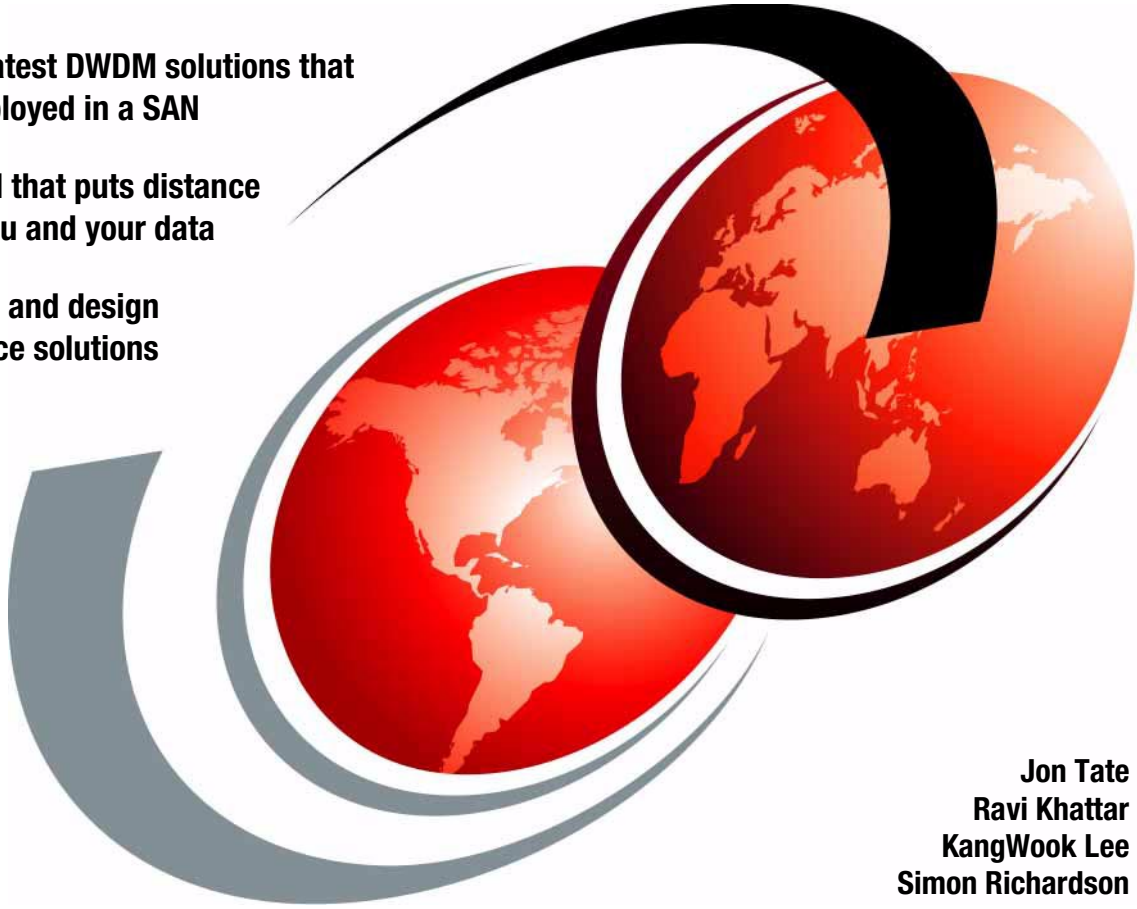
IBM

Introduction to SAN Distance Solutions

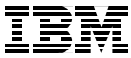
- Learn the latest DWDM solutions that can be employed in a SAN
- Build a SAN that puts distance between you and your data
- Understand and design SAN distance solutions



Jon Tate
Ravi Khattar
KangWook Lee
Simon Richardson

Redbooks

ibm.com/redbooks



International Technical Support Organization

Introduction to SAN Distance Solutions

January 2002

SG24-6408-00

Take Note! Before using this information and the product it supports, be sure to read the general information in “Special notices” on page 453.

First Edition (January 2002)

This edition applies to the hardware and software in a SAN environment.

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

Figures

Tables

Preface

The objective of this IBM Redbook is to provide information on the best configurations for distance solutions in a SAN environment. We show the particular business problems that distance solutions solve now; and with the future in mind, how these solutions can be expanded as the SAN world evolves.

We demonstrate the advantages that the IBM SAN and its OEM partners' and resellers' solutions bring to the marketplace. In addition, we provide information on the key factors to consider when choosing one particular solution over another in order to protect and maximize your return on investment.

The distributed environment, particularly SAN, has resulted in a significant increase in communications. Data storage requirements have exploded. With these two developments in mind, it is vital that we show how, where, and when data should be sent over distances quickly, and how to design and configure new and legacy systems while shaping them for the future.

This redbook introduces the storage solutions that IBM brings to the market. We introduce the concepts of Dense Wave Division Multiplexing (DWDM), and document the solutions that are incorporated in this redbook.

Attention: In this redbook we discuss technologies and products that may not be supported by IBM. However, given the scope of the IBM SAN portfolio, they may be encountered when designing or planning a distance solution based on these products. Their inclusion is based on this reason alone and should not be regarded as either a forward looking statement of support, or an assumption that support will be provided.

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 the photograph on the next page, we show the team overlooking Silicon Valley.

From left to right: Simon, Jon, Ravi, and KangWook

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, Jon worked at 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 he is an IBM SAN Certified Specialist.

Simon Richardson is an accredited Senior IT Specialist based in the United Kingdom. He has 14 years of broad experience in the IT industry, focusing during the last five years in the storage discipline. His key skills are in the SAN arena. Simon has been a driving force on many SAN engagements in the EMEA North region. His diversity of skills allows him to engage at any point in the project cycle — from proposal to scope, and design to implementation. While his primary role is within the ITS Storage and SAN Services Team, Simon often works as part of a larger, virtual, SAN consulting team.

Ravi Khattar is a Storage Pre-Sales Specialist with IBM in India. He has over 16 years of experience in the Information Technology field. He holds a degree in Electronics Engineering, and he is also a Microsoft Certified Systems Engineer (MCSE). His areas of expertise include Open Storage solutions and data management solutions on heterogeneous platforms.

KangWook Lee is a Senior IT Specialist at IBM in Korea. He is the Team Leader for Storage Solution Sales in the IBM Storage Systems Group. He has 14 years of experience as an IT Specialist, providing program service, customer support, project implementation for VSE and OS/390 users, and pre-selling storage solutions to all industry users, including UNIX, Windows 2000/NT, and mainframe customers. His areas of expertise include consulting for SAN design and storage solutions.

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

Scott Drummond
Sandy Albu
John Sing
IBM SSG

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

Ricardo Haragutchi
Walt Mostowy
IBM Global Services

Jim Baldyga
Omy Shani
Brocade Communications Systems

Ernie Swanson
Cisco Systems

Dave Burchwell
INRANGE Technologies Corporation

Karl Evert
Mindy London
Jim Straw
CNT Corporation

Chuy Perez
Henry Yang
Dave Ward
Bob Williamsen
McDATA Corporation

Michael Barrington
Nortel Networks

Brian Wood
Sorrento Networks

Special notice

This publication is intended to help planners, consultants, marketing and sales staff decide upon an appropriate SAN DWDM solution. The information in this publication is not intended as the specification of any programming interfaces. See the PUBLICATIONS section of the IBM Programming Announcement for the products we mention for more information about what publications are considered to be product documentation.

IBM trademarks

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

e (logo)® 
IBM ®
AIX®
AS/400®
CICS®
DFS™
DFSMS/MVS®
DFSMSdfp™
Enterprise Storage Server™
ES/3090™
ES/9000®
ESCON®
Extended Services®
FICON™
FlashCopy™
Footprint®
GDPS™
Geographically Dispersed Parallel IMS™
iSeries™
Magstar®
MVS™
Netfinity®
OS/390®
OS/400®

Redbooks Logo 
Redbooks™
Parallel Sysplex®
Perform™
pSeries™
RACF®
RS/6000®
S/390®
SANergy™
SP™
StorageSmart™
StorWatch™
Sysplex Timer®
Sysplex™
System/390®
Tivoli®
TotalStorage™
VM/ESA®
VSE/ESA™
Wave®
xSeries™
z/OS™
z/VM™
zSeries™
3090™

Comments welcome

Your comments are important to us!

We want our IBM 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

Products

In this first part of the book we introduce the products and technology associated with DWDM distance solutions.

1

Introduction

Until recently, disaster planning for businesses focused on recovering centralized data centers following a catastrophe, either natural or man-made. While these measures remain important to disaster planning, the protection they provide is far from adequate for today's distributed computing environments. The goal for companies today is to achieve a state of business continuity, where critical systems and networks are always available. To attain and sustain business continuity, companies must engineer availability, security and reliability into every process from the outset.

A sensible disaster recovery plan will incorporate some form of distance solution in the event of a disaster or tragedy striking.

In this chapter we will also describe some of the other business drivers that shape the IT and Storage needs.

We look at the IT challenges facing the businesses. We drill into the issues related to data and server consolidation and the need for data sharing and system interoperability.

1.1 Today’s business needs

As businesses become more and more dependent on information technology to conduct their operations and stay competitive, the availability of their processing facilities becomes crucial. Today, most businesses require a high level of availability, which extends to continuous availability, 24 hours a day and seven days a week operation. A lengthy outage could lead to significant financial losses, loss of credibility with customers, and maybe even a total failure of business. Therefore, the ability to provide continuous availability for the major applications is more often than not a necessity for business survival.

With today’ s increasingly sophisticated applications, such as business intelligence and e-business, information must be accessible to anyone who needs it.

The challenge of managing and moving data to accommodate all planned and unplanned events affecting a computer installation is overwhelming. The requirements to protect critical data from loss and deliver functions that remove the risk and anxiety associated with change are forever increasing.

Data loss through system failure, user error, theft, or natural disaster can cripple a company and cause thousands of dollars in lost revenues, reduced workforce efficiency, and missed opportunities for new products or services. Studies have shown that the average firm loses two to three percent of its gross sales within eight days of a computer outage. If the outage lasts longer than 10 days, more than half of these companies will go out of business within five years.

With the rapid growth of network computing, the reliance on the Internet and intranet technologies, and the emergence of large databases, data marts and data warehouses, data has become an increasingly vital asset for most organizations. Properly managing storage resources helps increase an organization's efficiency, productivity, and profitability by enabling it to turn data into information. Fully leveraging the information's value means anyone who needs it — employees, customers, suppliers, and business partners — must be able to access it whenever and wherever they need it.

Consider this real-life scenario: several years ago, a large airline's reservation system went down for 17 hours. This loss of function cost the company five percent of its market share for six months. This amount was equivalent to the purchase price of five new jumbo jets at that time. Now consider that 80 percent of all data loss on a daily basis is due to user error, and it becomes very clear why protecting enterprise network data is so critical.

1.1.1 IT challenges

As one of the driving forces of the changes in business, the IT landscape has itself been transformed several times over. Today, these are just some of the challenges of IT organizations:

- Rapid development and deployment of new applications
- Responding to increased complexity and pace of change
- Global reach of systems
- e-business or e-commerce creates unbounded opportunity
- Technology integration is fundamental to success
- Capitalize on fast-paced advances in technology
- Flexibility in technology investment
- Consistent drive for cost efficiencies
- Role of IT is visible and critical to customers
- Users demand more information more quickly
- Decisions increasingly dependent on information from databases
- IT as a competitive weapon to be exploited
- Data and server consolidation

Enterprises today are motivated by a need to more efficiently manage their human, financial and IT resources. Data and server consolidation helps in:

- Reduction of data management and systems management complexity
- Increased physical and data security
- Reduction of administrative staff
- Optimization of capacity utilization
- Improved reliability and availability
- System scalability
- Direct and indirect costs

1.1.2 Reduction of systems management complexity

The core issue in most organizations is fragmentation of data. The problems generated are most obvious when there are large numbers of distributed servers and storage systems.

Access to data located on multiple servers or storage systems is commonly reported to be difficult. Companies desire to obtain a unified view of enterprise data.

As experienced in many organizations, complex multi-tier architectures based on multiple servers and storage systems clearly contributed to:

- High costs
- Manageability problems
- Availability problems

- Data access or protection problems

Reducing system complexity leads to reducing system management and data management efforts.

1.1.3 Increased physical and data security

Data protection issues, including data integrity, local backup and recovery, and security issues, are also commonly cited. Management is concerned about the dangers of business disruption, customer lawsuits, and regulatory action in the event of severe data loss. Management needs to implement effective disaster recovery procedures.

1.1.4 Reduction of administrative staff

Consolidating servers and storage systems concentrates the number of administrative tasks and so decreases the number of administrative people required.

1.1.5 Optimization of capacity utilization

Servers and storage consolidation improve an enterprise's system capacity utilization. Indeed, in order to manage performance and have a level of acceptable, consistent response times, enterprises typically manage response time during peak workloads by configuring the system to run at 50 to 60% utilization, thus leaving capacity for the peak workload. The corresponding 40 to 50% of non-utilization has to be multiplied by the number of servers. The same has to be considered for storage systems.

1.1.6 Improved reliability and availability

One of the most important customer requirements is availability. Historically, most installations only considered unplanned-outages when measuring availability, but as today's systems are much more reliable, the major reason for having an outage today is a planned outage. Strategic systems, which can be internal applications, enterprise and inter-enterprise e-mail, or Web sites, must be up and running 24 hours a day and seven days a week. Availability encompasses aspects of reliability (an unreliable system is likely to become unavailable at the most inconvenient time), scalability (systems can be brought down by too much access), manageability (knowledge that systems have failed or, even better, are about to fail is key to maintaining availability), and planned maintenance (it is necessary to install new versions of hardware or software as well as applying maintenance to installed configurations).

1.1.7 System scalability

Another important requirement is to be able to quickly grow the number of users, number of applications, and size of applications with relatively little pain. This involves the ability to easily add capacity or move to larger servers or storage systems without serious disruption. This also involves the ability to deal with a peak of usage without crashing or seriously degrading performance.

Scalability can also play a key role in manageability and cost of ownership. Companies frequently have to add new servers and storage systems; this leads to increased management complexity and directly raises the staffing costs of administering and managing the environment. Consolidating servers on a scalable platform can eliminate or strongly reduce the need for adding new servers.

1.1.8 Direct and indirect costs

The following direct and indirect costs give an estimation of the total cost of ownership (TCO):

- Software-costs: Are estimated to be the result of multi-year licensing rates established under typical corporate volume purchasing environments.
- Server management costs: Can vary, depending on the complexity and costs associated with the operation of servers. The underlying value proposition of TCO is to move complexity from the small servers to a consolidated server.
- Storage and file management costs: Can vary, depending on the complexity and costs associated with the operation of both types of servers and applications. The underlying value proposition of TCO is to move from distributed storage systems to centralized storage systems.
- Operations labor costs: Can vary, depending on the complexity and costs associated with the IT operation.
- Help Desk (Tier 1 Support): Can vary, due to complexity, use of applications, and costs associated with the IT help desk operation.
- Communication costs: Typically include initial remote access, and telephone charges, and use of paid networks for remote access.
- Development costs: Can vary, depending on the application's complexity, software changes, and costs associated with software application development.
- End-User support costs: Can vary, depending on the complexity and costs associated with the IT end-user support operation.

1.2 Increased system interoperability and data sharing

Businesses have always relied on information in order to operate. The basic information that they needed was typically the market, the product, the source of the product, and how to get the product to the market. The competitive business climate today has made it critical to acquire and utilize much more information in order to survive.

Interoperability of systems has come to be recognized as a necessity in many areas worldwide, spurred on by the ever-increasing number of manufacturers and products. This interoperability between applications and servers permits reduction of purchase cost and support cost (because it's more easy to manage) and increases flexibility in the long term.

The potential user or consumer of this multiplicity of products is faced by an often bewildering choice, posing difficult investment decisions, particularly for those who are responsible for the acquisition of computer hardware and software products. Sophisticated computer installations usually require significant investments to implement, and consequently it is particularly important to ensure a satisfactory return on this type of investment.

Proprietary computer hardware and operating systems usually provide relatively better performance, but even if it is possible to standardize on a single vendor to provide solutions to the current information processing needs and requirements of an organization, change is inevitable, and future mergers or acquisitions may result in the situation where different proprietary computer systems are required to operate together effectively.

Data sharing: Different operating systems have to access to a single repository and share data. Enhanced intra-organizational and extra-organizational data sharing often requires the consolidation and integration of disparate databases.

1.3 Challenges of storage systems

In the last 30 years the goals of storage solutions have not really changed all that much. The qualitative merits of good storage solutions are the same as they were in the early days of mainframe glass houses. In 1968, 1998, and probably in 2008 there were, are, and will be five basic dimensions of value of a storage strategy. These dimensions are:

- Growth
- Access to Data
- Data Movement
- Security

- Management

Even as the goals of storage systems have remained basically the same, the competitiveness of the market has been consistently pushing them to a higher degree. This market dynamics is what makes them moving targets.

1.3.1 Growth and scalability

A storage environment must accommodate growth in data, in compute power, application types, and consumers and generators of data. Storage consolidation, the advent of e-business, and application integration have all contributed to the higher growth rate of databases.

1.3.2 Access to the data

Data must be physically located in the right place in order to be available to consumers when and where they need it, in the right format and at the required service level. Competition fuels the need for faster data access to achieve better response times.

1.3.3 Data movement

Backup copies need to be sent to other locations, and data must be migrated to new, more reliable storage hardware. Increased dependence on information requires that storage systems be more reliable. This has made the use of multi-level caching and quick backups a requirement.

1.3.4 Security

Data must be safe from disasters, caused either by nature, human error, or vandalism. It must also be protected from unauthorized access of people inside and outside the organization. While new technologies have made it possible to make business transactions more convenient, it has also made possible unauthorized access to data using new techniques. Therefore, more sophisticated ways of protecting data have to be developed.

1.3.5 Storage management

An organization has to ensure that all the desired attributes of its data are in place and dependable with minimum requirement for manual intervention or technical skill. Redundancy must be in place, and performance targets are met. The increasing cost of managing storage has made it necessary for storage designers to come up with ways to facilitate this task. That is why today, there are storage subsystems which can be managed from the network (Internet).

The challenge for storage managers is to meet all these targets without interrupting access to the data or slowing response.

1.4 Today's realities associated with storage

Traditional applications which used to be the bread and butter of corporations are now being replaced by newer, more productive applications. These new applications are able to generate product awareness and interest. They are used to effectively capture market share.

1.4.1 Applications shift

Data storage has moved upward in the hearts and minds of IT managers today. This is due to the demand for more storage generated by new applications. A user survey shows the increasing requirement for capacity of newly developed applications as compared with the traditional ones.

These new applications are:

- Data warehousing
- Decision support systems
- Multimedia applications
- e-commerce
- Intranet implementations
- Increasing use of e-mail and the Internet

1.4.2 Performance of storage systems

Hand-in-hand with the growth requirement of capacity, storage performance requirement is also rising rapidly. Complex applications, larger data sets and server consolidation — more users and more applications per server — are key factors driving the need for higher performance storage subsystems.

Easy and quick retrieval of information by applications is taken for granted. Poor response can cause customer dissatisfaction and drive away potential business opportunities.

1.4.3 Reliability of storage systems

With today's increasingly sophisticated applications — such as business intelligence, e-business, and collaborative groupware solutions — reliable data storage is more critical than ever. To keep the business running smoothly, one needs to have flexible storage and data sharing solutions that can keep pace

with the changing requirements brought on by server consolidation efforts and new applications. As applications are added or expanded, the ability to add or reconfigure storage quickly without disrupting business operations becomes critical for business success.

Users today demand higher system availability with almost any application. Core business applications, such as enterprise financial packages, have always required high availability. But today, e-mail, intranet access, and even regular file servers are among the new set of mission-critical applications.

Quick backups and quick restores of data are the norm. Online windows are expanding while batch windows are shrinking. Business executives have long realized that when end users are not productive, the bottom line is affected. This leaves no room for unavailability of data.

Some find it easy to quantify the amount of business lost for a given amount of downtime. Not all situations involving downtime of information systems are easy to measure, but ensuring data availability is critical. Businesses expect and depend on rapid data and information exchange within and across networks. This makes it essential for an enterprise to have a dependable technology for doing data backup, recovery, and storage management.

1.4.4 The budget for storage

This is an age where the cost per gigabyte of storage has significantly decreased and is still decreasing. However, the demand for greater storage capacity is such that storage is now regarded as the largest single IT hardware expense. Market analysts at Dataquest noted in a 1997 report that, “Storage has become the most costly hardware component in the IT environment”. It takes up 35% of the average hardware budget, and it keeps growing.

Even as the budget for storage hardware has grown, the cost of maintaining it is even higher. Management costs dwarf the purchase price for storage. Industry analysts agree on probably the most significant item with regard to storage is total cost of ownership (TCO):

- Storage management costs greatly exceed the capital investment in most cases.
- For decentralized storage, management expense is estimated to be up to seven times greater than the purchase price.

1.5 Summary

Old technologies and architectures usually evolve to meet new business requirements. New technologies and architectures are constantly being developed also for this purpose. To effectively leverage the information resource, it is wise to view these technical developments from a business perspective and keep the following reminders in mind:

- Understand the business requirements
- Keep an application view
- Avoid a technology focus
- Check short term cost and implementation speed
- Consider long term cost and benefit
- Maintain flexibility
- Ensure easy data access and protection
- Monitor performance from the end user's view
- Maintain high availability
- Keep management cost and effort at a profitable level

There are several choices for solutions, servers, storage devices and the related technology. The correct choice of a business solution can mean the success of an enterprise for several years. On the other hand, an ill-advised choice can mean the unavailability of data when or where it is needed. This can result in missed business opportunities and lost market share. In the final analysis, the criteria for selection should be based on business needs of the present, the short term benefits and long term projections of the enterprise.

With the products and solutions that we describe in this redbook, we will maintain our focus on the one item that is an integral part of disaster recovery. This is the ability to put distance between your primary and secondary sites.

2

Extending fiber over distance with DWDM

This chapter gives an overview of attaining higher throughput within a fibre network environment. We describe the technology and methods used to extend fibre networks over distance. Initially it is important to understand the distances and terminology used to describe the geography that we could extend over. We then describe the components that enable this today. Technologies that we include in this overview are traditional time-division multiplexing (TDM), wavelength division multiplexing (WDM) and dense wave division multiplexing (DWDM).

This chapter gives an understanding of what this means for the storage area networking (SAN) arena today.

2.1 Why we need fiber saving devices

It's the need for speed and more bandwidth in the same over-subscribed fibre (laying new fibre is very expensive), and driving existing fibre harder is an economically attractive solution (especially when you can mix protocols).

There are two ways of increasing the effective capacity of already deployed fibre. You can increase the bit rate of existing systems or increase the number of wavelengths being transmitted down the fibre. In this chapter we will discuss both methods here. We will also discuss both the architecture and resilience as these infrastructures are likely to be mission critical to your business.

By its very nature, today's storage area networking (SAN) environments continue to evolve. This continues to be driven by many diverse business factors. We will consider many of these driving factors and how each solution addresses these problems. We will demonstrate how specific vendors solutions integrate with IBM SAN solutions, and the business problems they address.

Graphically dispersed networks historically have been core to edge networking architectures that have evolved from LAN to MAN and/or WAN. This redbook is based around extending SAN over distance to facilitate geographically dispersed location. This book will be addressed from a storage center solution perspective, working from storage up to servers, not from servers down.

