-000110-100
- *ES-3016 Switch User Manual*, 620-000111
- *ES-3016 Switch Installation and Service Manual*, 620-000112
- *ES-3032 Switch Planning Manual*, 620-000118-000
- *ES-3032 Switch User Manual*, 620-000117-000
- *ES-3032 Switch Installation and Service Manual*, 620-000116-000
- *ES-1000 Switch Planning Manual*, 620-000102-000
- *ES-1000 Switch User Manual*, 620-000103
- *ES-1000 Switch Installation and Service Manual*, 620-000105

These QLogic publications are also relevant as further information sources:

- *QLA2200 Hardware Manual*, FC0151103-00
- *QLA2200 Hardware Manual*, FC0151103-00
- *QLA2100 Software Manual*, FC0153301-00
- *QLA2100 Hardware Manual*, FC0151102-00
- *QMS V1 Installation Guide*, FC0051104-00
- *QLview for Fibre Operations Guide*, FC0051101-00

- *QLconfig Operations Guide*, FC0051102-00

These Compaq publications are also relevant as further information sources:

- *64-Bit PCI to Fibre Channel Host Bus Adapter Release Notes for Tru64 UNIX and OpenVMS*, AV-RLLUA-TE
- *64-Bit PCI to Fibre Channel Host Bus Adapter User Guide,* AA-RKPDB-TE

Referenced Web sites

These Web sites are also relevant as further information sources:

- www.storage.ibm.com/ibmsan/index.htm
  IBM Enterprise SAN
- www.pc.ibm.com/ww/netfinity/san
  IBM Storage Area Networks: Nefinity Servers
- www.storage.ibm.com/hardsoft/products/sangateway/supserver.htm
  IBM SAN Data Gateway
- www.storage.ibm.com/hardsoft/products/tape/ro3superserver.htm IBM SAN Data Gateway Router
- www.storage.ibm.com/hardsoft/products/fcss/fcss.htm
  IBM Fibre Channel RAID Storage Server
- www.storage.ibm.com/hardsoft/products/ess/ess.htm
  Enterprise Storage Server
- www-1.ibm.com/servers/eserver/pseries/library/hardware_docs/index.html
  IBM eServer pSeries and RS/6000 Hardware Documentation
- techsupport.services.ibm.com/server/fixes?view=pSeries
  IBM eServer pSeries and RS/6000 Support
- techsolutions.hp.com
  Hewlett-Packard Support
- www.software.hp.com
  Hewlett-Packard Software Depot
- www.redhat.com
  Red Hat, Inc.
- www.suse.com
  SuSE Inc.
- www.caldera.com
  Caldera International, Inc.

- www.turbolinux.com
  Turbolinux Inc.
- www.cdp.com
  Columbia Data Products, Inc.
- www.emulex.com
  Emulex Corporation
- www.fibrechannel.com
  Fibre Channel Industry Association
- www.jni.com
  JNI Corporation
- www.mcdata.com
  McDATA Corporation
- www.peakresources.com
  PEAK Resources, Inc.
- www.pathlight.com
  Pathlight
- www.qlogic.com
  QLogic Corporation
- www.sanergy.com
  Tivoli SANergy
- www.snia.org
  Storage Networking Industry Association
- www.tivoli.com
  Tivoli
- www.t11.org
  Technical Committee T11
- www.vicom.com
  Vicom Systems
- www.vixel.com
  Vixel
- www.scsita.org
  SCSI Trade Association
- www.futureio.org
  InfiniBand (SM) Trade Association
- www.nsic.org
  National Storage Industry Consortium

- www.ietf.org
  Internet Engineering Task Force
- www.ansi.org
  American National Standards Institute
- www.standards.ieee.org
  Institute of Electrical and Electronics Engineers
- www.pc.ibm.com/us
  US Personal Systems Group
- www.compaq.com/products/storageworks/adapters.html
  Compaq Storage - Host Bus Adapters

How to get IBM Redbooks

Search for additional Redbooks or redpieces, view, download, or order hardcopy from the Redbooks Web site:

ibm.com/redbooks

Also download additional materials (code samples or diskette/CD-ROM images) from this Redbooks site.

Redpieces are Redbooks in progress; not all Redbooks become redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

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

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

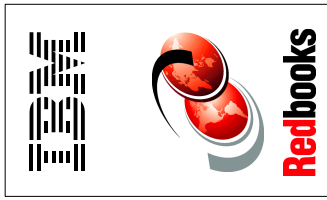
U

V

W

X

Z



Designing and Optimizing an IBM Storage Area Network

(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

Designing and Optimizing an IBM Storage Area Network

Using real life case studies, we show how to build a SAN

Review SAN designs and their application

Discover 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 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.

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

SG24-6419-00 ISBN 0738425311