To facilitate this we will consider the core to edge architecture and how it fits in the distance market place. For the purpose of this book, SAN will be regarded as the center of the enterprise. This seems in line with current industry trends as enterprises find the value and size of data to be growing exponentially.

2.1.1 Storage Area Network

In this e-commerce economy, it is no secret that data is business. Because of the increased dependence on information, the storage needs of companies are growing exponentially. Research shows that storage needs for traditional bricks-and-mortar companies are doubling every year, and every 90 days for dot coms. This growth raises new concerns for the maintenance and protection of valuable data resources.

Historically in open system storage environments, physical interfaces to storage consisted of parallel SCSI channels supporting a small number of SCSI devices. Storage Area Networks use new technologies to connect greater numbers of servers and devices. Deployments of SANs today are exploiting the storage-focused capabilities of Fibre Channel. The Fibre Channel SAN consists of hardware components such as storage subsystems, storage devices, and servers that are attached to the SAN via interconnect entities (host-bus adapters,

bridges, hubs, switches). Another hardware element being deployed in SANs is commonly referred to as a SAN appliance or SAN server. These SAN appliances are computing elements attached directly to the SAN or installed in the storage data path. These SAN appliances are responsible for managing the Fibre Channel topology and additionally providing an abstraction of storage.

The heart of the storage management software is the virtualization. The term virtualization, when it pertains to disk storage, refers to the representation of a storage unit or data to the operating system and/or application running on an application server. The storage unit or data presented is decoupled from the actual physical storage where the information may be contained. There is some method for providing the translation between the logical and physical storage. Once the storage is abstracted, storage management tasks can be performed with a common set of tools from a centralized point, which will greatly reduce the cost of administration.

2.1.2 Metropolitan Area Network

Metropolitan Area Networks (MAN) will typically be a city wide solution, connecting systems across areas of 50 to 100 km. These networks will typically carry traffic between office locations within a metropolitan area. A MAN will often carry many diverse networking protocols, and these protocols will all have their own speed and performance characteristics. It can also be an intermediate media interfacing between a Storage Area Network (SAN) and a Wide Area Network (WAN). This begins to build a hierarchical model of the SAN to MAN to WAN. MANs have traditionally been based on Synchronous Optical Network (SONET) or Synchronous Digital Hierarchy (SDH) technology using point-to-point or ring topologies with add/drop multiplexers (Aims).

The MAN has two critical goals to achieve. It must meet the needs created by the dynamics of the ever-increasing bandwidth requirements. It must also address the growing connectivity requirements and access technologies that are resulting in demand for high-speed, customized data services.

2.1.3 Wide Area Networks

Wide Area Networks (WAN) are at the very core of a global network. Like the MAN, the WAN's purpose is as a transport medium, and this means they need to be resilient with a high level of capacity. These networks are often provisioned by SONET or SDH technology. Due to the high demand for bandwidth these solutions are experiencing large demands on their fibre. WAN’s are likely to be public (in the case of the Internet) or private in the case of a company’s self

deployed dedicated WAN. Some WANs are based on an agreed bandwidth within a public network, and this leads to virtual private networks. Most large corporate customers will deploy one of these last two options, or a combination of the two.

2.2 Fiber extending device concepts

There are three methods that are employed to extend the distance of fibre SAN solutions. The first we will describe is channel extenders. These translate incoming signals to another protocol that can be sent further, often this utilizes compression algorithms to help achieve speed at long distances. The second is known as Time Division Multiplexing or TDM. Multiple channels are transmitted on a single carrier by increasing the modulation rate and allotting a time slot to each channel. However, increasing the bit rate of a system requires more sophisticated high-speed electronics at the transmitting and receiving ends of the communications link. And as the bit rate increases, inherent modulation limiting characteristics of optical fibers become dominant. Chromatic and polarization mode dispersion cause pulse spreading, which affects the signal quality over longer transmission distances.

2.2.1 Channel extenders

Channel extenders are used to provide longer cable distances. Most optical interfaces are multimode cable. Extenders convert the multimode interface to single mode and boost the power on the laser. Typically, an extender will provide a single mode cable distance of 30 km or 18 miles.

Compression

Compression is the reduction in size of data in order to save space or transmission time. For data transmission, compression can be performed on just the data content or on the entire transmission unit (including header data) depending on a number of factors. Content compression can be as simple as removing all extra space characters, inserting a single repeat character to indicate a string of repeated characters, and substituting smaller bit strings for frequently occurring characters. This kind of compression can reduce a text file to 50 percent of its original size. Compression is performed by a program that uses a formula or algorithm to determine how to compress or decompress data.

The net result of compression is less data. The distance we can send data is often gaited by the amount of space or memory that it takes up, so less information or data to send often means longer distances are achievable.

2.2.2 Multiplexers

Multiplexing is the process of simultaneously transmitting multiple signals over the same physical connection. There are two common types of multiplexing used for fiber optic connections:

- Time Division Multiplexing (TDM)
- Wavelength Division Multiplexing (WDM)
- Dense Wavelength Division Multiplexing (DWDM)

Time-Division Multiplexing

Time-division multiplexing (TDM) was created by the telecommunication industry to maximize the amount of effective traffic that could be carried over an existing fibre network. Previously in the telephone industry, every phone call needed its own discrete physical link. This system was obviously very costly to operate and limited for growth. The multiplexing facility enabled many phone circuits to be sent over a single link. In summary TDM enables many signals to be concentrated into a single fibre and all within the same wavelength.

TDM can be thought of as highway traffic. Traffic that needs to traverse from one city to another will start out on minor routes, it will then make it to a main (single lane) highway where it will join into a slot between other vehicles. The traffic will be controlled so that it is fair to all minor routes. Once the vehicles arrive at their destination they will be taken off to minor routes again.

This method is used in synchronous TDM devices. It increases the capacity of the fibre by reducing time slots into smaller intervals. This enables more bits in the same fibre which increases the effective bandwidth of the fibre. Within the TDM the input sources are multiplexed in a fair time share manner. Each signal will therefore have a packet space allocated to it whether it has signal input to send or not. This can be inefficient because there can be unused packets within the frame. Some protocols will reduce the effect of this by keeping data flowing into the channel. This is more likely in an asynchronous architecture.

Figure 2-1 shows a TDM and its method of combining several slower speed data streams into a single high speed data stream. Data from multiple sources is broken into portions (bits or bit groups) and these portions are transmitted in a defined sequence. Each of the input data streams then becomes a “time slice” in the output stream. The transmission order must be maintained so that the input streams can be reassembled at the destination.

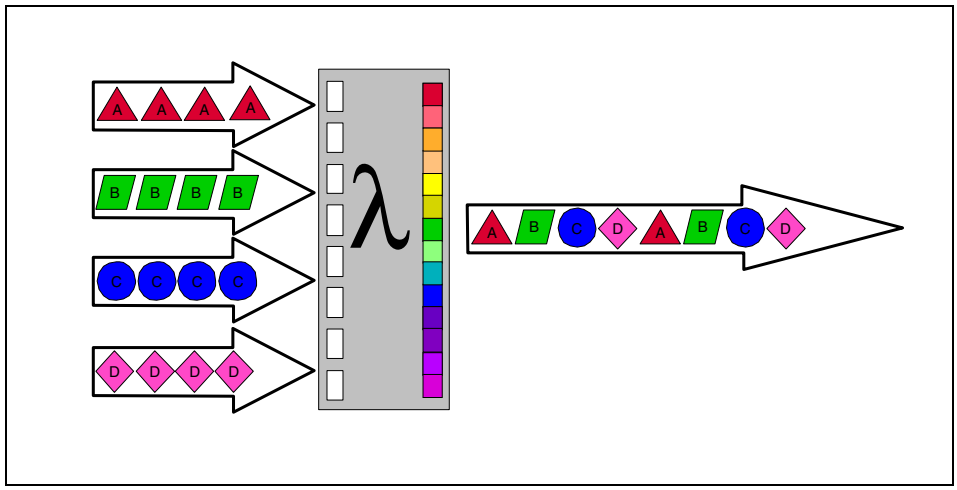


Figure 2-1 Time Division Multiplexer concepts

Wavelength Division Multiplexing

Wavelength Division Multiplexing (WDM) differs from TDM in that it does not use time to multiplex on — it uses the many wavelengths of light to multiplex on. WDM receives incoming optical signals from many sources (devices) which it converts to electrical signals, it then assigns them a specific wavelength (or lambdas or λ) of light and retransmits them on that wavelength. This method relies on the large number of wavelengths available within the light spectrum. You can think about WDM as though each channel is a different color of light; several channels then make up a “rainbow.” In summary WDM enables many signals to be concentrated into a single fibre all being sent at different wavelengths.

WDM allows the simultaneous transmission of a small number of data streams over the same physical fiber, each using a different optical wavelength. The advantages of WDM over TDM are that transmission order does not need to be maintained and that the information streams can use different protocols and bit rates. The WDM is sometimes described as the coarse wave division multiplexer.

Figure 2-2 shows each of the input signals coming into the WDM being multiplexed at different wavelengths on the output.

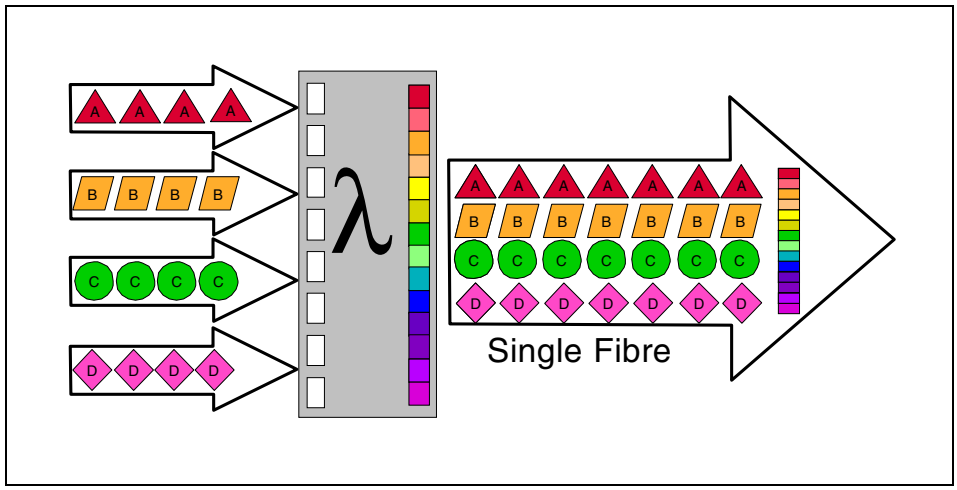


Figure 2-2 Wave Division Multiplexer concepts

This output is all multiplexed into the fiber. At the other end of the fiber, the signals are demultiplexed by the receiving WDM. This differs from the TDM method where time is displaced. Within the WDM environment each signal has its own wavelength. This means that they can all co-exist within the same fibre at the same time. This results in all channels having full bandwidth.

Dense Wave Division Multiplexer

Dense Wave Division Multiplexer (DWDM) uses the same design principles as the WDM, it can simply handle a much larger number of wavelengths. By reducing the spacing of these wavelengths more channels can be accommodated. As each channel maps to its own individual wavelength, the more wavelengths we have, the greater the total capacity or bandwidth of the device. We currently see DWDM devices that are capable of driving 256 discrete wavelengths along one single mode fibre.

DWDM is an approach to opening up the conventional optical fiber bandwidth by breaking it up into many channels, each at a different optical wavelength (a different color of light). Each wavelength can carry a signal at any bit rate less than an upper limit defined by the electronics, typically up to several gigabits per second. DWDM has all the advantages of WDM but with the added benefit of supporting far more independent transmissions over the same fiber. Due to the nature of these boxes, they are often considered transparent to protocol and bit rate.

Note: Some DWDM solutions use TDM within a channel (wavelength) for some low data rate protocols. This allows more efficient use of the available bandwidth.

The business drivers for DWDM are clear and deliver:

- **Bandwidth**

  If you currently have a pair of fibre installed that you are using for a single channel then by employing DWDM's that could bring you 255 extra channels per fibre.

- **Protocol independence**

  The DWDM is deployed as part of the physical layer. It is therefore independent of protocol, simply passing signal information in the format it is received. Examples of the protocols it can support are ATM, Gigabit Ethernet, ESCON, and Fibre Channel.

- **Growth on demand**

  A DWDM solution can preserve investment in already deployed fibre infrastructure, and be easily expanded to meet growing capacity.

- **Speed to market**

  Once a DWDM solution is deployed it can be used quickly and efficiently for new application, transparent of application and protocol. This enhances your ability to react to platform changes at an enterprise level. This can never be understated in the new e-business world.

DWDM Summary

Figure 2-3 shows a generic overview of the components within a DWDM. The incoming signals form inputs are varied, we have simply shown SAN, ESCON and other protocols. Other Protocols could include, Gigabit Ethernet, SONET (OC-3, OC-12, OC-48), SDH (STM-1, STM-4, STM-16), Fibre Channel (1 Gbps), ESCON, FICON, and more.

These are often vendor specific as to what is supported and should be checked at the product level. We show the translation within the Optical to Electrical to Optical converter or Transponder and this takes the input signal on a specific wavelength and converts it to electrical signal which is then re-modulated on the new frequency that it will use during transit within the dark fibre media. These new wavelengths should adhere to the ITU-T grid, however different vendors will often use different channels from this grid and may even skip channel.

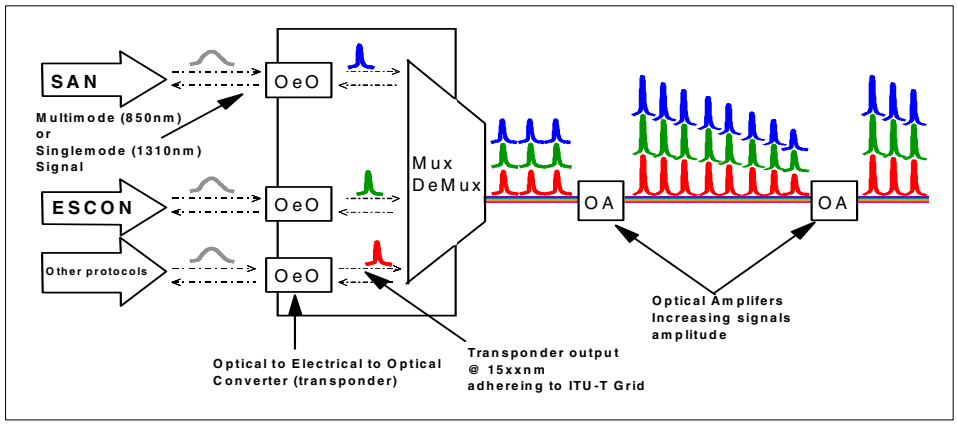


Figure 2-3 Generic example of DWDM architecture

Some of the differentiations within the DWDM market place are:

- Optical filters
  - Thin-film
  - Laser-welded
  - Fibre Bragg Grating
  - Planar-waveguide
- Channel expansion methods
  - Interleaving
  - Reduced channel sizes greater bandwidth
- Modulation techniques
  - Direct
  - Indirect

Many of these features have price and performance trade offs, however the comparison of these components within the DWDM industry is beyond the scope of this redbook.

2.2.3 Multiplexers and demultiplexers

The DWDM needs internal components that are capable of taking the many input signals and aligning them to the wavelength that they occupy within the fibre during transfer. This work is performed by the multiplexer, it does it by taking the signal wavelengths from the input fibres and converges them into one beam of light that is comprised of many wavelengths. This band of light then gets sent across the fibre to the receiving DWDM. At the receiving end, the opposite operation is performed, the signal is taken and split out at an optical level. This is

then sent to the appropriate receiving photo detector. DWDM architecture components include DWDM filter modules, transmitters, receivers, (DWDM-capable) optical amplifiers, integrated optoelectronics, tunable filters used to add or drop specific frequencies.

In the topics that follow we will describe some of the internal components.

Lasers

There are two types of light emitting devices that are used in optical transmission, light-emitting diodes (LEDs) and laser diodes. The LED's are typically used for slower designs, and are often found in multimode implementations with speeds up to 1 Gb/s. LED's are relatively inexpensive devices. Lasers diodes are more expensive but lend themselves better to single mode devices. Two types of lasers diodes are widely used, monolithic Fabry-Perot lasers, and distributed feedback (DFB) lasers. The latter type is particularly well suited for DWDM applications, as it emits a nearly monochromatic light, is capable of high speeds, has a favorable signal-to-noise ratio, and has superior linearity. DFB lasers also have center frequencies in the region around 1310 nm, and from 1520 to 1565 nm. The latter wavelength range is compatible with EDFAs.

There are many other types and subtypes of lasers. Narrow spectrum tunable lasers are available, but their tuning range is limited to approximately 100-200 GHz. Under development are wider spectrum tunable lasers which will be important in dynamically switched optical networks. Cooled DFB lasers are available in precisely selected wavelengths.

Photo detectors

The photo detector is necessary to recover the signals transmitted at different wavelengths on the fiber. Photo detectors are wideband devices and cannot identify which band it is detecting. This means that the optical signals have to be demultiplexed before they reach the detector. The industry today tends to use two types of photodetector, the positive-intrinsic-negative (PIN) photodiode and the avalanche photodiode (APD). PIN photodiodes work in a similar fashion to LED's except in reverse — light is absorbed rather than emitted and photons are converted to electrons which are transmitted as electrical signals. APDs are similar to PIN photodiodes, however they amplify the signal. This result in one photon releasing many electrons (that is to say, one-to-many). While PINs are cheaper and more reliable APDs have more sensitivity and accuracy.

In Figure 2-4 we show a multiplexer to demultiplexer.

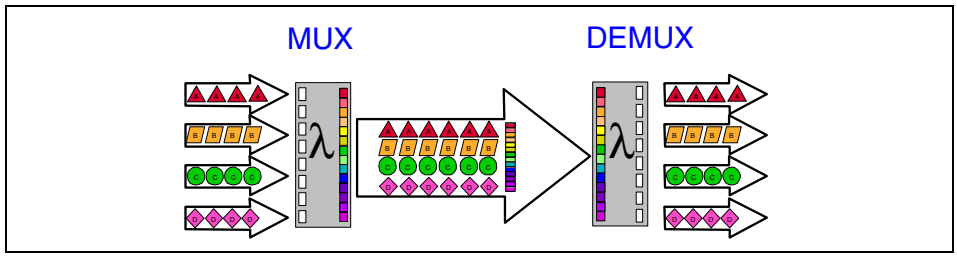


Figure 2-4 Multiplexer to demultiplexer

Different wavelengths are utilized to send traffic in opposing directions on the same fibre. This is achieved by using different wavelengths for each direction. In these instances there is a mux/demux function in the device at both ends of the fibre. This is shown in Figure 2-5.

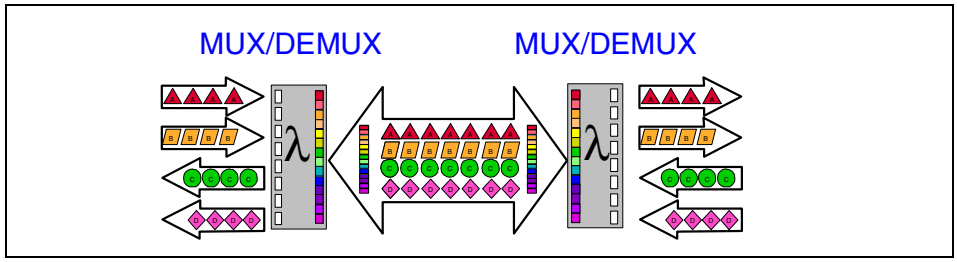


Figure 2-5 Both multiplexer and demultiplexer

Optical amplifiers can also be used to boost signal power after multiplexing or before the demultiplexer function.

2.2.4 Optical add/drop multiplexers

Another component that offers functionality in this arena is the optical add/drop multiplexers (OADM). These devices can be used at interim points between DWDM mux/demux units to split off or to inject signal wavelengths from or into the fibre. These units can be static devices or active.

The static first generation of OADMs have been built and configured to strip off (drop) or inject (add) specific wavelengths. The dynamic second generation of OADMs can be dynamically reconfigured to add or drop specific wavelengths to or from the wavelengths present in the fibre. Wavelengths that are not to be added or dropped from the traffic continue unchanged and are sometimes referred to as express channels for the purpose of that OADM.

In Figure 2-6 we show light being stripped out.

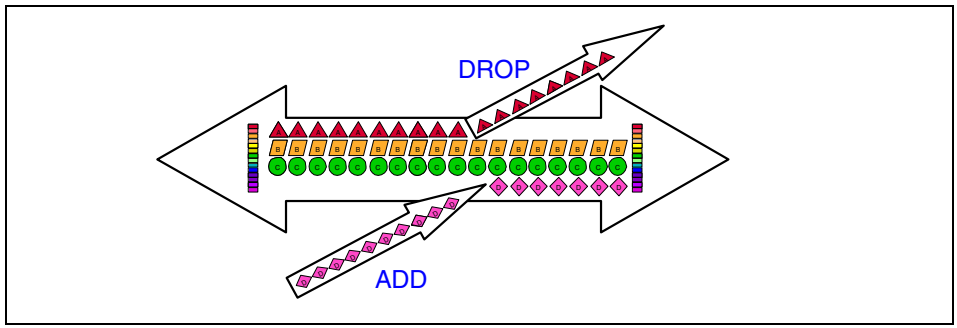


Figure 2-6 Light dropped and added

An OADM is used at intermediate stations. It removes (drops) a channel from a combined DWDM signal or adds a channel to a combined DWDM signal without interfering with the other channels on the fiber. After a channel has been dropped, the wavelength then becomes available to be reused by a different signal.

Figure 2-7 shows an OADM multiplexer/demultiplexer device. It has a crystal of transparent material with parallel sides on which a dielectric filter is deposited. The filter allows a single wavelength to be transmitted, reflecting all others. Therefore, a ray of light entering the device through a Graded Index (GRIN) lens will have one wavelength separated or demultiplexed from it. The device will operate in reverse as a DWDM multiplexer.

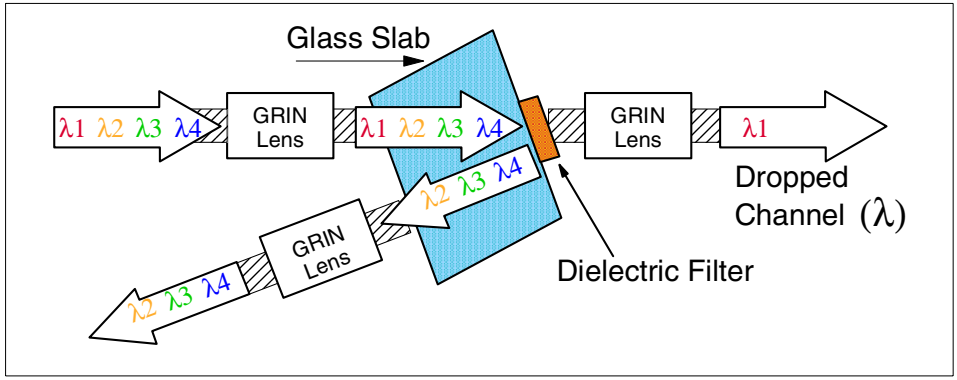


Figure 2-7 Example of OADM using dielectric filter

2.2.5 Optical amplifiers

Due to attenuation, the distance the signal on a fibre can propagate without loss of integrity is limited. We overcome this by using optical amplifiers to increase the signal strength. In the past we needed a repeater for each channel or signal transmitted, however this is now not true. The optical amplifier will amplify all the signals (wavelengths) as light and they will not need to be converted to from optical to electrical to optical in this process. Typically signals travel for up to 120 km between amplifiers, and longer distances are attainable, but the signal must be regenerated.

2.2.6 Regenerative repeaters

A regenerative repeater is a device that regenerates optical signals by converting incoming optical pulses to electrical pulses. It will clean up the electrical signal to eliminate noise, and reconvert them to optical pulses for output. This gives the ability to extend over very long distances.

2.3 DWDM topologies

DWDM can be implemented in more than one way and we describe these topologies:

- Point-to-point
- Linear
- Ring
  - Hubbed ring
  - Mesh ring

2.3.1 Point-to-point

This is the simplest of the three implementations. A point-to-point topology is a connection between two DWDM connections across a pair of single fibres. This is implemented with two Fibre Channels, one of which is considered an east link, and the other a west link.

Figure 2-8 shows a simple point-to-point connection.

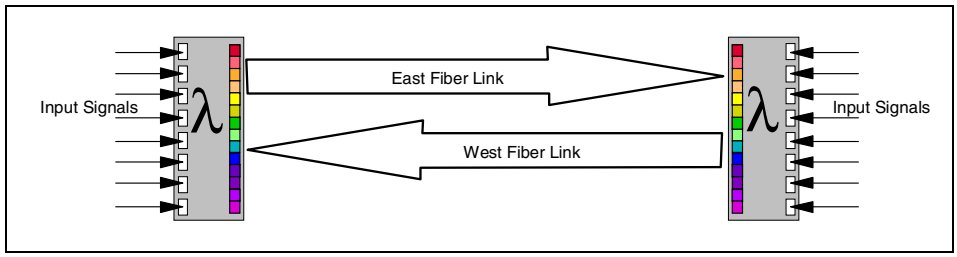


Figure 2-8 Point-to-point topology

2.3.2 Linear

A linear topology is a logical progression from the point-to-point architecture. It is a connection between DWDMs that are set out in a linear fashion. This is implemented with two Fibre Channels with one considered an east link and one a west link, between each DWDM station.

Figure 2-9 shows a simple linear connection.

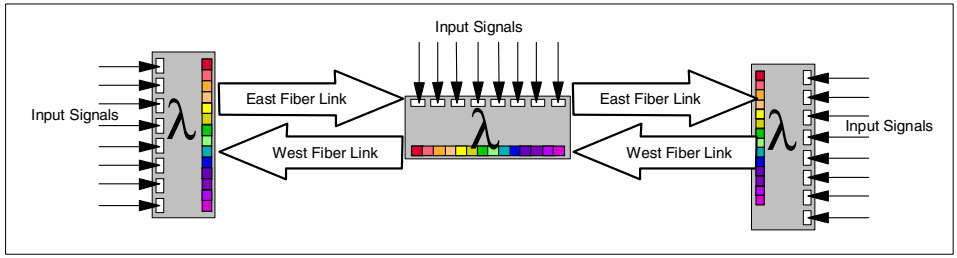


Figure 2-9 Linear topology between three locations

2.3.3 Ring

This topology is typically implemented where many geographically dispersed locations need to be connected. We will show a ring here with four points of presence. This solution could be implemented with only one or more DWDMs or could comprise of many components including OADM devices and Hubs. Channels can be dropped and added at one or more nodes on a ring. Rings have many common applications, including providing extended access to SANs where increasing the capacity of existing fiber is desirable.

Hubbed ring

A hubbed ring is composed of a hub node and two or more add/drop or satellite nodes. All channels on the ring originate and terminate on the hub node. At the add/drop node certain channels are terminated (dropped and added back) while the channels that are not being dropped (express channels) are passed through optically, without being electrically regenerated. This is shown in Figure 2-10.

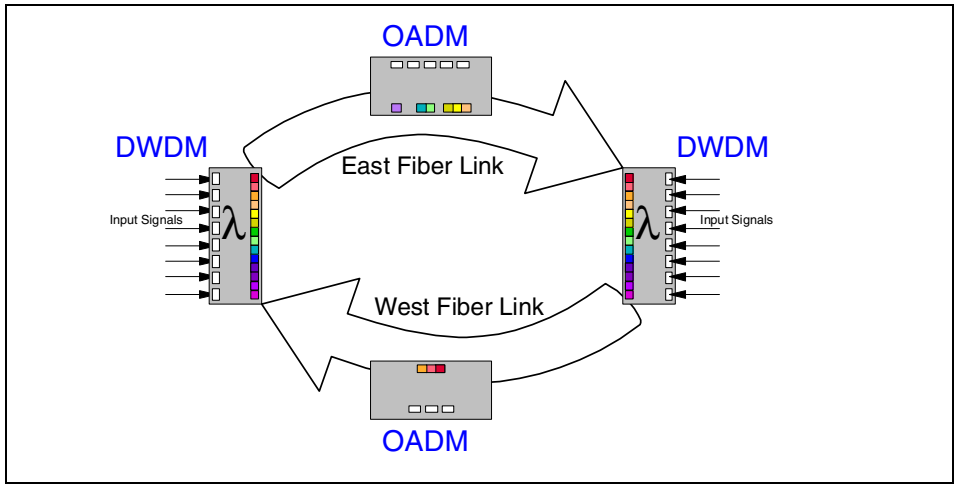


Figure 2-10 Ring topology using two DWDM and two OADM

Meshed ring

A meshed ring is a physical ring that has the logical characteristics of a mesh. While traffic travels on a physical ring, the logical connections between individual nodes are meshed.

Figure 2-11 shows a ring topology, and logical connections between some of the nodes can be thought of as meshed.

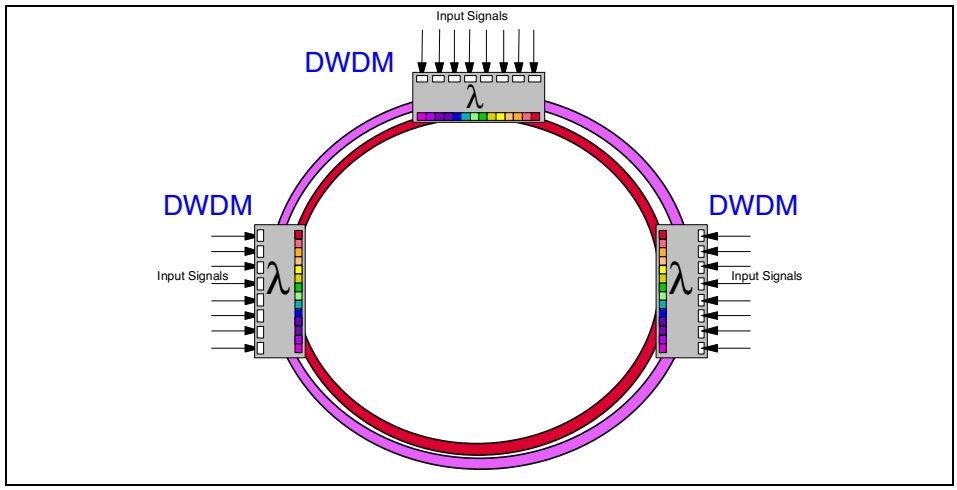


Figure 2-11 Ring topology with three DWDM

East and west

DWDM's are often composed of shelves each of which operate in a specified wavelength band, which is determined by the wavelength band of the optical modules installed in the shelf.

Figure 2-12 shows a DWDM shelf. Each shelf is divided into two parts, the west side and the east side. The two sides must have the same wavelength band.

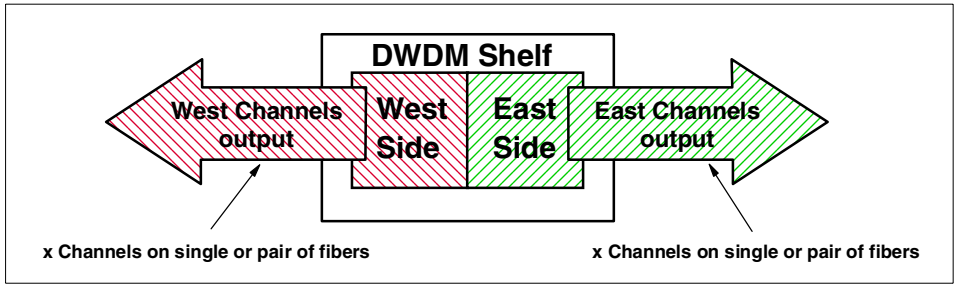


Figure 2-12 DWDM module showing east and west

A side can have many channels, each operating on a different wavelength within the shelf's wavelength band. Each west side channel has the same wavelength as the corresponding east side channel.

Figure 2-13 shows a DWDM comprising of four shelves, each with its own band.

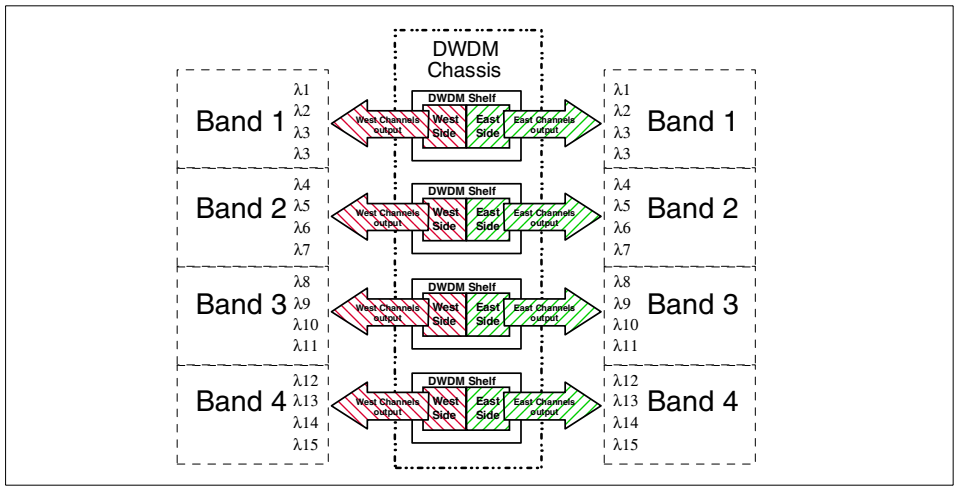


Figure 2-13 East and west must have same wavelengths within the same band

We also introduce the concept of those bands having many wavelengths. In this example we have four wavelengths (λ) per band. As described above, both east and west sides of the optical module must operate on the same wavelengths.

Protection

Protection within the DWDM devices comes in two flavors:

- **Internal protection**

  Internal protection is the ability to configure redundant components internal to the DWDM chassis. This protection is often at component level. Components that will often have internal protection include CPU modules, power supplies, fan assemblies.

- **External protection**

  External protection is the ability to configure redundant fibres external to the DWDM device. This protection is aimed at surviving fibre failure. Implementing protected circuits will often halve the number of channels that are available to you. This protection can be implemented by deploying redundant fibres. Many DWDMs support this option and will automatically sense and reroute their signal upon loss of the primary link.

  Redundant paths can be implemented using line protection or path protection.

  – Line Protection is when you have two lasers transmitting — one transmits east, one transmits west.

  – Path Protection is when you have one laser and the signal is optically split to the east and the west.

2.4 Factors that will affect distance

Fiber channel distances depend on many factors and include:

- Type of laser used, longwave or shortwave
- Type of fiber optic cable, multi-mode or single-mode
- Quality of the cabling infrastructure in terms of dB loss — connectors, cables, and even bends and loops in the cable can result in dB signal loss
- Native shortwave FC transmitters have a maximum distance of 500 m with 50 micron diameter, multi-mode, optical fiber
- Although a 62.5 micron, multi-mode fiber can be used, the larger core diameter has a greater dB loss and maximum distances are shortened to 300 meters.
- Native longwave FC transmitters have a maximum distance of 10 km when used with 9 micron diameter single-mode optical fiber.

Link extenders will provide a signal boost that can potentially extend distances up to about 100 km. These link extenders simply act as a very big, fast pipe. Data transfer speeds over link extenders depend on the number of buffer credits and efficiency of buffer credit management in the FC nodes at either end of this fast pipe. Buffer credits are designed into the hardware for each FC port.

FC provides flow control that protects against collisions. This is extremely important for storage devices, which do not handle dropped or out-of-sequence records. When two FC ports begin a conversation they exchange information on their buffer capacities. An FC port will send only the number of buffer frames for which the receiving port has given credit. This not only avoids overruns, but also provides a way to maintain performance over distance by filling the “pipe” with in-flight frames or buffers.

The maximum distance that can be achieved at full performance is dependent on the capabilities of the FC node that is attached at either end of the link extenders. This is very vendor specific. There should be a match between the buffer credit capability of the nodes at either end of the extenders. A host bus adapter (HBA) with a buffer credit of 64 communicating with a switch port with only eight buffer credits would be able to read at full performance over greater distance than it would be able to write. This is because on the writes, the HBA can send a maximum of only eight buffers to the switch port, while on the reads, the switch can send up to 64 buffers to the HBA.

Until recently, a rule of thumb has been to allot one buffer credit for every 2km in order to maintain full performance.

2.4.1 Light or link budget

It is important to understand the link budget terminology. The decibel (dB) is a convenient way of expressing an amount of signal loss or gain within a system or the amount of loss or gain caused by some component of a system. When signal power is lost, you never lose a fixed amount of power. The rate at which you lose power is not linear. Instead you lose a portion of power — one half, one quarter, and so on. This makes it difficult to add up the lost power along a signal's path through the network if measuring signal loss in watts.

For example, a signal loses half its power through a bad connection, then it loses another quarter of its power later on through a bent cable. You cannot add 1/2 plus 1/4 to find the total loss. You must multiply 1/2 by 1/4. This makes calculating large network dB loss time-consuming and difficult.

Decibels, though, are logarithmic. This allows us to easily calculate the total loss/gain characteristics of a system just by adding them up! Keep in mind that they scale logarithmically. If your signal gains 3dB, the signal doubles in power. If your signal loses 3dB, the signal halves in power.

It is important to remember that the decibel is a ratio of signal powers. You must have a reference point. For example, you can say, "there is a 5dB drop over that connection." But you cannot say, "the signal is 5dB at the connection." A decibel is not a measure of signal strength, but a measure of signal power loss or gain. A decibel milliwatt (dBm) is a measure of signal strength. People often confuse dBm with dB. Do not fall into this trap! A dBm is the signal power in relation to one milliwatt. A signal power of 0 dBm is one milliwatt, a signal power of 3 dBm is 2 milliwatts, 6 dBm is 4 milliwatts, and so on. Also, do not be misled by minus signs. It has nothing to do with signal direction. The more negative the dBm goes, the closer the power level gets to zero.

For example, -3 dBm is 0.5 milliwatts, -6 dBm is 0.25 milliwatts, and -9 dBm is 0.125 milliwatts. So a signal of -30 dBm is very weak.

2.4.2 Buffer credits

Buffer credits within the switches and directors have a large part to play in the distance equation. The buffer credits in the sending and receiving nodes heavily influence the throughput that is attained within the Fibre Channel. Fibre Channel architecture is based on a flow control that ensures a constant stream of data to fill the available pipe. A rule-of-thumb says that to maintain acceptable performance one buffer credit is required for every 2 km distance covered.

Buffers

Ports need memory, or “buffers”, to temporarily store frames as they arrive and until they are assembled in sequence, and delivered to the upper layer protocol. The number of buffers, that is the number of frames a port can store, is called its “Buffer Credit”.

BB_Credit

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

EE_Credit

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

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

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

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

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

This is particularly true with distance. At 1 Gb/s a frame occupies 4 km of fiber. In a 100 km link we can send 25 frames before the first one reaches destination. We need an ACK (acknowledgment) back to start replenishing EE_Credit. We will be able to send another 25 before we receive the first ACK. We need at least 50 buffers to allow for non stop transmission at 100 km distance.

2.4.3 Quality of fiber

The optical properties of the fibre will influence the distance that can be supported. There is a decrease in signal strength along a fiber. As the signal travels over the fibre, it is attenuated, and this is caused by both absorption and scattering; and this is usually expressed in decibels per kilometer (dB/km). Some early deployed fibre was designed to support the telephone network and this is sometimes of insufficient specification for the new multiplexed environments. If you are being supplied dark fibre by another party you will normally specify that they must not allow more than xdB loss in total.

2.4.4 Cable types

Light is sent into the fibre by the emitter and it propagates in optical pulses and is detected by the detector. In Figure 2-14 we show this.

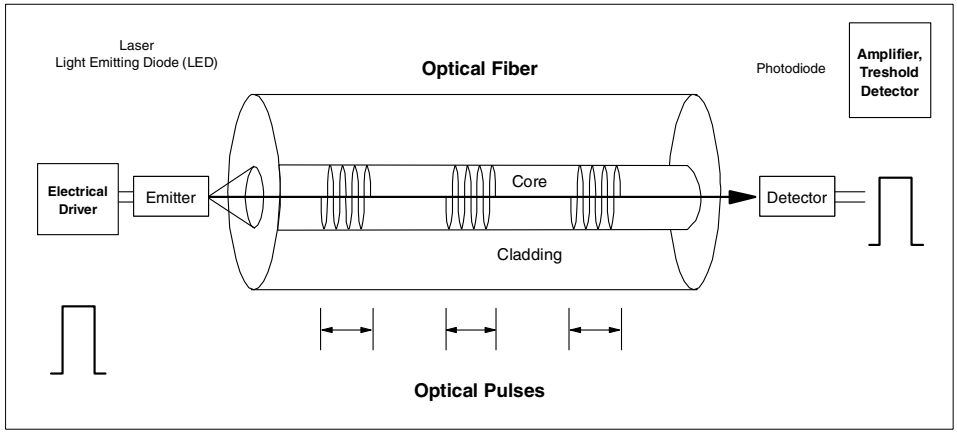


Figure 2-14 Light propagation through fiber

Fibre cables within the SAN environment break down into two categories:

- Single mode fibre
- Multimode fibre

Note: DWDM applications use single mode fibre on the transport or dark fibre side.

In most cases, it is impossible to distinguish between single-mode and multi-mode fiber with the naked eye (unless the manufacturer follows the color coding schemes specified by the Fibre Channel physical layer working subcommittee of orange for multi-mode and yellow for single-mode). There may be no difference in outward appearance, only in core size.

Both fiber types act as a transmission medium for light, but they operate in different ways, have different characteristics, and serve different applications. We show the light propagation characteristic of single mode fibre (Figure 2-15) and in multi mode fibre (Figure 2-16).

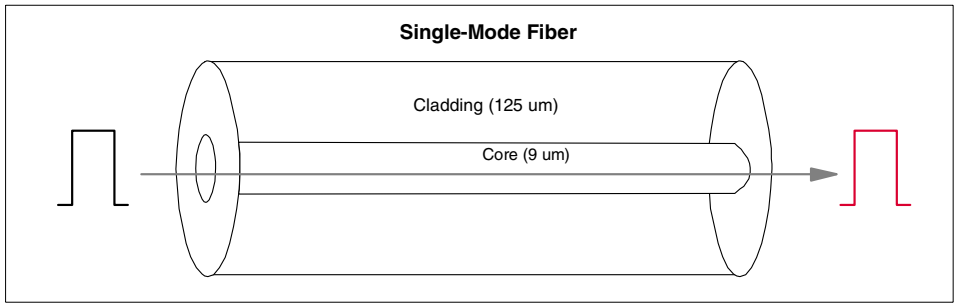


Figure 2-15 Light propagation in single-mode fiber

Single-mode (SM) fiber allows for only one pathway, or mode, of light to travel within the fiber. The core size is typically 8.3 µm. Single-mode fibers are used in applications where low signal loss and high data rates are required, such as on long spans between two system or network devices, where repeater/amplifier spacing needs to be maximized.

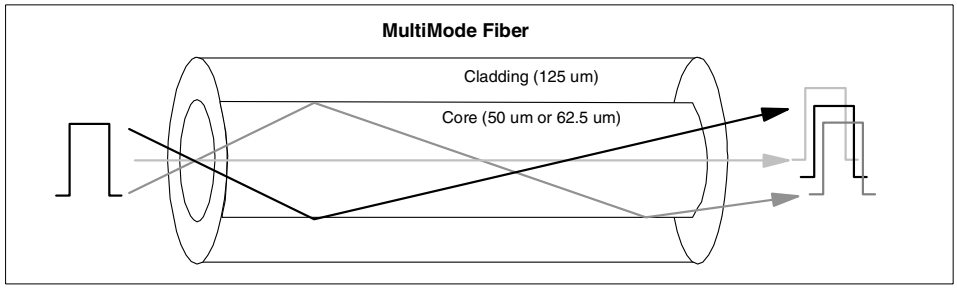


Figure 2-16 Light propagation in multi-mode fiber

Multi-mode (MM) fiber allows more than one mode of light. Common MM core sizes are 50 µm and 62.5 µm. Multi-mode fiber is better suited for shorter distance applications. Where costly electronics are heavily concentrated, the primary cost of the system does not lie with the cable. In such a case, MM fiber is more economical because it can be used with inexpensive connectors and laser devices, thereby reducing the total system cost. This makes multi-mode fiber the ideal choice for short distance under 500 meters from transmitter to receiver (or the reverse).

2.4.5 Droop

Droop will effect performance and it is experienced when a critical distance is exceeded. Factors that affect it are the laws of physics, the speed of light in fibre, the link data rate and the available buffering within the sending and receiving devices. Droop begins when a links distance reaches a point where the time the light takes to make one round trip on the link equals the time it takes to transmit the number of bytes that fit in the receiver's buffer.

In Figure 2-17 we show an ESCON estimate of how the effective data rate decreases as the path length increases. At a distance of 9 km, performance begins to decrease precipitously. This data point is referred to as the distance data rate droop point.

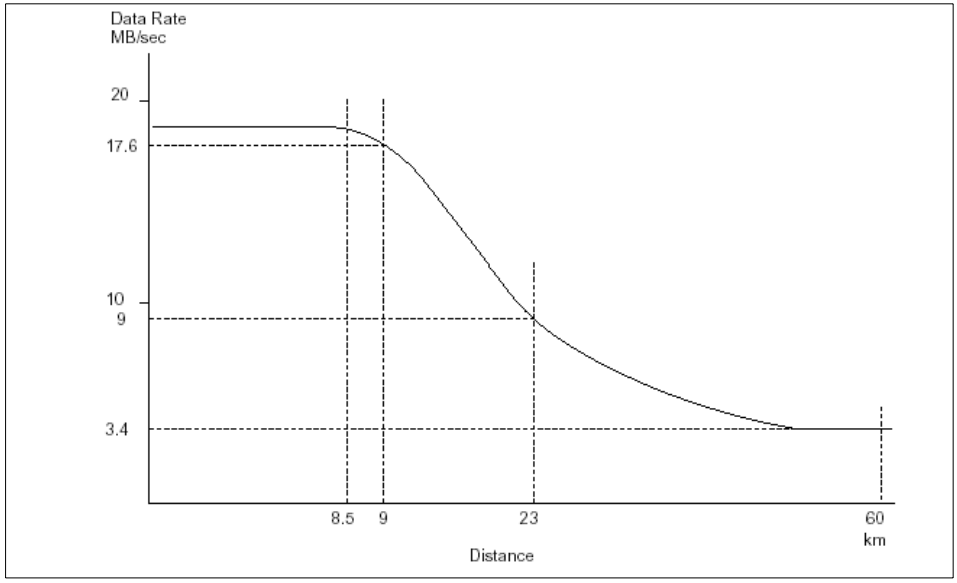


Figure 2-17 ESCON droop example

2.4.6 Latency

Within the SAN arena, longer distances introduce other factors to consider in the SAN design, one of which is latency. A key contributor to latency is distance, latency increases with distance since the time for the signal to travel the longer links has to be added to the normal latency introduced by switches and/or directors. This discussion leads to time-outs and buffer credits, and these should

allow for increased travel times. Latency may or may not be a problem, it is that ability of the fabrics components to tolerate it that will cause issues. The fabric components, applications and operating systems ability to tolerate it will determine its effects.

2.4.7 Sizing

Sizing will be site specific based on the number of channels that need to be sent from location to location. The number of fibres needed will be influenced by whether protection is needed and the topology deployed.

2.4.8 Hops

The hop count as such is not increased by the DWDM architecture, the DWDM components are transparent to the number of hops. The hop count limit within a fabric is set by the fabric devices (switch or director) operating system and it is used to derive a frame hold-time value for each fabric device. This hold-time value is the maximum amount of time that a frame can be held in a switch before it is dropped or the fabric is busy condition is returned. For example, a frame would be held if its destination port is not available. The holdtime is derived from a formula using the error detect time-out value and the resource allocation time-out value.

The discussion on these fabric values is beyond the scope of this book, however further information can be found in *Designing an IBM Storage Area Network*, SG24-5758-00*,* or *IBM SAN Survival Guide*, SG24-6143-00.

If these times become excessive, the fabric will experience undesirable time outs. It is considered that every extra hop adds about 1.2 microseconds latency to the transmission.

We are therefore saying that the number of hops that can be achieved in a SAN fabric is a function of time. DWDM solutions introduce long fibre distances and increased latencies, and this may well affect the number of hops that can be supported. The amount of data that can be in transit is related to buffer credits (we describe these in “Buffer credits” on page 31). As a working rule of thumb we can say that we should allow 1 buffer credit per 2 km of fibre.

2.4.9 Physical location of repeaters

In the past it has been necessary to deploy repeaters at regular intervals to attain long distances. These would have been in the order of 40 km in many examples. DWDM's make it possible to extend over longer distances with less repeaters or optical amplifiers (distances in the order of 120 km). This means that you do not need convenient points en route to accommodate repeater or optical amplifying kits. Remember it may not be as simple as setting up a repeater every 40 km. It will depend on many factors.

2.4.10 Standards

The International Telecommunication Union (ITU), headquartered in Geneva, Switzerland is an international organization within which governments and the private sector coordinate global telecom networks and services. The ITU Telecommunication Standardization Sector (ITU-T) is one of the three Sectors of the International Telecommunication Union and its mission is to ensure an efficient and on-time production of high quality standards covering all fields of telecommunications. The ITU has standards for communications equipment based on the specific wavelength that lasers operate at (the ITU recommendations for optical interfaces for multichannel systems with optical amplifiers G.692 (10/98)). Their Web site can be found at:

http://www.itu.int

2.4.11 Terminology

In this section we will introduce some of the commonly encountered terminology.

Nanometer

The word "nano" means 10^{-9}, so a nanometer is one billionth of a meter.

A nanometer is a unit of spatial measurement that is 10^{-9} meter, or one millionth of a millimeter. It is commonly used in nanotechnology which is the building of extremely small machines.

The wavelengths of light are measured in nanometers.

Wavelength

The technical definition of a wavelength is the distance measured in the direction of propagation of a light wave between two successive points in the wave that are characterized by the same phase of oscillation. This is expressed as nanometer (nm), where one nm is the equivalent of one one-millionth of a millimeter between the two successive points.

In non-technical terms, lower wavelength values have a longer range, are brighter, and generally have a larger spot size than a higher wavelength value.

For example, a 635 nm laser product is brighter with greater range than a 650 nm laser product when both have the same the output power (as defined below). Combined with the appropriate output power, wavelength is an important consideration when using a laser product outdoors or in brightly lit interior environments.

Bandwidth

Bandwidth (the width of a band of electromagnetic frequencies) is used to mean how fast data flows on a given transmission path and, somewhat more technically, the width of the range of frequencies that a signal occupies on a given transmission medium. Any digital or analog signal has a bandwidth. In optical networks bandwidth is defined as the range of frequencies within which a fiber optic waveguide or terminal device can transmit data or information.

Channel

In telecommunications in general, a channel is a separate path through which signals can flow. In optical fiber transmission using dense wavelength-division multiplexing (DWDM), a channel is a separate wavelength of light within a combined, multiplexed light stream.

In IBM mainframe systems, a channel is a high bandwidth connection between a processor and other processors, workstations, printers, and storage devices within a relatively close proximity. It is also called a local connection as opposed to a remote (or telecommunication) connection.

DWDM

Dense wavelength division multiplexing (DWDM) is a technology that puts data from different sources together on an optical fiber, with each signal carried at the same time on its own separate light wavelength. Using DWDM, up to 80 (and theoretically more) separate wavelengths or channels of data can be multiplexed into a light stream transmitted on a single optical fiber. Each channel carries a time division multiplexed (TDM) signal. In a system with each channel carrying 2.5 Gbps (billion bits per second), up to 200 billion bits can be delivered a second by the optical fiber. DWDM is also sometimes called wave division multiplexing (WDM).

Since each channel is demultiplexed at the end of the transmission back into the original source, different data formats being transmitted at different data rates can be transmitted together. Specifically, Internet (IP) data, Synchronous Optical Network data (SONET), and asynchronous transfer mode (ATM) data can all be traveling at the same time within the optical fiber.

2.4.12 Protocol definitions

In this section we describe some of the commonly encountered protocols.

ATM

ATM (asynchronous transfer mode) is a dedicated-connection switching technology that organizes digital data into 53-byte cell units and transmits them over a physical medium using digital signal technology. Individually, a cell is processed asynchronously relative to other related cells and is queued before being multiplexed over the transmission path. Because ATM is designed to be easily implemented by hardware (rather than software), faster processing and switch speeds are possible. The pre-specified bit rates are either 155.520 Mbps or 622.080 Mbps. Speeds on ATM networks can reach 10 Gbps. Along with Synchronous Optical Network (SONET) and several other technologies, ATM is a key component of broadband ISDN (BISDN).

Gigabit Ethernet

The Ethernet protocol is the worlds most popular LAN (local area networking) protocol. This standard has evolved from the original shared 10 megabit per second technology, developed in the 1970's, to the recently completed Gigabit Ethernet standard, (the first Gigabit Ethernet standard (802.3z) was ratified by the IEEE 802.3 Committee in 1998). Gigabit Ethernet is the newest version of Ethernet and it supports data transfer rates of 1 Gigabit (1,000 megabits) per second.

SONET (OC-3, OC-12, OC-48) and SDH (STM-1, STM-4, STM-16)

SONET and SDH are a set of related standards for synchronous data transmission over fiber optic networks. SONET is short for Synchronous Optical NETwork and SDH is an acronym for Synchronous Digital Hierarchy.

SONET is the United States version of the standard published by the American National Standards Institute (ANSI). SDH is the international version of the standard published by the International Telecommunications Union (ITU).

Fiber Channel

Fibre Channel is a technology for transmitting data between computer devices at a data rate of up to 2 Gbps, or one billion bits per second. A data rate of 10 Gbps has been proposed by the Fibre Channel Industry Association.

Fibre Channel is especially suited for connecting computer servers to shared storage devices and for interconnecting storage controllers and drives. Since Fibre Channel is three times as fast, it has begun to replace the Small Computer System Interface (SCSI) as the transmission interface between servers and clustered storage devices. Fibre channel is more flexible; devices can be as far

as ten kilometers (about six miles) apart if optical fiber is used as the physical medium. Optical fiber is not required for shorter distances, however, because Fibre Channel also works using coaxial cable and ordinary telephone twisted pair.

Fibre Channel offers point-to-point, switched, and loop interfaces. It is designed to interoperate with SCSI, the Internet Protocol (IP) and other protocols, but has been criticized for its lack of compatibility — primarily because (like in the early days of SCSI technology) manufacturers sometimes interpret specifications differently and vary their implementations. Standards for Fibre Channel are specified by the Fibre Channel Physical and Signalling standard, and the ANSI X3.230-1994, which is also ISO 14165-1.

ESCON

ESCON is a 200 Mbps unidirectional serial bit transmission protocol used to dynamically connect mainframes with their various control units. ESCON provides non blocking access through either point-to-point connections or high speed switches, called ESCON Directors. ESCON performance is seriously affected if the distance spanned is greater than approximately 8 km. For instance, measurements have shown that ESCON performance at 20 km is roughly 50 percent of maximum performance. Performance degradation continues as distance is further increased.

FICON

FICON is the next generation bidirectional channel protocol used to connect mainframes directly with control units or ESCON aggregation switches (ESCON Directors with a bridge card). FICON runs over Fibre Channel at a data rate of 1.062 Gbps. One of the main advantages of FICON is the lack of performance degradation over distance that is seen with ESCON. FICON can reach a distance of 100 km before experiencing any significant drop in data throughput.

3

ESS copy solutions at a distance

With the rise of the Internet and the growth of online applications to support e-business, enterprises are increasingly looking for ways to improve availability, avoid outages and minimize the adverse effects on business critical applications when they do occur.

One of ways that you can implement a process for disaster recovery is using enterprise disk copy solutions. IBM TotalStorage Enterprise Storage Server (ESS) provides you with PPRC, XRC, GDPS, and Split mirror backup/recovery which we will cover in the following topics.

However, there are some considerations that you must take into account when you implement a disaster recovery solution.

Those topics include knowing the:

- Types of outages
- Business objectives for disaster recovery
- Business recovery options
- Synchronous versus asynchronous copy, and short versus long distance PPRC, XRC, GDPS, and split mirroring with FlashCopy

3.1 The need for remote copy solutions

In case of the loss of data at one site from any disaster, you need to prepare the method for recovering data.

One of the solutions to easily recover data is by using IBM remote copy solutions for both S/390 and open systems.

3.1.1 Types of outages

So what do we mean by disaster and what are the causes of outages that may need to be taken into account? Disaster often refers to an event having fatal or ruinous results. It generally implies great destruction, hardship, or loss of life.

In Figure 3-1 we show some examples of disasters that can impact your system and lead to an outage.

A/C Failure	Evacuation	Low Voltage	Sprinkler Discharge
Acid Leak	Explosion	Microwave Fade	Static Electricity
Asbestos	Fire	Network Failure	Strike Action
Bomb Threat	Flood	PCB Contamination	S/W Error
Bomb Blast	Fraud	Plane Crash	S/W Ransom
Brown Out	Frozen Pipes	Power Outage	Terrorism
Burst Pipe	Hacker	Power Spike	Theft
Cable Cut	Hail Storm	Power Surge	Toilet Overflow
Chemical Spill	Halon Discharge	Programmer Error	Tornado
CO Fire	Human Error	Raw Sewage	Train Derailment
Condensation	Humidity	Relocation Delay	Transformer File
Construction	Hurricane	Rodents	UPS Failure
Coolant Leak	HVAC Failure	Roof Cave In	Vandalism
Cooling Tower Leak	H/W Error	Sabotage	Vehicle Crash
Corrupted Data	Ice Storm	Shotgun Blast	Virus
Diesel Generator	Insects	Shredded Data	Water (Various)
Earthquake	Lighting	Sick building	Wind Storm
Electrical Short	Logic Bomb	Smoke Damage	Volcano
Epidemic	Lost Data	Snow Storm	

Figure 3-1 Types of disasters

Additionally, and although not a disaster, you sometimes need to shutdown your system for hardware maintenance, application update, and periodical backup. Although this is done at periods of low activity in your system, this is still an outage that needs to be planned for.

3.1.2 Outage management

For planned outages you must balance the cost of delaying the change with the cost of shutting down the system to make the change. You must also balance the cost to the business of avoiding the failure against the cost of the failure of an unplanned outage.

Planned outages

Some companies simply can't tolerate any down time, and will compromise, by scheduling a planned outage for a time, when the fewest users will be affected. This means that necessary or application change is postponed until such a time when the system can be offline.

More importantly, in today's competitive and changing business environment, the cost of delaying change can be the most critical of IT business concerns. Business requirements to improve or modify critical systems for business advantage features, are often delayed, waiting for the official change window.

Sometimes these needs can be resolved with an alternative system or mirrored data. This occurs when business critical applications like online transaction processing (OLTP) are required 24 hours a day, 7 days a week.

Responsiveness to business needs is a trade-off against high availability requirements.

Unplanned Outage

Because of the panic nature of this outage, a great potential exists that significant delays may be caused by errors that occur when people are trying to get the service back up. What is the impact of lost or deferred business? What is the cost of the employees who are sitting idle?

For planned and unplanned outages, an IT planner needs to improve the IT infrastructure to one that is highly available and responsive to achieve continuous availability.

3.1.3 Business objectives for disaster recovery

The following objectives must be determined before you select how to prepare for disaster. And you had better establish a Recovery Time Objective (RTO), Recovery Point Objective (RPO), Network Recovery Objective (NRO) very early on for successful disaster recovery.

Recovery time objective

This question asks you how long you can afford to be without your system. Of course, it depends on your business types. Generally, the cost of a disaster recovery solution is in inverse proposition to backup strategy cost.

The following table is an example of recovery time that a certain government is guiding some limitation time to financial companies till recovery according to the impact of business.

In Table 3-1 we show an example of recovery time.

Table 3-1 Example of recovery time

Company	Recovery Time (hours)
Banks, Securities	2.0
Cards	4.0
Insurances	10.0

Recovery point objective

This objective asks you how much data you can afford to lose when it is recovered. You can decide whether you use synchronous or asynchronous remote copy solutions with performance factors. It may not mean that you actually lose data, but at what point will you be able to start your critical applications, while in the meantime you are recovering the not-so critical data.

Network recovery objective

This asks you how much of the network must be restored after the disaster. Will it need to be all or just enough to get along?

3.1.4 Business recovery options

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.

In the following sections we describe different recovery solutions and define them as different tiers.

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 tier 1, an installation would need to develop a contingency plan, back up required information and store it in contingency storage (an off-site location), 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 3-2 we show a tier 1 recovery solution.

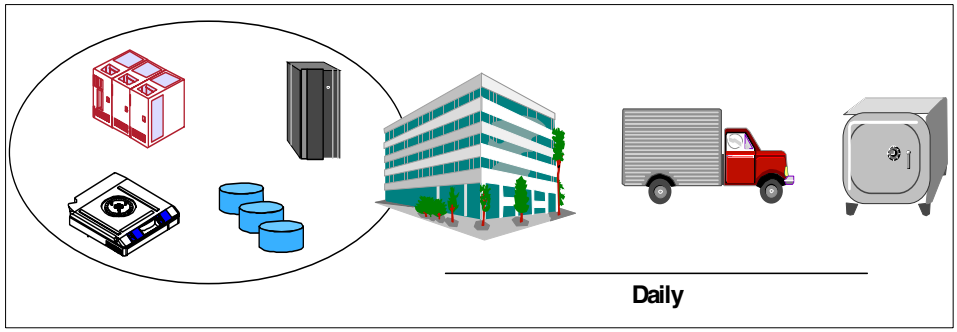


Figure 3-2 Tier 1 recovery solution

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 3-3 we show a tier 2 recovery solution.

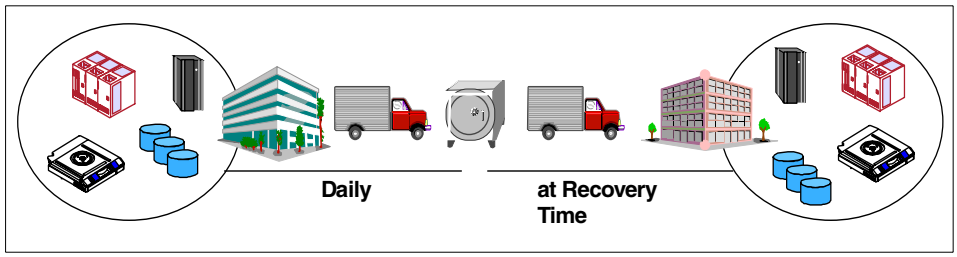


Figure 3-3 Tier 2 recovery solution

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 3-4 we show a tier 3 recovery solution.

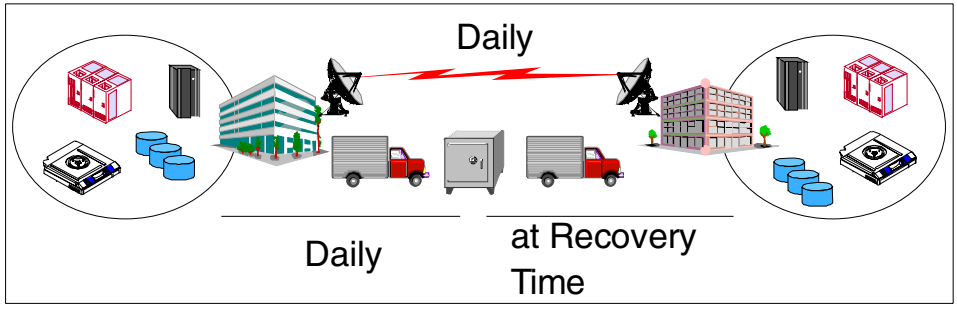


Figure 3-4 Tier 3 recovery solution

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 3-5 we show a tier 4 recovery solution.

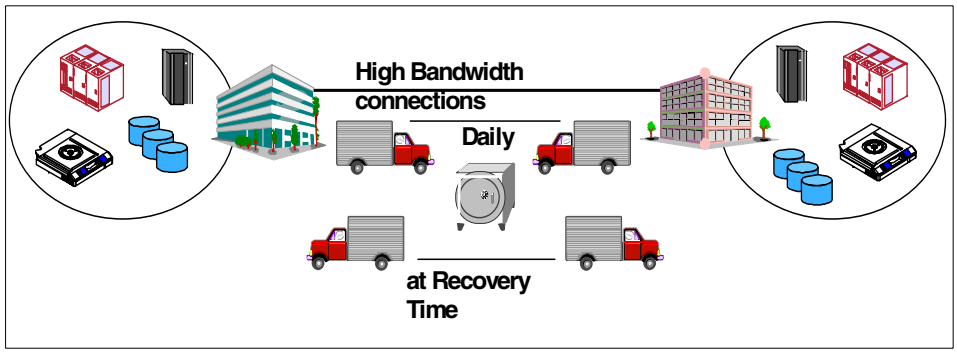


Figure 3-5 Tier 4 recovery solution

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 3-6 we show a tier 5 recovery solution.

Figure 3-6 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 3-7 we show a tier 6 recovery solution.

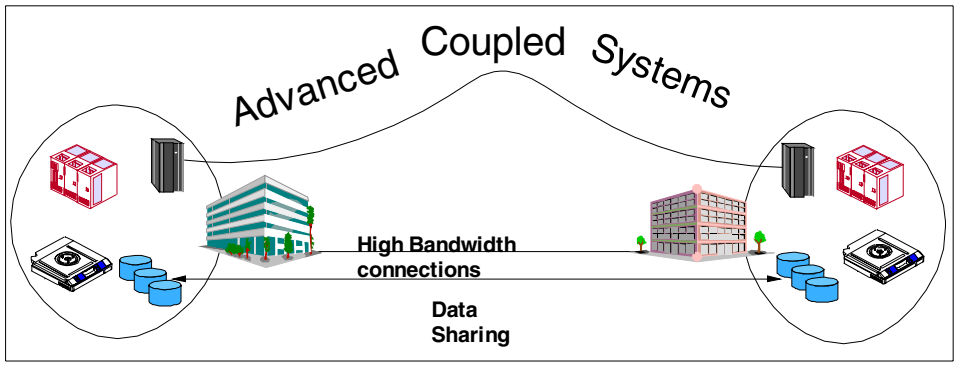


Figure 3-7 Tier 6 recovery solution

3.1.5 Business impact/cost analysis for distance and copy method

Obviously, there will need to be a business case presented that performs some analysis on the impact versus cost.

We detail some of the considerations here.

Short versus long distance

In Table 3-2 we show some considerations for either a short or long distance solutions.

Table 3-2 Short versus long distance

Short Distance (without DWDM)	Long Distance (with DWDM)
Less expensive Single staff can support both sites Clustering of servers easy Logistical consideration minimized No need of DWDM Max. 14 km (16 km by RPQ)	More expensive Need staff at secondary site Coverage against regional disasters Need DWDM for fast response Performance consideration needed

Synchronous versus asynchronous copy

In Table 3-3 we show some of the considerations for synchronous (SYNC) versus asynchronous (ASYNC) copy solutions.

Table 3-3 Synchronous versus asynchronous copy

Synchronous Copy	Asynchronous Copy
Use when no data loss is required Use when response time impact is acceptable Use for log mirroring Minimize database mirroring for long distance Use DWDM for long distance	Data loss possible at the time of failure Know well recovery ways for loss of data Use when long distance is required Use for very large scale applications

3.1.6 Summary

Remote copy solutions may not be easy to implement, and may be expensive to build up from scratch. So you have to consider many things before you start to purchase equipment and other items, and you have to implement it methodically to hopefully avoid errors.

In order to implement the correct remote copy solution, you must understand the considerations in terms of outage management, what the business objectives for disaster recovery are, what the business recovery options are, what synchronous/asynchronous copy is, and what short and long distance is.

3.2 Peer-to-Peer Remote Copy (PPRC)

PPRC is an established data mirroring technology that has been used for many years in mainframe environments. It is used primarily to protect an organization's data against disk subsystem loss or, in the worst case, complete site failure.

It is a synchronous protocol that allows real-time mirroring of data from one Logical Unit (LUN) to another LUN. LUNs can be in the same IBM TotalStorage Enterprise Storage Server or in another ESS located at another site some distance away.

PPRC is application independent. Because the copying function occurs at the disk subsystem level, the application has no knowledge of its existence.

The PPRC protocol guarantees that the secondary copy is up to date by ensuring that the primary copy will be written only if the primary receives acknowledgement that the secondary copy has been written.

More in-depth details about PPRC are available in *Implementing ESS Copy Services on S/390*, SG24-5680, and *Implementing ESS Copy Services on UNIX and Windows NT/2000*, SG24-5757.

Figure 3-8 shows the sequence of events.

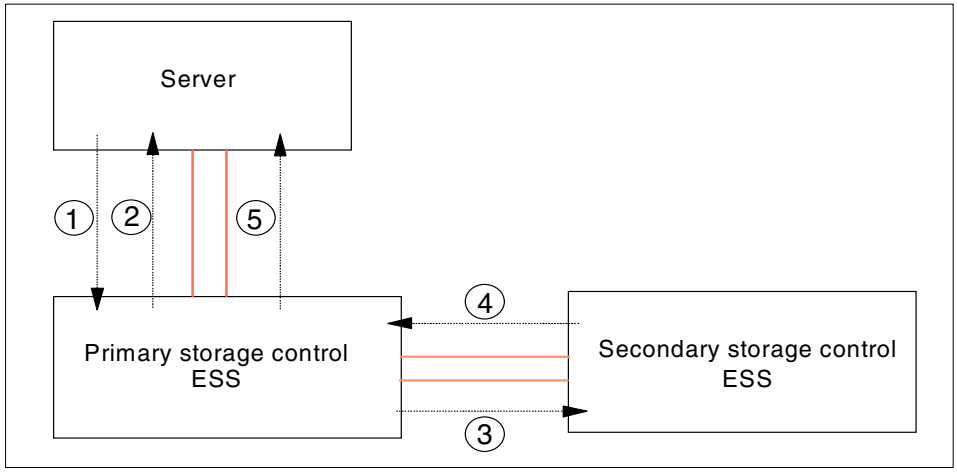


Figure 3-8 PPRC write cycle

This is what happens at each stage:

1. The server requests a write I/O to the primary ESS. The write is staged through cache into Non-volatile storage (NVS).
2. Once the data has been transferred to the ESS's cache and NVS, channel end status is issued to the server.
3. PPRC dispatches the write over an ESCON channel to the secondary ESS's cache and NVS.
4. The secondary site ESS signals "write complete" to the primary site ESS when the updated data is in its cache and NVS.
5. When the primary site ESS receives the "write complete" from the secondary site ESS, it returns device end status to the server.

Destage from cache to the disk drive modules on both the primary and secondary site ESS is performed asynchronously.

If acknowledgement of the remote write is not received within a fixed period of time, the write is considered to have failed, and is rendered ineligible for destage to disk. At this point, the application receives an I/O error, and in due course, the failed write I/O is "aged-out" of each NVS.

3.2.1 Planning for PPRC on an ESS

The following section describes the important areas you should consider when planning for PPRC on an IBM TotalStorage Enterprise Storage Server.

Hardware and software requirement

PPRC is possible only between Enterprise Storage Servers. Other disk storage units that support PPRC can also communicate to the same type of unit only. But the servers are independent on PPRC, which means you can use any servers including xSeries.

You need to have purchased the PPRC feature and PPRC capable microcode activated on all ESS units that will be used for PPRC.

PPRC operates at a volume level from one LSS to another LSS. That means you need to have the target volumes available on the secondary ESS and you need to identify the LSSs where the primary and secondary volumes are.

ESCON connections have to be configured between the units. There can be up to eight ESCON links between the subsystems. A primary ESS can communicate to up to four secondary ESS. A secondary ESS can be connected to any number of ESS primary subsystems. You will need to purchase ESCON cables and possibly some other equipment depending on the distance between the primary and the secondary ESS.

Physical ESCON connections between the ESSs can be direct fiber optic connections, or through ESCON Directors. If the application and recovery site are far removed from each other, then you can use channel extenders such as a a pair of DWDMs.

If the StorWatch ESS Specialist Copy Services Web browser interface is used to manage PPRC, Ethernet connections are required between the ESS subsystems.

The secondary volume must have the identical track capacity and number of tracks per cylinder, and the same or larger volume capacity as the primary.

If you plan to use CLI, install the Java Developers Kit (JDK) on the machines that will run the CLI commands.

Resource planning

When planning your secondary ESS volume layout for PPRC, optimize your disk capacity. It is important to realize that the capacity needed on the secondary ESS for disaster recovery may not have to be initially as large as the primary ESS. A disaster recovery plan (DRP) requires significant investment financially in terms

of technology, people, and process. Every company will be different, but the I/T components of a disaster recovery plan are essentially driven by the applications and data you require for business continuity, should a disaster occur. Some applications and data will be more critical than others. An organization will typically require its core business systems to be available in a short time, whereas less critical systems quite possibly could be restored over a number of days.

Bearing in mind that the disk space you need for PPRC secondary volumes is real disk space, size your secondary ESS based on your critical business requirements, possibly with some headroom for applications of intermediate importance. Create PPRC pairs for the critical data so that is copied in real time. Then, if a disaster happens, you will have the core systems available on the secondary copies. After the initial recovery priorities have been handled, you can add more disk ranks for the applications of lower importance and restore them from tape.

Data consistency considerations

In any recovery situation, including disaster recovery scenarios, you may be exposed to something called lost writes. These are the “in-flight” transactions that have not been committed from memory to the ESS’s non-volatile storage (NVS). You should expect that a few uncommitted transactions will be lost. On the other hand, data that was transferred to the ESS and confirmed back as written into the NVS (or the secondary ESS in case of PPRC), will be destaged to disk.

Invariably, a host server will check its file systems after recovering from a crash. At a DR site, the host servers may be operational when the primary site fails. So it is important to perform a full file system check on all PPRC secondary volumes before you start using them. Of course, rebooting the servers will achieve the same result.

When your database restarts, normal database recovery commences and any partially committed transactions will be rolled back.

Test plan and disaster recovery plan

DRP is complex — nothing can understate the importance of rehearsals and testing your environment. You only get one shot at getting it right when a real disaster hits. Carefully set up your PPRC tasks for establishing and terminating pairs. Ensure that they are well tested and documented. Prepare your documentation as if it were intended for someone else; you may not be around when a disaster strikes.

Make sure you understand any operating system specific issues related to bringing your PPRC secondaries online.

3.2.2 Distance configurations

In this section we present the most common ESCON distance configurations regarding to PPRC connections between two ESSs.

ESCON configuration up to 3 km

Figure 3-9 shows a direct point-to-point connection between a primary and a secondary ESS. Because standard ESCON adapters are used in the ESS, this connection can only be a multi-mode connection. Depending on the fibers you use, the maximum distance can be 2 km (if 50/125 µ fibers are used) or 3 km (if 62.5/125 µ fibers are used).

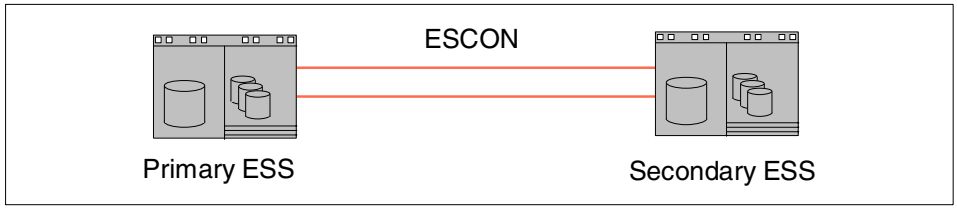


Figure 3-9 Point-to-Point PPRC configuration

ESCON configuration up to 6km

Figure 3-10 shows a connection through an ESCON Director. In terms of distance, there is no difference between a static and a dynamic connection through an ESCON Director. The maximum distance between each ESS and the ESCON Director can be 2 km (50/125 µ fiber) or 3 km (62.5/125 µ fiber), resulting in a maximum distance between both ESSs of 4 km or 6 km, respectively.

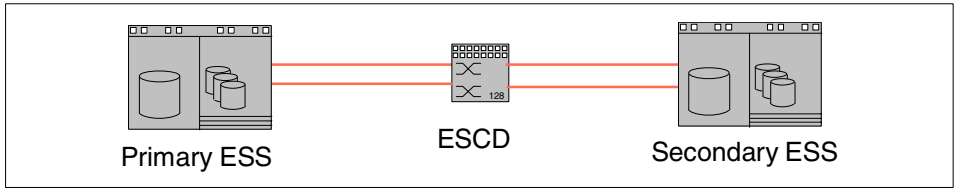


Figure 3-10 Configuration with one ESCON director

ESCON configuration up to 26 km

Figure 3-11 illustrates a configuration with two ESCON Directors. On such configurations, only one of the ESCON Directors can be configured for a dynamic connection. Distances of up to 26 km between the two ESSs can be realized with this configuration.

The distance between an ESS and an ESCON Director is 2 km (50/125 µ fiber) or 3 km (62.5/125 µ fiber).

The distance between the two ESCON Directors can be 2 km for 50/125 µ fiber, 3 km for 62.5./125 µ fiber, or 20 km for 9/125 µ fiber.

The 9/125 µ fiber requires an ESCON-XDF adapter installed on the ESCON Directors.

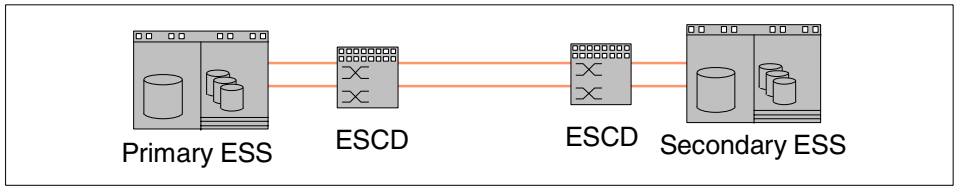


Figure 3-11 Configuration with two ESCON directors

ESCON configuration with single pair of DWDM

The configuration in Figure 3-12 allows a maximum distance of 53 km. The DWDM connects one side to a standard ESCON adapter and the other side to a pair of dark fibers, that are optical fibers provided by telephone companies or other providers. The distance between each DWDM can be up to 50 km. So, in such a configuration, it is possible to have a distance of 53 km between the primary and the secondary ESS.

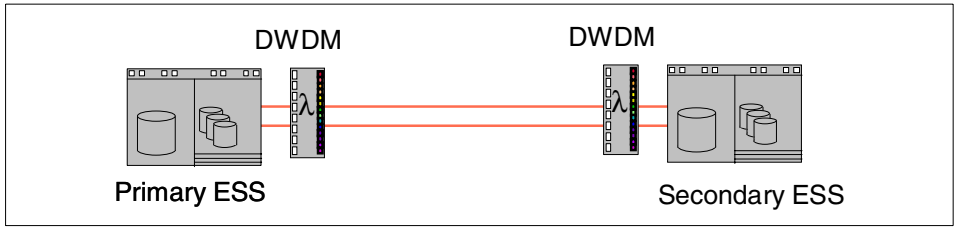


Figure 3-12 Configuration with DWDM

ESCON configuration with two pairs of DWDM

The configuration in Figure 3-13 allows a maximum distance of 103 km. The DWDM connects one side to a standard ESCON adapter and the other side to a pair of dark fibers, that are optical fibers provided by telephone companies or other providers. The distance between each DWDM can be up to 50 km. So, in such a configuration, it is possible to have a distance of 103 km between the primary and the secondary ESS.

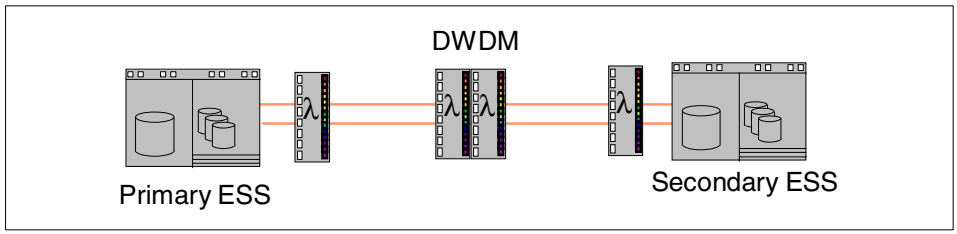


Figure 3-13 Configuration with four DWDMs

3.2.3 Connectivity on the ESS

As with other PPRC implementations, you can establish PPRC pairs only between storage control units of the same type, which means that you can only connect an ESS to another ESS. ESCON links between ESS subsystems are required.

There have been many enhancements to the way two ESS control units communicate over ESCON links compared to the PPRC implementation on an IBM 3990 Model 6. The ESCON protocol is streamlined, less handshaking is done, and larger ESCON frames are transmitted between two ESSs. These enhancements now allow an extended distance between two ESSs of up to 103 km, when using DWDM.

Up to 32 ESCON links are supported between two ESS storage subsystems. The local storage control is usually called primary storage control if it contains at least one PPRC source volume, while the remote storage control is called secondary storage control if it contains at least one PPRC target volume. A storage control can act as primary and secondary at the same time if it has PPRC source and target volumes. This mode of operation is called bi-directional PPRC.

A primary ESS can be connected to up to four secondary ESS storage subsystems as shown in Figure 3-14.

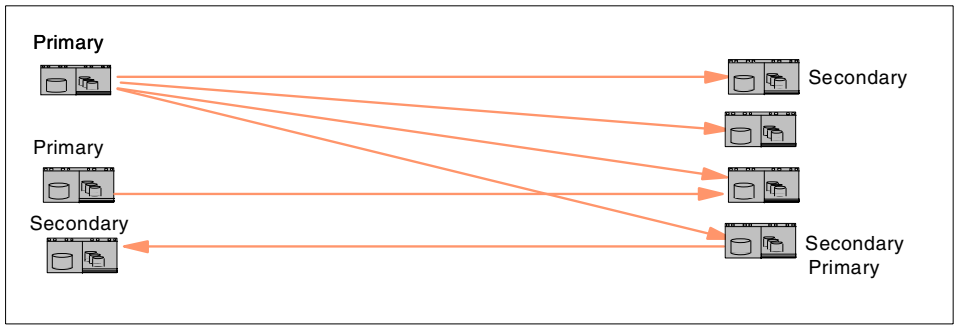


Figure 3-14 PPRC connection options with IBM ESS

A secondary ESS can be connected to as many primary ESSs as there are ESCON links available.

PPRC restrictions

There is a limitation of a maximum 2000 volumes per Copy Services server. This number includes all the primary and secondary PPRC volumes *and* all the source and target FlashCopy volumes.

3.3 Extended Remote Copy (XRC)

This section provides an overview of Extended Remote Copy (XRC) on the IBM TotalStorage Enterprise Storage Server (ESS). You can get more details from *Implementing ESS Copy Services on S/390*, SG24-5680.

XRC is the asynchronous remote copy solution offered on the ESS. It is a combined hardware and software solution that offers the highest levels of data integrity and data availability in a disaster recovery, workload movement, and disk migration environment.

3.3.1 Overview

XRC is a software centric remote copy implementation. A DFSMSdfp component called System Data Mover (SDM) will copy writes issued to primary volumes by primary systems, to the secondary devices. Although the main XRC implementation consists of host resident software, special XRC support is required in the ESS that attach the XRC primary volumes.

When ESS is used as the primary storage subsystem, XRC now also supports unplanned outages. ESS will maintain a hardware bitmap of the tracks changed on primary volumes by the primary systems. If an outage occurs, only changed tracks need to be copied to the secondary volumes when the connection between the ESS and the SDM is re-established, and by this a full resynchronization of the volumes is avoided.

The SDM will also exploit new ESS command control words (CCW) for improved performance. When ESS is running in XRC toleration mode, no hardware bitmap is maintained, and the XRC support is like the support in 3990-6 (XRC version 2).

Figure 3-15 represents a logical overview of the XRC components and the basic data flow.

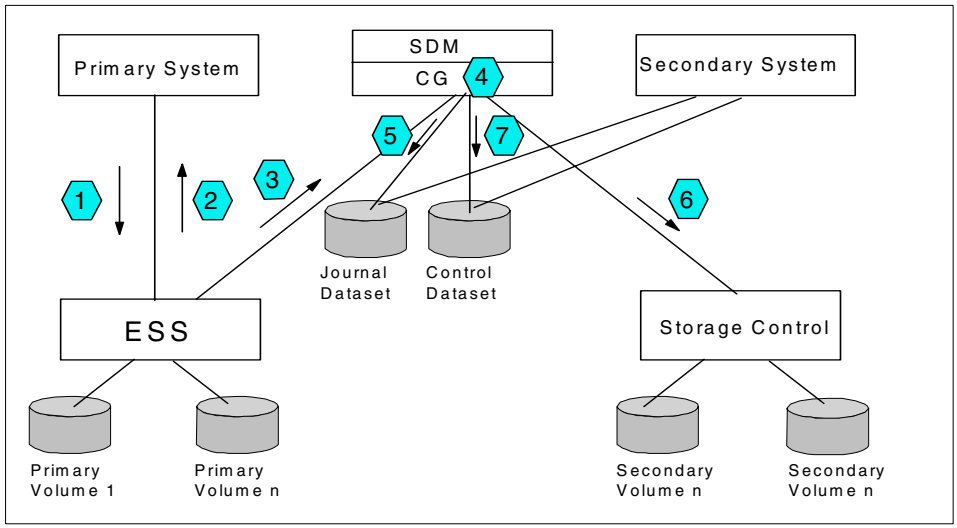


Figure 3-15 XRC data flow

XRC is implemented in a cooperative way between the ESSs on the primary site and the DFSMSdfp host system component called System Data Mover (SDM):

1. The primary system writes to the primary volumes.
2. The application I/O is signaled complete when the data is written to primary ESS cache and NVS, that is, channel end and device end are returns to the primary system.
3. The ESS groups the updates into record sets which are asynchronously off-loaded from the cache to the SDM system. As XRC uses this asynchronous copy technique, the performance impact on primary applications is minimal.

4. The record sets, perhaps from multiple primary storage subsystems, are processed into consistency groups (CGs) by the SDM. The CG contains records that have their order of update preserved across multiple LCUs within an ESS, across multiple ESSs and across other storage subsystems participating in the same XRC session. This preservation of order is absolutely vital for dependent write I/Os, such as databases and logs. The creation of CGs guarantees that XRC will copy data to the secondary site in real time with update sequence integrity for any type of data.
5. When a CG is formed, it is written from the SDM real storage buffers to the journal data sets.
6. Immediately after the CG has been hardened on the journal data sets, the records are written to their corresponding secondary volumes. Those records are also written from SDM's real storage buffers. Because of the data in transit between the primary and secondary sites, the currency of the data on secondary volumes lags slightly behind the currency of the data at the primary site.
7. The control data set is updated to reflect that the records in the CG has been written to the secondary volumes.

3.3.2 System Data Mover

System Data Mover (SDM) is part of DFSMSdfp and must have connectivity to the primary volumes and to the secondary volumes. When primary systems write to the primary volumes, SDM manages the process of copying those updates to the secondary volumes.

Only one XRC session can be in effect per OS/390 system, but multiple instances of SDM (on separate OS/390 images) are possible. Each SDM will have one XRC session, being responsible for a group of volumes. SDM maintains update sequence consistency for the volumes participating in the XRC session, across LCUs in the ESS and across ESSs (and other primary storage subsystems which support XRC).

The SDM for XRC operates in two system address spaces, ANTAS000 which is automatically started during IPL, and ANTAS001, which is started by the XSTART TSO command. These address spaces are unswappable.

ANTAS000 handles TSO commands that control XRC. If this address space is cancelled, it is automatically restarted with no impact on XRC operations. If the address space for some reason is not automatically restarted, you can submit a job to restart it.

ANTAS001 manages the movement from primary to secondary volumes. It manages the journal data sets and controls the application of updates to secondary volumes. If this address space is cancelled, the XRC session for this SDM will be terminated.

3.3.3 Consistency Groups

The Consistency Group (CG) contains records that have their order of update preserved across multiple LCUs within an ESS, across multiple ESSs and across other storage subsystems participating in the same XRC session.

Time-stamping process

Maintaining the update sequence for applications whose data is being copied in real time is a critical requirement for applications that execute dependent write I/Os. If data is copied out of sequence, serious integrity exposures could render the recovery procedures useless. XRC uses special algorithms to provide update sequence consistency for all data. The starting point for maintaining update sequence integrity is when a record is first written by an application system to a primary volume of an XRC-managed pair. When the record is written, the LCU maintains the data (including the time-stamp information) and transfers it to SDM along with other updated records. This process is shown in Figure 3-16.

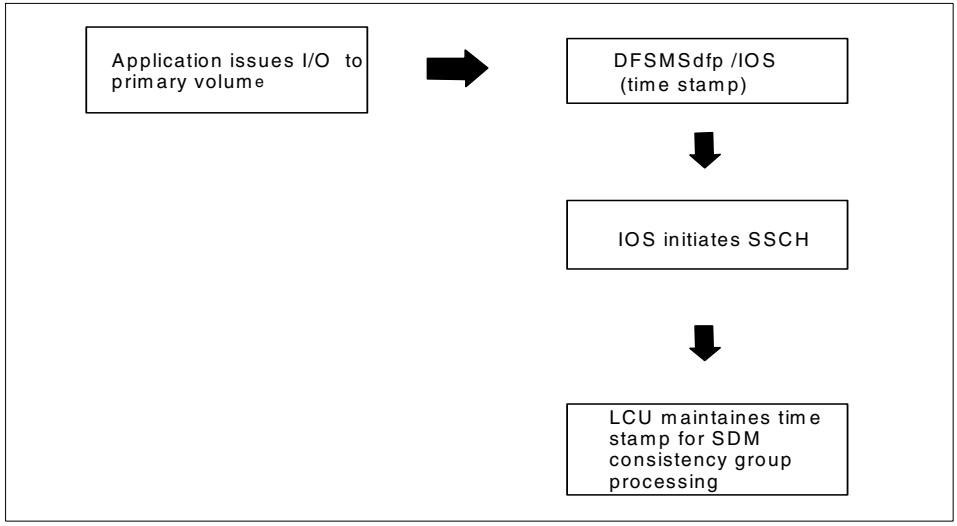


Figure 3-16 XRC time-stamping process

Whenever an XRC-managed pair of volumes is established by the XADDPAIR command, all attached primary systems are informed of their duplex status. Therefore, DFSMSdfp can be selective when performing the time-stamping process.

Only those records that are being written to XRC primary volumes are time stamped for future analysis by SDM. Identification of records that require this special treatment is a result of the pack change interrupt, which is issued when the volumes come under XRC control after the XADDPAIR command is executed. The pack change interrupt allows all attached hosts to update their internal control blocks to reflect the XRC status of the volumes, and therefore provides an efficient way for DFSMSdfp to identify which records should be time stamped and which should not.

The time-stamping code is an extension of the IOS SSCH code, so it applies to all data. Deferred writes held in main storage buffers for database applications are processed in SSCH order. This ensures that the time stamping process delivers update sequence integrity support accurately and efficiently.

XRC update sequence consistency example

For those of you who want a more detailed explanation of how the consistency groups are created, we have included the example shown in Figure 3-17.

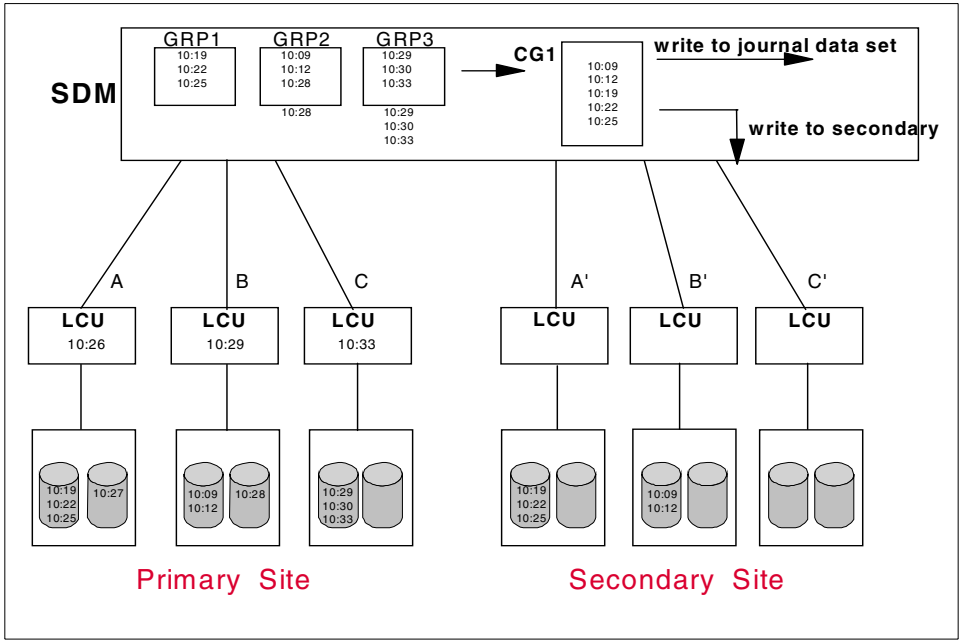


Figure 3-17 Creation of contingency group

This shows a configuration with six primary volumes attached to three LCUs at the primary site, and six secondary volumes attached to three LCUs at the secondary site. The SDM system has connectivity to the three LCUs at the primary site and the three LCUs at the secondary site. We assume a one-for-one relationship between the LCUs and the primary volumes at the primary site and the LCUs and secondary volumes at the secondary site (A, B, C, copied to A', B', C', respectively).

The applications running on the primary system update records on the primary volumes at certain times. In our example, the first volume attached to LCU(A) at the application site will have three records updated at 10:19, 10:22, and 10:25. Similarly, the times recorded for the other five volumes signify the time when their records will be updated (ignore the secondary site disk for the time being).

Two record updates are highlighted on two volumes at the primary; we will use these two records as our dependent writes. The first record is the log update on the second volume attached to LCU(A), which takes place at 10:27. The second record is the database update on the second volume attached to LCU(B), which takes place at 10:28.

Now we define a point in time when SDM reads the updated records from each primary LCU. Notice the time in each of the three LCUs at the primary site. We assume that this is the time that SDM will read the records that have been updated from that LCU.

So, at 10:26, SDM will read updates from LCU(A); at 10:29, from LCU(B); and at 10:33, from LCU(C). We will look at each of the LCUs individually and describe the record transfer to SDM and the CG processing that takes place when all LCUs have sent their updates to SDM.

SDM reads from LCU(A) at 10:26

At 10:26, SDM will read three records from the LCU(A) — the records written at 10:19, 10:22, and 10:25. SDM will not read the highlighted log record, because it has not been written at this point in time (the log will be written at 10:27). We are deliberately stress testing SDM by creating a scenario that could easily result in copying data out of sequence. This group of records will be stored in SDM' s real storage buffers and is designated as GRP1 on the diagram. SDM will not write this data yet, because it must determine whether the other LCUs have dependent record updates that are included in the session.

SDM reads from LCU(B) at 10:29

At 10:29, SDM will read three records from LCU(B) — the records written at 10:09, 10:12, and 10:28. Note that the highlighted database update (10:28) is included in this group of records, which is described as GRP2 in the diagram. This poses a potential problem because SDM has read the database update but not the log update. If SDM simply writes data as it receives it, the serious data integrity exposure described previously would prevail. SDM's ability to avoid these exposures becomes clear as we proceed with this example.

SDM reads from LCU(C) at 10:33

At 10:33, SDM will read three records from LCU(C) — the records written at 10:29, 10:30, and 10:33. This group of updated records is called GRP3. At this stage, SDM has received responses from all three LCUs that have volumes in the XRC session. The next phase of providing the remote copy function is to perform the update sequence integrity check for all of the updated records read from all of the primary storage subsystems in the XRC session.

SDM creates CG

SDM now has three groups of records that it uses to create a CG. The CG is the unit of data transfer to the journal data sets at the secondary site. It contains records that SDM has determined can be safely written to the secondary site without risk of out-of-sequence updates. To compile the CG, SDM uses all of the record groups read from the LCUs as input. In our case, we have three record groups that SDM uses as input to produce a single CG output.

SDM compares the maximum time stamp for each individual record group (GRP1 = 10:25, GRP2 = 10:28, GRP3 = 10:33) and selects the smallest value (10:25) to calculate which records should be included in the CG. SDM calculates that all records written at or before this smallest value time can be written together to the remote site journals as a CG.

In our example, five records (10:09, 10:12, 10:19, 10:22, and 10:25) qualify for inclusion in the first CG — CG1. Using this calculation, SDM ensures that dependent write records are not written "ahead" of time as demonstrated by the fact that the database update at 10:28 has not been included in the CG. This is vital because the log update (10:27) is located behind a different LCU(A) and has not yet been read into SDM's real storage buffers for inclusion in the "consistency group creation algorithm".

Now we see the importance of using the maximum time stamp in each individual record group, and selecting the minimum of all of the maximums across record groups as the CG "upper limit" time stamp.

SDM has “retained” the database update along with other possibly dependent-write-sensitive records in its real storage until it calculates that it is safe to include them in a future CG. In the diagram, these four records are listed beneath their respective original record groups (10:28 in GRP2 and 10:29, 10:30, and 10:33 in GRP3).

SDM writes CG to journal data sets

Having created a group of records that can be safely written together, SDM then writes this group to the journal data sets at the secondary site. Therefore, SDM can harden the updates as quickly as possible in one I/O to a single data set.

SDM writes the updates to the secondary volumes

After the CG has been written to the journal data set, SDM immediately writes out the individual records to their appropriate secondary volumes. This transfer takes place from SDM’s real storage buffers (the journal data sets are only read during recovery using the XRECOVER command). Updating the secondary volumes could involve several I/Os, because the record updates could be directed to several volumes located behind several LCUs. In this example, the CG comprises five records directed to two volumes attached to two LCUs. The figure illustrates that the five records have been successfully copied to their corresponding secondary volumes (10:19, 10:22, and 10:25 written to the first volume of LCU(A’), and 10:09 and 10:12 written to the first volume of LCU(B’)).

Figure 3-18 concludes the example on update sequence consistency by stepping forward in time and describing how the dependent write updates are copied to the secondary site.

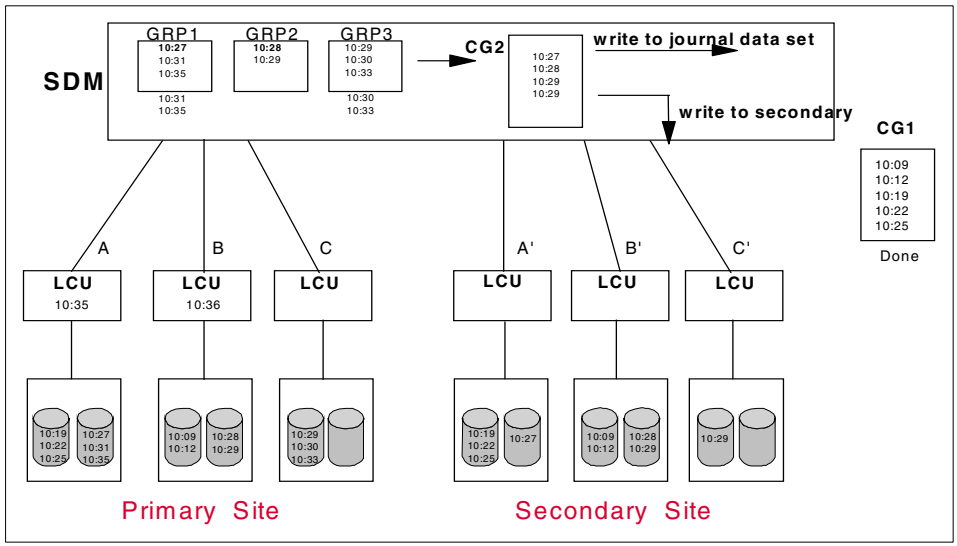


Figure 3-18 Creation of consistency group

SDM copies data in real time, so it constantly transfers updated records from perhaps multiple LCUs in an XRC session. In our example, we simulate moving the “clock” forward in time and tracing the steps taken by SDM in the production of its second CG.

Notice that LCU(A) and LCU(B) at the primary site now have different times indicating when SDM reads the updated records. We proceed as before:

SDM reads from LCU(A) at 10:35

At 10:35, SDM reads three records from LCU(A) (10:27, 10:31, and 10:35). Note that the log update (10:27) has now been read. This group of three records has been designated GRP1 in the diagram.

SDM reads from LCU(B) at 10:36

At 10:36, SDM reads one record from LCU B (10:29). This record can now keep the database update record (10:28) company in SDM real storage (these two records are in GRP2 in the diagram). Remember that the database update record (10:28) was not included in the previous CG and was kept in real storage by SDM along with other “residue” records.

SDM detects NULL response from LCU(C)

Finally, SDM completes its cycle of updated record collection by establishing that the third LCU(C), which also connects volumes in this XRC session, has “nothing to report”. The NULL response must be sent even if the LCU has no updates, because it indicates that the LCU has not suffered an outage and that the record groups retrieved to date from other LCUs can be used to build a valid CG.

SDM creates CG

SDM creates the second CG using the same algorithm as before, and we can see that the two dependent writes (the log update at 10:27 and the database update at 10:28) have been “captured” in this CG.

Notice that the record group transferred from LCU(B) produced a maximum time stamp, which also turned out to be the smallest of the three GRP maximums. Notice also that this value (10:29) enabled all records written before or at this time to be included in the CG — therefore allowing the log update and the database update to accompany each other on the journey to the secondary site when written to the journal data sets.

The diagram shows that when SDM writes the records to the secondary volumes, the relationship between the contents of the log and the database, which is vital for recovery, has been protected. The record residue (10:31 and 10:35 from GRP1, and 10:30 and 10:33 from GRP3) will be processed during the next SDM record group collection cycle.

3.3.4 XRC requirements

There are the software and hardware requirements when using XRC with ESS.

Hardware requirements

XRC hardware requirements are:

- ESA/390 hardware must be used.
- For the primary ESSs, the XRC feature must be enabled. Support is provided by XRC-capable Licensed Internal Code (LIC). For Model E10 and E20 you have to order feature codes according to the total installed capacity on the ESS. For Model F10 and F20 you order feature codes according to the capacity of the OS/390 portion of the ESS.

  The following feature codes are applicable to both F and E models:

  - **1810** up to 0.5 TB
  - **1811** up to 1 TB
  - **1812** up to 2 TB
  - **1813** up to 4 TB

– **1814** up to 8 TB
– **1815** larger than 8 TB

There will be two different LIC levels:

– XRC toleration support (XRC support as in 3990-6)
– XRC exploitation support (hardware bitmap support)

Primary systems must have a common time reference. XRC uses time stamped record information to ensure that updates are not copied out of sequence at the secondary site.

In environments where the primary systems are on different CECs, IBM 9037 Sysplex Timer, or equivalent, is required.

When you have only one primary system, or the primary systems are running in LPARs on the same CEC, the system time-of-day clock is used.

A compatible secondary volume must be available for each primary volume you want to copy. The secondary volume must have the identical track capacity and number of tracks per cylinder as the primary volume. The capacity of the secondary volume can be the same or larger than that of the primary volume.

Software requirements

DFSMS/MVS software support for ESSs with XRC LIC is provided as program temporary fixes (PTFs). There are two types of support:

- XRC toleration support
- XRC exploitation support
- XRC toleration support

With this support, the XRC support is the same as for the 3990-6.

XRC toleration support is provided by DFSMS 1.3 and above, plus the following APARs:

- OW33608
- OW38296
- OW37431
- OW36948
- OW43843

XRC in toleration mode can run on OS/390 systems with:

- ESS transparency support
- ESS toleration support
- ESS exploitation support
- XRC exploitation support

This will support ESSs with XRC exploitation LIC level.

XRC exploitation support is provided in DFSMS 1.3 and above, plus the following APARs:

- OW43315

  XRC in exploitation mode can run on OS/390 systems with ESS exploitation support

You should review the PSP 2105DEVICE for the latest ESS maintenance. We also recommend for you to review the information APARs II08303 and II11778. Those APARs have information about the latest maintenance for copy services.

3.3.5 Planning for a primary and secondary ESS

In the topics that follow we discuss the planning steps needed for successful implementation of primary and secondary ESSs.

Primary ESS

You have to assess the capacity of the primary ESS. In addition to reads and writes issued from the primary systems, the ESS LCU must also be able to handle the reads associated with SDM's off loading process. The data written to primary volumes by primary systems will be grouped into record sets in cache, and those record sets will be off-loaded by the SDM system with the **`read record set`** channel command.

For more information about ESS performance, see the ESS performance white paper at:

`http://www.ibm.com/storage/ess`

The cache requirement should also be evaluated. There is no separate copy in the cache of the data to be off-loaded, the ESS merely builds a logical queue using a directory structure.

When the SDM operation is balanced, meaning that the SDM does not get delayed when retrieving data from the primary ESSs and committing those updates to the journal data sets and secondary volumes, the primary ESS uses minimally more cache than what it normally needs to satisfy primary application performance requirements.

As long as the SDM manages to retrieve the updates before the ESS caching algorithms would have removed them out of cache, minimal additional cache is required.

However, if the primary systems write rate is (temporarily) more than the SDM can absorb, or if there is a (temporary) disruption in the SDM data flow, the cache will be used as a buffer.

Our recommendation is that you evaluate the cache size required to satisfy your primary application needs, and then plan for an equal amount of cache for XRC buffers. This may mean you should install more cache (if possible), or increase your cache-to-backstore ratio by spreading the volumes across multiple ESSs.

Secondary ESS

The secondary storage subsystem can be any subsystem supported by the SDM system, as long as it provides volumes with the same track capacity, the same number of tracks per cylinder, and at least the same capacity as its corresponding primary volume.

If a secondary volume has more capacity than a primary volume, it has to be carefully considered if a copy-back procedure is implemented.

The secondary site needs the same number of XRC volumes as the primary site.

Although it is no requirement to use ESSs as secondary storage subsystems, doing so will give you all the benefits of ESS's performance characteristics. In addition, it also makes it possible to use XRC in a copy-back implementation. With ESSs at the secondary you can also combine XRC and FlashCopy functions.

Ideally the number of LCUs at the secondary site should match the number of LCUs at the primary site to provide configuration symmetry.

The secondary site storage controls must be configured to handle all of the primary site LCU's primary volume writes plus the I/Os to the journal, control and state data sets as a minimum. But they must also be capable to support the I/O activity related to the primary application's requirements in a disaster recovery situation.

3.3.6 Selecting volumes to remote copy

In the topics that follow we detail the volume selection process.

Volume level copy

XRC copies the entire contents of a volume. As a result, all data sets on a primary volume are copied to the secondary volume. This support is for all data types, independent of the applications.

Which data sets are essential for your XRC solution?

Before you start up your XRC environment, you have to identify the data you need on the secondary site. This may differ depending on whether you use XRC for data or workload migration, or for a disaster recovery solution.

In the remainder of this section, we are assuming that XRC is used for disaster recovery.

You should identify the type of data required for a successful recovery. Many installations with an existing secondary site may have completed this task already.

The SYSRES, master catalog, SPOOL volumes, as well as other data sets required to initialize the secondary system (including those used to start up JES, TSO, and RACF) could be copied by XRC.

However, the focus for remote copy is on your application data sets that are updated regularly and are essential for your company. If a disaster occurs, critical application volumes must be made available to your primary application as quickly as possible, depending on the speed of recovery required. XRC offers a single-command recovery strategy, which makes it the fastest method for recovering data at the secondary site. A single XRECOVER command recovers all the volumes in an XRC session at once, and all of the volumes are consistent to a single point in time.

Application system volumes

Different application systems at the primary site may have different priorities. Some of these application systems may not warrant the investment required for remote copy. If only a subset of application systems is to be copied, the volumes belonging to those applications must be identified and included in the XRC configuration. Even if all application systems are required at the recovery site, there still may be volumes that do not need to be copied.

Because applications deal with data sets and not volumes, multi-volume data sets require special attention. Multi-volume data set types include data sets that reside on multiple volumes, striped data sets, and VSAM spheres. If you want a copy of a multi-volumes data set at the secondary location, you therefore have to copy all volumes on which this data set resides.

System volumes

The following data sets should be given special consideration as they might be eliminated from a particular installation's XRC configuration:

Page

Page data sets are of no use during recovery at the recovery site. They are owned by the host application site.

Spool
The spool data sets should be copied to the recovery site if it is necessary for recovery. Copying the spool increases the amount of copy activity to the secondary. If the contents of the spool are easily recreated, exclude spool volumes from the XRC configuration.

Temporary data sets
Volumes containing temporary data sets (that is, those that exist for the duration of a job or step) can be recreated at the secondary by resubmitting jobs. These volumes should be excluded.

SYSRES volume
The change activity against data sets residing on the SYSRES often does not affect the applications. However, some customer installations share the SYSRES volume with active data, so the decision on whether to copy the SYSRES volume can depend on the data stored there. Because XRC copies data at the volume level, all data residing on that volume is copied. If the SYSRES volume is not copied by XRC operations, another method must be used to ensure that a copy is available at the secondary site.

Master catalog and user catalogs
Catalog changes are limited, but they are essential for recovery and therefore they must be considered for XRC copying.

Program libraries
All program libraries that contain modules for XRC applications are essential for recovery and should be considered for XRC copying.

Control data sets
Control data sets, such as RACF, SMS, HSM, and RMM, are essential for recovery and should be considered for XRC copying.

HSM migration level 1 volumes
Consider whether migrated data is required at the secondary site. If not, exclude HSM migration level 1 volumes. If migrated data is required but is not critical, these volumes could be copied daily by other means to reduce the amount of copying XRC has to do as migrations and recalls occur.

3.4 Geographically Dispersed Parallel Sysplex (GDPS)

IBM's Geographically Dispersed Parallel Sysplex (GDPS) is the highest application availability solution available in the marketplace today. GDPS combines software and hardware to provide the means of managing a complete switch of all resources from one site to another automatically providing the highest level of availability available in the industry today.

GDPS can be used to manage all resources required for a set of application(s) in the event of a planned or unplanned requirement to switch the applications from one site to another.

GDPS enhances the remote disk mirroring available with IBM TotalStorage Enterprise Storage Server (ESS). With remote disk mirroring, the time required to switch from one site to another is typically on the order of two to four hours. With GDPS, installations have a total multi-site resource management capability which enables an installation to perform a site switch in under one hour. In the event of an unplanned situation, GDPS, through its "Freeze Triggers", insures all data at the secondary site is I/O consistent to a single point in time, yielding the ability to perform a database RESTART. Further, GDPS automation manages the switch of ALL resources required to run the applications at the remote site. Disk mirroring can only manage disk resident data, GDPS manages processors, LPARS, CF structures, Network resources, etc. Further, GDPS performs the entire site resource switch, and brings up the application environment as well as the application(s) at the remote site automatically, yielding the highest level of availability available in the industry today.

3.4.1 GDPS business option

When a potential failure event occurs GDPS multi-site monitoring, invokes a system-wide data "freeze". GDPS provides several business options to be taken after the cross site data freeze, including the ability of having no data loss. This initially attractive business option becomes a "business impact" trade off that we will further investigate.

Most installations consider use of remote disk mirroring, commonly known as Remote Copy, as the first line of defense against loss of data in the event of a failure. This storage subsystem feature's scope is the data within the specific storage subsystem or group of storage subsystems.

GDPS incorporates the storage subsystem remote copy implementation into a complete multi-site resource management solution. GDPS utilizes IBM's Peer to Peer Remote Copy (PPRC) and IBM's Extended Remote Copy (XRC) architecture that is implemented on several IBM Storage subsystems including IBM TotalStorage Enterprise Storage Server (ESS).

GDPS extends remote disk mirroring by providing, through multi-site monitoring software, a data freeze capability across all volumes on all storage subsystems at the primary production site on the first indication of a failure. GDPS makes use of special I/O commands to affect the cross volume, cross storage subsystem data freeze. In order for a storage subsystem to implement the data freeze, the storage controllers must be capable of responding to these I/O commands. Therefore, when considering a disk subsystem acquisition, the GDPS “data freeze” function imposes considerations which are more significant than just the selection of a particular disk vendor's mirroring implementation. But any vendor’s disk subsystem can be used in a GDPS installation. Most GDPS implementations today have multiple vendor’s disk subsystems installed. Therefore, the GDPS architecture helps enterprises move to an industry wide “open” storage subsystem remote copy implementation.

GDPS data freeze function

The data freeze “trigger” can be generated by a number of different system components, including:

- A specific disk subsystem
- The SYSPLEX facility and Coupling Facility links
- Coupling Facilities (CF) (Future)
- Tape subsystems (Future)
- The fabric — some examples of what is meant by fabric triggers are:
  - Storage subsystem to storage subsystem
  - Host to storage subsystem
  - ESCON Director ports

Therefore, the scope of a GDPS Data Freeze extends across the entire S/390 SYSPLEX. This is not possible with disk mirroring implementations which are confined only to the storage subsystem(s). Triggers outside of the disk subsystem are essential since in the event of a disaster one cannot depend on the disk subsystem to be the first to recognize a failure situation. Further, as CF Mirroring and Tape subsystem mirroring is brought into the GDPS solution, a cross ‘data’ subsystem ‘data freeze’ is required to maintain data consistency across multiple data storage mediums (disk, tape and CF data structures).

When a GDPS trigger event occurs, all data at the secondary site is frozen to the point in time of the first failure event. All secondary site data is preserved thus providing complete data integrity and data consistency across all volumes on all storage subsystems (including in the future tape and CF structures).

The ability to execute a quick database restart at the secondary site in response to a system/resource failure at the primary site, is the principle objective of companies that have implemented GDPS. This secondary site-wide data freeze is exactly what data bases require to be able to invoke a quick data base restart. The Data Freeze capability is the key component in the ‘less than one hour secondary site recovery’ requirement, as it insures that once the infrastructure to support the data base and application(s) at the secondary site is in place via the site switch automation, a restart of the business application(s) is insured.

GDPS provides the installation with three business policy options which determine how to continue processing after a secondary site ‘data freeze’ occurs. The three options are:

- Freeze and Go
- Freeze and Stop
- Freeze, Stop Conditional.

Freeze and go

The business objectives with this option are to provide minimal impact to the primary application environment, while maximizing the ability to perform a business application(s) RESTART at the secondary site. In other words, to get the application(s) back up and running as quickly as possible in the event of an unplanned site switch.

This business policy states that after the data freeze has occurred at the secondary site, GDPS will raise an alert and then permit production to continue at the primary site. In the event of a subsequent “full” disaster and a GDPS invoked site switch, all transaction that completed in the window between the cross site data freeze and the actual site switch are lost. The Cross Site Data Freeze on the first failure (start of a disaster), preserves the cross volume, cross storage subsystem data integrity/data consistency. This provides the ability to perform an application restart when GDPS switches all resources to restart production at site 2.

The cost of implementing this business option is the loss of transactions/data that occurs between the time that the secondary site data is frozen and the time that the site switch is initiated. Any completed transactions/data updates in this window are lost and need to recreated. Some techniques used by various customers to manage lost transactions/data are outlined below.

Freeze and stop

This option will cause all production systems to be stopped after the cross site data freeze.

Alerts are raised to the operations staff and all production is stopped.[1]

Based on the GDPS TM Alert that occurs on the cross site data freeze, the Operations staff must

decide to either:

- Restart production at the primary site and then re-synchronize the remote mirrored disk pairs in both sites
- Switch production to the secondary site.

No transaction/data loss occurs, and the ability to restart the business application(s) at the secondary site is preserved. However, the business trade-off with this GDPS option is availability of the production application(s).

All production is impacted and an outage is taken until the failure is analyzed and an action is taken. Data Freeze triggers can occur for events that might be the first of a series of events that would lead to a full site disaster, but they also might be “Operational” errors.

Freeze, stop conditional

This option is a popular choice. The system performs the data freeze at the secondary site. Then one of two actions are taken.

- Stop all production images if the cause of the data freeze has to do with the primary site or intra-site connectivity.
- Continue production at the primary site if the cause of the cross site data freeze was a failure at the secondary site.3

This approach is a reasonable subset of the ***Freeze and Stop*** scenario. However, as noted above, if an operational failure occurs at the primary site all production application(s) are stopped until some analysis of the failure is completed and an action is taken.

3.4.2 GDPS/PPRC requirements

GDPS with PPRC requires at least he following hardware and software.

Hardware requirements

- Parallel sysplex two sites
- Replicate hardware across two sites
  - Coupling Facility
  - IBM 9037-2 Sysplex Timer
  - Processors
  - ESS with PPRC code
  - Tape

- HMC automation infrastructure

Software requirements

- Z/OS V1R0 or OS/390 V2R6 or higher
- System automation for OS/390 V1R6 or higher
- Tivoli Netview for OS/390 V1.2 higher or Netview 3.1

3.4.3 GDPS/XRC requirements

Hardware requirements

- Production Parallel Sysplex
- System Data Mover Parallel Sysplex
  - Production and SDM parallel sysplex may be the same
  - SDM parallel sysplex may be at any distance
- Replicate hardware across sites for redundancy
  - Coupling Facility
  - IBM 9037-2 Sysplex Timer
  - Processors
  - Disks including ESS with XRC code
  - Tape
  - HMC automation infrastructure

Software requirements

- Z/OS V1R0 or OS/390 V2R6 or higher
- System automation for OS/390 V1R6 or higher
- Tivoli Netview for OS/390 V1.2 higher or Netview 3.1

3.4.4 Summary

IBM TotalStorage Enterprise Storage Server's remote copy solutions bring the Peer-to-Peer Remote Copy (PPRC) architecture as well as Extended Remote Copy forward to the disaster recovery. These ESS features can be used in conjunction with Parallel Access Volume (PAV) and Multiple Allegiance (MA) to provide IT infrastructure flexibility that is required to meet the demands of today's e-business.

GDPS is a multi-site application availability solution that provides the capability to manage the remote copy configuration and storage subsystems, automates Parallel Sysplex operational tasks, and performs failure recovery from a single point of control, thereby improving application availability. GDPS supports all transaction managers, such as CICS TS, IMS TM, and database managers, and is enabled by means of key IBM technologies:

- Parallel Sysplex

- System Automation for Z/OS or OS/390
- IBM TotalStorage Enterprise Storage Server
- XRC
- DWDM

3.5 Split mirroring and FlashCopy

In the following topics we detail split mirroring and FlashCopy.

3.5.1 Split mirroring

PPRC by itself is a solution for protecting data against hardware outages or environmental disasters, but it does not protect data against user or application logical errors. For example, in the case of a user or application error, a hot standby database would be in the same inconsistent state as the live database.

Split mirror backup/recovery functions as a high availability backup/recovery scenario, where the backup is taken on a remote disk subsystem that is connected to an application disk subsystem through PPRC function. Normally, the PPRC connection is suspended (mirror split) and will only be resumed for the resynchronization of the primary and the secondary volumes.

You do not have to stop your applications while the database backup is being taken. In the case of a user or application error, the primary database is available for analysis while a secondary database is recovered to a consistent point-in-time.

Figure 3-19 shows the example of a split mirror backup/recovery configuration. The most current version of relational database has a function that allows it to temporarily freeze and resume all the writes to the database, except MS/SQL database up to now.

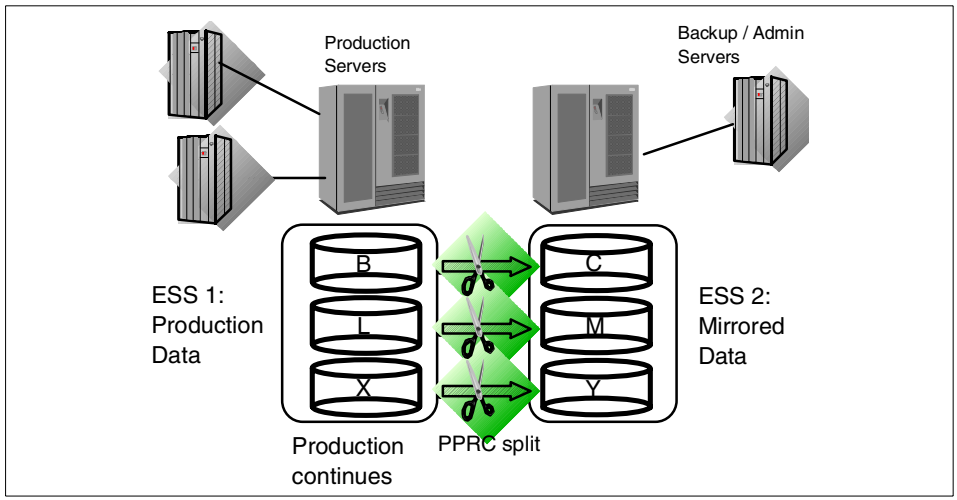


Figure 3-19 Split mirror backup/recovery configuration

You can copy source volumes which are volume B, L, X to volume B', L', X' in ESS1 and volume C, M, Y to volume C', M', Y' in ESS2 by FlashCopy for easy backup or batch job without stopping production online.

3.5.2 FlashCopy

Today, more than ever, organizations require their applications to be available 24 hours per day, seven days per week (24x7). They require high availability, minimal application downtime for maintenance, and the ability to perform data backups with the shortest possible application outage.

The prime reason for data backup is to provide protection in case of source data loss due to disaster, hardware failure, software failure or user errors.

Data copies can also be taken for the purposes of program testing, data mining by database query applications. However, normal copy operations take a long time requiring the prime application to be offline. In addition to the need for 24x7 data processing, it is also necessary to have an instant copy of the data.

FlashCopy allows you to move effectively towards such solutions.

Overview

FlashCopy provides an instant or point-in-time copy of an ESS logical volume. The point-in-time copy functions give you an instantaneous copy, or "view", of what the original data looked like at a specific point-in-time. This is known as the T_0 (time-zero) copy. This is shown in Figure 3-20.

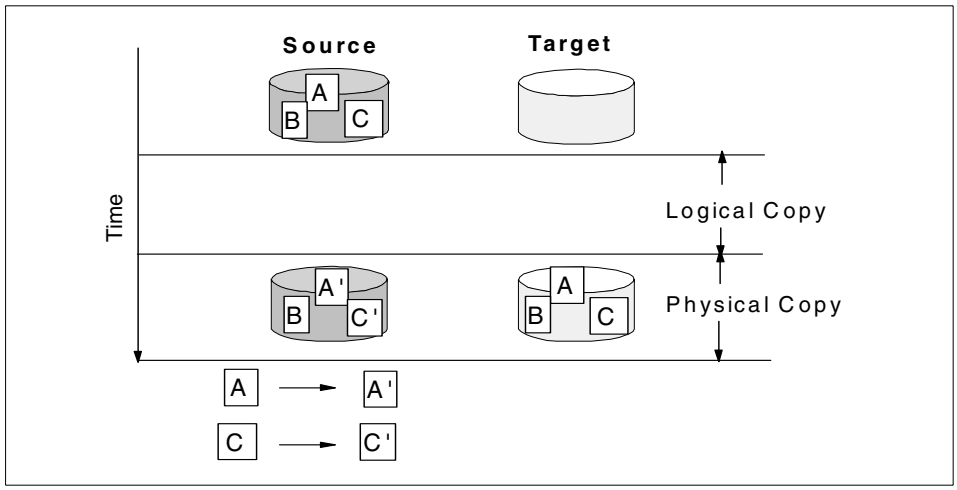


Figure 3-20 FlashCopy

When a FlashCopy is invoked, the command returns to the operating system as soon as the FlashCopy pair has been established and the necessary control bitmaps have been created. This process takes only a few seconds to complete. Thereafter, you have access to a T_0 copy of the source volume. As soon as the pair has been established, you can read and write to both the source and the target volumes.

The point-in-time copy created by FlashCopy is typically used where you need a copy of production data to be produced with minimal application downtime. It can be used for online backup, testing of new applications, or for creating a database for data-mining purposes. The copy looks exactly like the original source volume and is an instantly available, binary copy. FlashCopy is possible only between disk volumes. It also requires a target volume to be defined within the same Logical Subsystem (LSS) as the source volume. A source volume and the target can be involved in only one FlashCopy relationship at a time. When you set up the copy, a relationship is established between the source and the target volume and a bitmap of the source volume is created.

Once this relationship is established and the bitmap is created, the target volume copy can be accessed as though all the data had been physically copied. While a relationship between source and target volumes exists, a background task copies the tracks from the source to the target. The relationship ends when the physical background copy task has completed.

You can suppress the background copy task using the *Do not perform background copy (NOCOPY)* option. This may be useful if you need the copy only for a short time, such as making a backup to tape. If you start a FlashCopy with the *Do not perform background copy* option, you must withdraw the pair (a function you can select) to end the relationship between source and target.

At the time when FlashCopy is started, the target volume is basically empty. The background copy task copies data from the source to the target. The FlashCopy bitmap keeps track of which data has been copied from source to target. If an application wants to read some data from the target that has not yet been copied to the target, the data is read from the source; otherwise, the read is satisfied from the target volume. When the bitmap is updated for a particular piece of data, it signifies that source data has been copied to the target and updated on the source. Further updates to the same area are ignored by FlashCopy. This is the essence of the T_0 point-in-time copy mechanism.

Before an application can update a track on the source that has not yet been copied, the track is copied to the target volume. Reads that are subsequently directed to this track on the target volume are now satisfied from the target volume instead of the source volume. After some time, all tracks will have been copied to the target volume, and the FlashCopy relationship will end.

You cannot create a FlashCopy on one type of operating system and make it available to a different operating system. You can make the target available to another host running the same type of operating system.

Planning for FlashCopy on ESS

Because FlashCopy invariably will be used on production systems, you should carefully plan the setup of your environment and test it thoroughly. This is an important step to minimize the possibility of error and potential rework.

Hardware and software requirements

To use FlashCopy, you need to comply with the following prerequisites:

Have a FlashCopy feature purchased and enabled on your IBM TotalStorage Enterprise Storage Server (ESS) by the Customer Engineer (CE). The feature code is dependent on the total disk capacity of your ESS, rather than on the capacity of the volumes that will use FlashCopy.

On the server that will have the FlashCopy target volumes attached, you need to have enough SCSI target IDs and/or SCSI or Fibre Channel LUNs available (not occupied by volumes). The ESS can have up to 15 SCSI target IDs each with up to 64 LUNs on one SCSI channel and up to 4095 LUNs on a Fibre Channel port.

You need TCP/IP connectivity between the ESS and the host system that will initiate FlashCopy (usually this is the system that will access the FlashCopy target) in order to use Copy Services Command Line Interface (CLI). You can achieve that by connecting the ESS to the company intranet. You have to install the CLI on the host that will be using it.

- If you have Independent Software Vendor (ISV) software installed that writes directly to disk, you need to contact the ISV regarding their support for ESS Copy Services.
- Review your volume manager software considerations for FlashCopy:
  - AIX LVM
  - Veritas VxVM
  - HP SAM
  - Sun Solaris DiskSuite
- The IBM Subsystem Device Driver fully supports HACMP clusters in both concurrent and non-concurrent access modes. Subsystem Device Driver (SDD) works with FlashCopy volumes.

Configuration planning

The most important consideration is to have an available volume (LUN) in the logical subsystem (LSS) where the source volume resides. The target LUN has to be of the same size as the source or bigger. The space for target data has to be available even if only the *Do not perform background copy* option will be used.

Resource planning

When planning your ESS volume layout, it is important to consider the capacity you may need for FlashCopy targets. Bear in mind that the disk space you need is real disk space. You must also consider that a FlashCopy target is restricted to the same LSS as its source volume. So, when you allocate additional storage for FlashCopy targets, consider how much space in each LSS you need to leave unallocated for them.

You cannot initiate a FlashCopy session on a source and target that are already in a FlashCopy session. You need to wait for the FlashCopy task to complete, or until you can withdraw the pair manually. If you have used the *Do not perform background copy* option, you always need to withdraw the pair.

Data consistency considerations

It is very important to verify that the copy of the data you will be using is fully consistent by using a proper file system check procedure, as provided by your operating system. If you are going to automate your FlashCopy procedures, consider including this check each time when you make the FlashCopy target available to the host. In all cases, before starting the FlashCopy procedure, the target volume must be unmounted; this ensures that there is no data in any system buffers that could be flushed to the target and potentially could corrupt it.

Test plan and disaster recovery plan

If you plan to use FlashCopy, you need to test your setup. Do not forget that you are dealing with a binary copy of the data which was done out of control of your operating system. Prepare a test plan and, if you are using FlashCopy for backup/restore, a recovery plan also.

Monitoring and managing FlashCopy pairs and volumes

FlashCopy pairs and tasks can be managed by both the ESS Specialist Copy Services Specialist Web panel and the Command Line Interface (CLI) on the host.

The ESS Specialist Web Interface allows you to manage FlashCopy volumes and tasks. You can establish and withdraw a FlashCopy by clicking on the graphical representations of the volumes in the Copy Services Specialist. If you wish to perform a FlashCopy from the CLI, you must create a FlashCopy task within the Specialist and save it. You can either execute your tasks from the Specialist, or you call them with the `rsExecuteTask` command in the CLI.

Using the CLI with predefined tasks enables automation, and it minimizes the danger of a human error when handling physical volumes by their volume numbers or names from the ESS Specialist Copy Services Specialist Web panel.

Using a FlashCopy target volume

Remember that if you have established a FlashCopy with the *Do not perform background copy* option, you need to withdraw the FlashCopy pair after you have finished using the FlashCopy target volume. If you choose to perform a full copy, the relationship will be withdrawn automatically when the background copy task ends.

If you will be using FlashCopy for data backup purposes, change your recovery procedure so that you will be able to recover even when the data has been backed up by a different backup client than the original owner of the LUN, or it has been backed up from a different location in the file system (the target mount point).

Of course, you can perform the FlashCopy from the restored volume to the original LUN using the full copy option that will perform the actual data copy to the target volume.

Automation

Different operating systems allow different levels of automation. The automation can be done using batch or script files executed before and after the application that uses the FlashCopy target.

3.5.3 Summary

Split mirroring and FlashCopy is one of solutions that we recommend for maximizing IBM remote copy solutions using IBM TotalStorage Enterprise Storage Server. The benefit of this is you can reduce back-up and batch window time, and recovery time without stopping your system while your system is running.

4

Tape solutions at a distance

Typically most companies will take a copy of their data on tape and put it in an off-site location for disaster recovery purposes in case of failure of the local site.

Today's 24x7 forever kind of application environments have also meant that these companies have deployed SAN's at their local sites for tape backup and archiving using SAN applications like LAN-free backup, and server-less backup.

These near-zero backup window environments have further built the need for having the backup and archiving happen at remote and/or off-site locations using the extended distance SAN solutions. Companies are also looking for ways to consolidate the daily backup and archiving activities that happen at multiple sites, into a centralized off-site location, therefore also allowing for disaster recovery incase of failure of any one of the primary sites.

This is also called Remote Tape Vaulting. An extension of Remote Tape Vaulting is Remote Tape Disaster Tolerance where there is a tape library also at each of the sites.

IBM offers a comprehensive range of tape storage solutions and Figure 4-1, shows some of the solutions we discuss in this chapter.

Figure 4-1 IBM TotalStorage tape solutions

In this chapter we explore the distance SAN tape concepts and solutions and the IBM tape solutions and their application in SAN distance solutions.

We begin with terminology and then describe the extended SAN tape solution concepts. After this, we then detail IBM tape solutions, such as LTO, Magstar MP, Magstar 3590, Magstar 3494 and IBM VTS B10 and B20, Peer to Peer VTS and extended SAN tape solutions examples.

4.1 Terminology

There is no industry standard definition for terms such as tape library sharing and tape drive pooling, so we will define what we mean by them in this chapter.

4.1.1 Tape library

A tape library consists of the physical robotics that move cartridges, one or more tape drives, and slots for tape storage. It must also have a mechanism for controlling the robotics (a library controller), and may also have a library manager, which maintains inventory and mediates sharing. In most cases, a library does not have a built-in library manager, so server-based software has to provide the library management function. As examples, the IBM 3584 and IBM 3494 have built-in library managers, whereas the IBM 3583 does not.

4.1.2 Tape library sharing

This is multiple servers attached to a tape library sharing the robotics. The tape drives and slots within the library may or may not be shared among the attached servers.

Tape library sharing has been practiced for some time by partitioning a physical library into multiple logical libraries. Alternatively, the library can appear to be shared by multiple hosts when, in reality, one of the hosts (the library manager) is issuing all the library commands both for itself and for the other hosts (clients), but all of them have direct access to the tape drives (tape pooling).

Tape drive pooling or tape drive sharing

Providing tape drives to each server is costly, and also involves the added administrative overhead of scheduling the tasks and managing the tape media. Tape pooling is the ability to allow two or more servers to logically share tape drives within a tape library. In this case, the servers need to be attached to the same SAN as the tape drives, and there needs to be some software management to control who owns the drives and tape cartridges at any one time.

Tape pooling across a SAN is supported by Tivoli Storage Manager, starting with Version 3.7. The nature of the SAN is not critical to this software solution, because the control of the actual access is done through the Tivoli Storage Manager management software. Physical access to each tape drive is required which the SAN topology allows and the dynamic sharing of one or more tape drives between multiple servers.

When using a tape library with an integrated library manager, like the IBM 3494, the tape drives can be shared without one of the Tivoli Storage Manager servers taking the role of the arbiter. The tape drives can be shared using SCSI Reserve/Release functions (dubbed auto-share, a Magstar 3590 feature) and the tape media is controlled by the library and assigned to a specific host, therefore eliminating the need to manage the tape inventory using application software.

Tape library partitioning

This is the ability to partition tape drives and slots to create logical libraries within the same physical library. The server attached to each logical library has no knowledge of any drives or slots outside the partition, and the partitions are fixed.

Multipath tape library

This is an architecture of the tape library whereby the tape library has multiple paths to the robotics controller or SCSI Medium Changer (SMC), as shown in Figure 4-2.

The IBM 3584 and IBM 3575 are multi-path libraries. A multipath architecture is a prerequisite for partitioning. The IBM 3583 is a single path architecture.

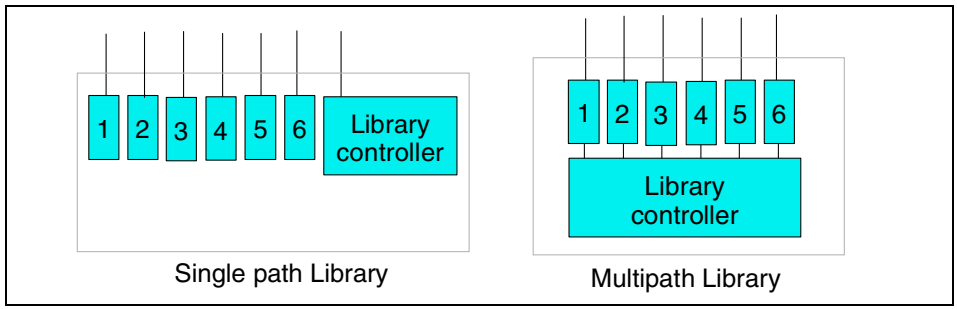


Figure 4-2 Single path and multiple-path tape library

In a SAN environment, for example, shown in Figure 4-3, the three hosts see their respective logical tape library — host 1 sees logical library 1, host 2 sees logical library 2 and so on.

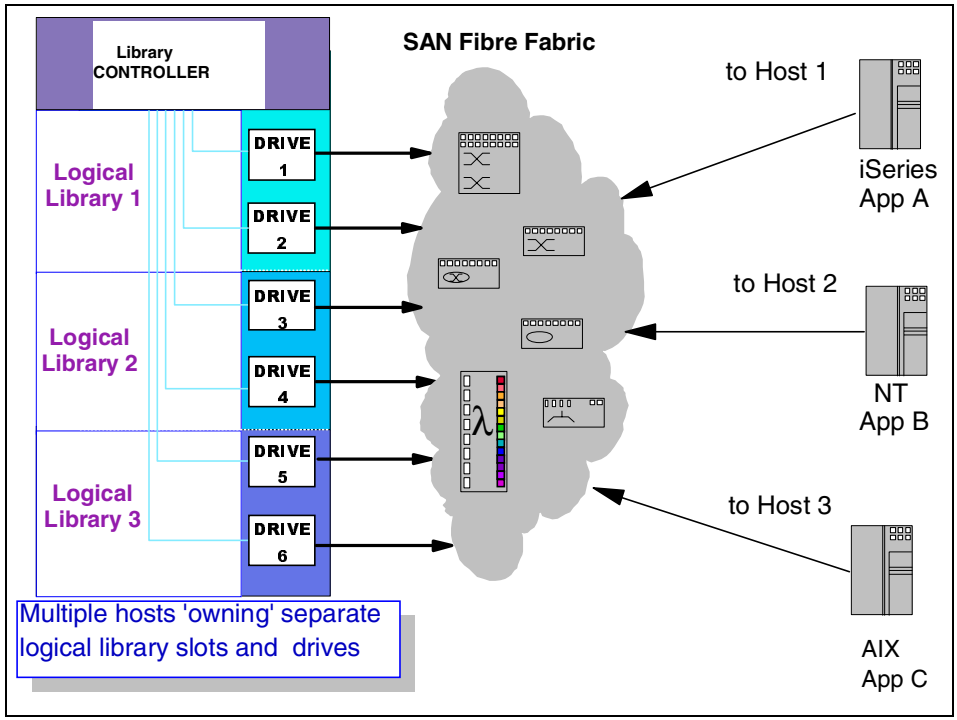


Figure 4-3 Example of multipath library in SAN environment

4.1.3 Remote tape vaulting

Remote Tape vaulting consists of electronically transmitting and creating backup tapes at a secure off-site facility, moving mission-critical data off-site faster and with greater frequency than traditional data backup processes. The traditional backup process is where the tape is actually created in a locally attached tape library, then ejected from the library and finally removed to an off-site location, as shown in Figure 4-4.

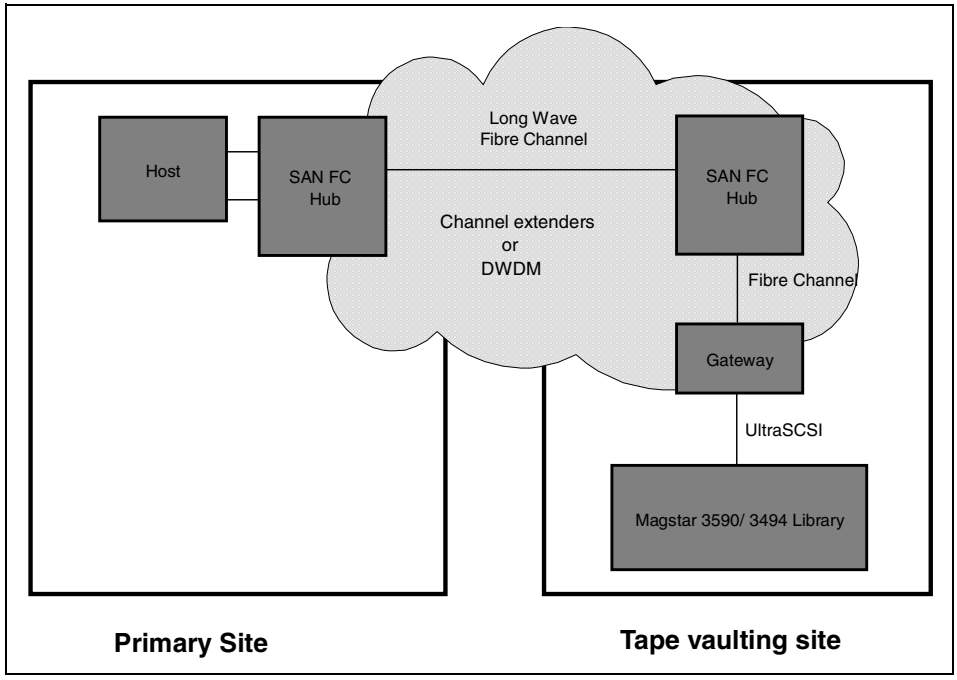


Figure 4-4 Remote tape vaulting

4.1.4 Remote tape disaster tolerance

This is basically an extension of remote tape vaulting where there is a tape library at each of the sites, both local and off-site. This is shown in Figure 4-5.

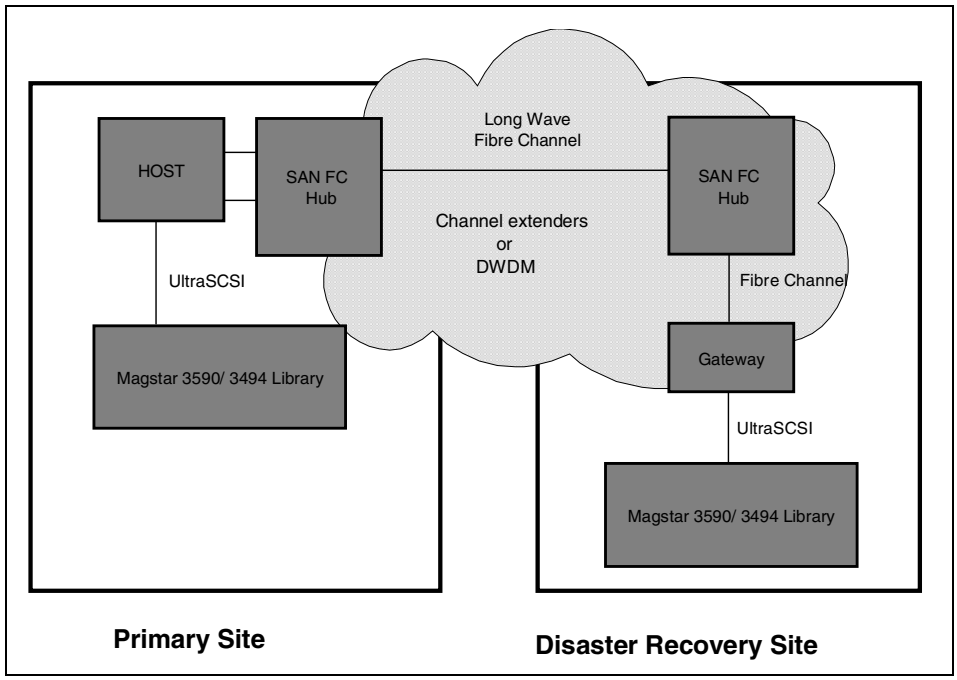


Figure 4-5 Remote tape disaster tolerance

4.1.5 Virtual tape library

In a virtual tape library, such as the IBM 3494-B10 Virtual Tape Server (VTS), data is stored at high speed onto disk and is then intelligently staged to a small number of physical tape drives. The performance and intelligent staging allows the VTS to appear as a large number of virtual tape drives.

This is particularly useful for OS/390 environments where tape utilization can be poor. In the open server environment, this is not a problem; however if you already have a VTS installed, it can be partitioned and connected to an open system SAN.

4.1.6 Backup server and backup client

A backup server is a server running the master or control version of the backup software. The entire server may be dedicated to the backup task or it may be running other applications as well.

A backup client is a server running productive applications with the backup client software installed.

4.1.7 LAN-free backup

Backup is traditionally done by transferring the data to be backed up over the LAN. The advent of SANs enables data to be transferred directly over the SAN from disk storage to the backup server, and then directly over the SAN to tape.

Tivoli Storage Manager version 3.7 or higher supports a LAN-free backup mode of operation. A LAN-free backup is done by a backup server using the SAN topology and functions of FC to move the backup data over the SAN, therefore eliminating the LAN from the data flow.

This does two things: first, the LAN traffic is reduced, and secondly (and most importantly), the traffic through the backup server is reduced. This traffic generally is processor-intensive because of TCP/IP translations. With LAN-free backup, the backup server orchestrates the data movement, manages the tape library and drives, and tells the clients what data to move. The client is connected to the SAN; its data can be on the SAN or the data can be on storage directly attached to the server as shown in Figure 4-6.

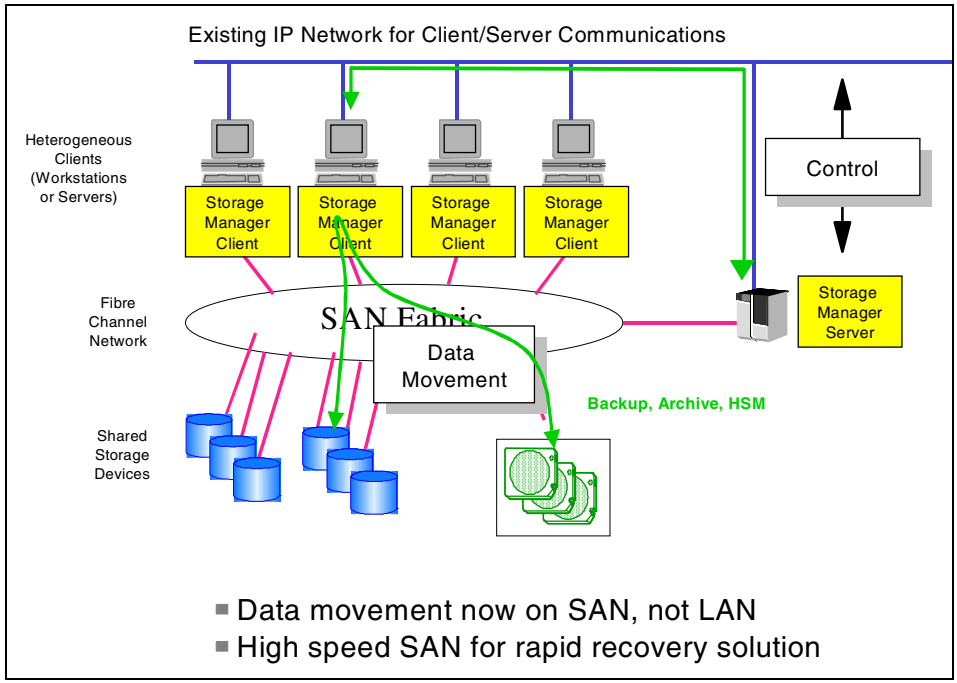


Figure 4-6 Lan-free backup

The LAN is still used to pass metadata (data about the data itself) back and forth between the backup server and the client. However, the actual backup data is passed over the SAN. The metadata is the data needed by the backup server to manage the entire backup process, and includes things like the file name, the file location, the date and time of the data movement, and where the new copy resides. The metadata is small compared to the actual client data being moved.

4.1.8 Server-less backup

Server-less backup refers to the ability to take a snapshot of the data to be backed up with minimal or no disruption to productive work, then move it intelligently between tape and disk without the data going through a server. This is shown Figure 4-7.

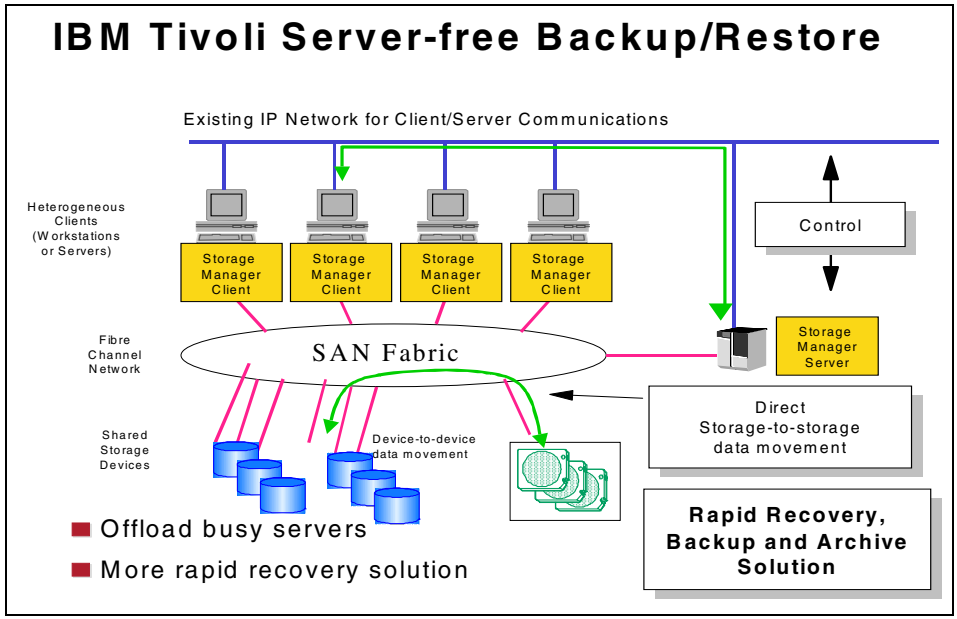


Figure 4-7 Server-less backup

All the elements to achieve this are available and there are currently several software solutions available. The key elements are:

- **Creating an instant copy**. This can be done either by host-based software or by copy functions in the disk storage server itself, such as FlashCopy in the Enterprise Storage Server (ESS) or the Modular Storage Server (MSS).
- **Mapping physical blocks to logical files.** The physical data blocks must be mapped to the file system or database and the mapping process managed.

- **Moving the data directly between the disk and tape.** This is done using the SCSI extended copy command. The component that performs the SCSI commands can theoretically be in one of the SAN fabric components, or it can be software that runs on a server.

4.2 IBM tape solutions

IBM offers the most comprehensive range of tape products using various tape technologies features, functionality and performance characteristics. For distance SAN attached tape solution examples we will consider only the midrange and the enterprise tape libraries, however the standalone tape drives and autoloaders can also be used.

Currently most midrange and enterprise class tape subsystems offer native or direct Fiber attachment, usually using a FC-AL arbitrated loop connection. However, there are many drives and libraries that might otherwise be suitable for use initially in a SAN which have SCSI interfaces and, for these, bridges or gateways can be used to convert SCSI to Fibre Channel.

A SAN can provide extended distance for tape applications such as remote vaulting; however, it also provides much greater flexibility for sharing. This has two implications: you need to consider the sharing capability of a library when you are deciding what type of library to use; and, high levels of sharing can make it highly attractive to consolidate tape work from many small, unshared libraries to a larger, more capable library.

4.2.1 IBM LTO tape solutions

Founded in 1997, The Linear Tape Open (LTO) program is a joint IBM, HP, and Seagate initiative to establish a new, open format for high capacity, high performance tape storage products. LTO technology has developed into two open tape format specifications, Accelis and Ultrium.

IBM currently offers several Ultrium format tape products designed to meet different levels of automation with SCSI or native fiber interfaces, depending on model. The Ultrium tape format is the LTO technology implementation optimized for high capacity and performance with outstanding reliability, in either a standalone or an automated environment. The Ultrium tape format uses a single reel cartridge to maximize capacity. It is ideally suited for backup, restore, and archival applications.

IBM 3583 Ultrium Scalable tape library

A mid-range tape library of the IBM Ultrium family is the IBM 3583 Ultrium Scalable tape library. The tape library automates the retrieval, storage, and control of LTO Ultrium cartridge tapes. Cartridges can be mounted and dismounted on tape drives by, using supporting software from the host without operator intervention.

Figure 4-8 IBM LTO 3583 tape library

The tape library is designed for easy expansion. It can accommodate from one to six tape drives and has cartridge storage configurations that hold 18, 36, or 72 cartridges. Any starting configuration can be upgraded to the maximum configuration of six drives and 72 cartridges.

The tape library input/output (I/O) station enables cartridges to be inserted and removed without disrupting library operation. There are two I/O station options, a single slot option, and a 12 slot option. The 12 slot I/O station is required to achieve the 72 cartridge maximum library configuration. With the 12 slot I/O station feature the library can be configured as 72 storage slots, or 60 storage slots and 12 I/O slots.

The tape library is a stand-alone unit with an optional rack mount kit feature. The library can be installed into an Electronic Industries Association (EIA) standard EIA-310-D 19-inch rack.

The Remote Management Unit (RMU) comes standard in every library. Library status can be sent to the network as Simple Network Management Protocol (SNMP) traps. The StorWatch Specialist enables network access to the library for more detailed status and control. The StorWatch Specialist provides network access to all library operator panel functions.

The SAN Data Gateway Module is another available library feature. The gateway provides the SCSI library with an avenue into a SAN infrastructure. With 2 Gigabit (Gbit) port speeds this module is compatible with 2 Gbit Fibre Channel components. It is also backward compatible with 1 Gbit Fibre Channel hardware. Two Fibre Channel ports make multiple attachments easy and support fail-over redundancy. The gateway also has four Ultra2 LVD SCSI ports for attachment of tape drives and medium changer.

SAN data gateway module with StorWatch Specialist

The gateway enables the library to connect to a SAN with simple Fibre Channel connectivity and enhances the following features:

- Assorted device and interface connectivity
- Storage network management
- Access control
- SCSI extended copy

The gateway is an essential component of the SAN infrastructure. The gateway integrates a number of different technologies such as Fibre Channel point-to-point, Fibre Channel arbitrated loop (FC-AL), and LVD SCSI.

The gateway maps addresses across and between different interfaces. The gateway maintains the persistency of the address maps across power ups of attached systems, devices, and the gateway. The gateway supports up to 255 unique device addresses across multiple interfaces.

The gateway has full knowledge of the SCSI-3 and SCSI-2 protocols for disk, tape, and medium changer devices.

The gateway includes support for remote management and event notification over Ethernet using the IBM StorWatch SAN Data Gateway Specialist. The gateway has internal event logging and analysis and runs periodic health checks for predictive failure analysis. All of these management, configuration, and remote notification capabilities are accessible using the industry-standard SNMP protocol. The StorWatch Specialist is a software package that provides remote management and configuration capabilities. The gateway provides access control capabilities. The gateway monitors hosts and devices attached to its

interfaces and controls access between ports. Access control applied across different interfaces is a requirement for multi-initiator SAN solutions. Using the StorWatch Specialist you can partition the SAN for different levels of access and performance.

The gateway also includes a data mover option that enables server-free backup. You can move data from disk to tape (backup) or from tape to disk (restore) without tying up valuable system resources.

The IBM 3583 tape library can be attached to:

- AS/400 and iSeries with OS/400 4.4 or later
- RS/6000, RS/6000 SP and pSeries with AIX 4.3.2
- Sun servers with Solaris 2.6, Solaris 7 or 8
- HP A, L, N and V-class servers with HP-UX 11.0
- Netfinity, xSeries and other PC servers supporting Microsoft
- Windows NT 4, Windows 2000 or Red Hat Linux 7.0 (either 2.2.16 or 2.4.2 kernel)

IBM 3584 Ultrium Ultra scalable tape library

An enterprise modular tape library of the IBM Ultrium family is the IBM 3584 Ultrium ultra scalable tape library. The tape library automates the retrieval, storage, and control of LTO Ultrium cartridge tapes. Cartridges can be mounted and dismounted on tape drives by using supporting software from the host without operator intervention. It is the largest member of the IBM Ultrium family of tape library storage solutions. It is shown in Figure 4-9.

Figure 4-9 IBM 3584 LTO tape library

The IBM 3584 Ultra Scalable Tape Library is a stand-alone device that provides reliable, automated tape handling and storage for unattended mid-range systems and network servers. The basic library is a single storage unit known as the base frame. The library's scalability allows you to increase capacity by adding up to five additional storage units, called expansion frames. Each frame in the library may contain up to 12 Ultrium Tape Drives or DLT 8000 Tape Systems, but may not contain a mix of both.

To match your system capacity and performance needs, you can tailor the 3584 Ultra Scalable Tape Library to take advantage of the following features:

- Use of up to 72 Ultrium Tape Drives or 60 DLT 8000 Tape Systems
- Aggregate sustained data transfer rate from 108 GB to 7.8 TB per hour for Ultrium Tape Drives (at 2:1 compression)
- Compressed data capacity of 496.2 TB for Ultrium Tape Drives (at 2:1 compression)
- For the IBM Ultrium Tape Drive, support of any combination of interfaces, including Fibre Channel, Low Voltage Differential (LVD) Ultra2 SCSI, and High Voltage Differential (HVD) Ultra SCSI
- For the DLT Tape System, support of the Fast/Wide LVD and HVD SCSI interfaces
- Multi-Path Architecture that enables a single library to be shared by multiple homogeneous or heterogeneous applications
- Support of any appropriate combination of frames that use Digital Linear Tape (DLT) or Linear Tape-Open (LTO) Ultrium media

The 3584 Ultra Scalable Tape Library features the SAN-ready Multi-Path Architecture, which allows homogeneous or heterogeneous open systems applications to share the library's robotics without middleware or a dedicated server (host) acting as a library manager. The SAN-ready Multi-Path Architecture makes sharing possible by letting you partition the library's storage slots and tape drives into logical libraries. Servers can then run separate applications for each logical library. This partitioning capability extends the potential centralization of storage that the SAN enables. The Multi-Path Architecture is compliant with the following attachment interfaces:

- Small Computer Systems Interface (SCSI)
- Fiber Channel

Whether partitioned or not, the 3584 Tape Library is certified for SAN solutions (such as LAN-free backup).

The Multi-Path Architecture also lets you configure additional control paths for any one logical library. A control path is a logical path into the library through which a server sends standard SCSI Medium Changer commands to control the logical library. Additional control paths allow the cartridge inventory of the 3584 Ultra Scalable Tape Library to be shared by multiple IBM e-server iSeries and AS/400 servers, or by other open systems hosts that run the same applications. Additional control paths also reduce the possibility that failure in one control path will cause the entire library to be unavailable.

The IBM 3584 is supported in these environments:

- AS/400 and iSeries with OS/400 4.4 or later
- RS/6000, RS/6000 SP and pSeries with AIX 4.3.2
- Sun servers with Solaris 2.6, Solaris 7 or 8
- Hewlett-Packard A, L, N and V-class servers with HP-UX 11.0
- Netfinity, xSeries and other PC servers supporting Microsoft Windows NT 4 or Windows 2000

Summary of the IBM Ultrium family

The 3580, 3581, 3583 and 3584 are part of the same family, meaning that the drive, cartridge technology and formats are the same, so the cartridges are interchangeable between the libraries as dictated by the LTO format standards. However, the machines (for example, 3583 and 3584) are not upgradeable from one to another, nor can the Ultrium drives be exchanged between different libraries.

The IBM LTO family of products is sold directly through IBM and its business partners, and are sold to other manufacturers under the IBM StorageSmart Solutions name, for integration into their library solutions.

For the latest information about the IBM Ultrium family, refer to *The IBM LTO Ultrium Tape Libraries Guide*, SG24-5946, or to:

`http://www.ibm.com/storage/lto`

Attaching LTO systems to a SAN environment

All four LTO Ultrium tape systems are available with both SCSI LVD and HVD interfaces. The 3583 and 3584 LTO tape library is also available with a direct FC-AL attachment.

There are two possible ways of attaching an Ultrium tape library into a SAN environment.

1. Direct attachment using a native Fibre Channel interface is available on the 3584. It enables you to attach each single FC-AL LTO tape drive, via the FC patch panel inside the 3584 to:

   – IBM 2103 hub (for distance extension only)
   – IBM 3524 managed hub
   – IBM 2109 switch
   – IBM 2031-L00 McDATA ES-1000 loop switch
   – IBM 2032-001 and 2032-064 McDATA directors with 2031-L00 Loop switch
   – IBM 2042 INRANGE FC/9000 Fibre Channel director

   A native Fibre Channel connection has the advantage that you will have an increased peak data rate of up to 100 MB/second. You can use the tape units at greater distances without additional hardware and it will also allow you greater systems configuration flexibility.

2. For drives with LVD interfaces, attachment can be through the SAN Data Gateway Router, 2108-R03. For drives with HVD interfaces, attachment will be through either the SAN Data Gateway Router, 2108-R03, or the SAN Data Gateway, 2108-G07. In each case, the server will need an appropriate Fibre Channel adapter supported by these gateways.

   As the 3583 has an optional integral Fibre Channel connection, there are additional options, as shown in Table 4-1.

Table 4-1 Connection options

Model Number	**3583 Model L18, L36, L72 feature code 8003**	**3583 Model L18, L36, L72 feature code 8004**
Connection type	Ultra2/Wide SCSI LVD interface	Ultra/Wide SCSI HVD interface
SAN connectivity	1. IBM SAN Data Gateway Router, 2108-R03, with feature code 2840, part number 2108R3L 2. Internal SAN gateway	1. IBM SAN Data Gateway Router, 2108-R03, with feature code 2830, part number 2108R3D 2. IBM SAN Data Gateway 2108-G07 3. Internal SAN gateway

As the 3584 has native Fibre Channel drives, there are additional options, as shown in Figure 4-10.

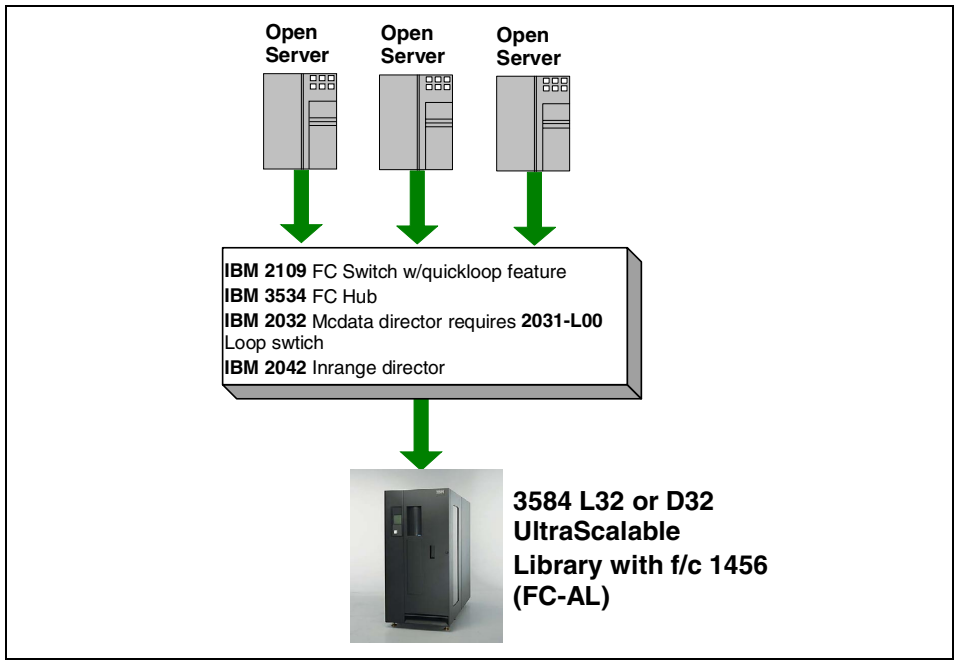


Figure 4-10 IBM 3584 tape library SAN connection

In Table 4-2 we show the 3584 LTO tape library SAN connections.

Table 4-2 IBM 3584 LTO tape library SAN connections

Feature code	1454	1455	1456
Connection type	Ultra2/Wide SCSI LVD interface	Ultra/Wide SCSI HVD interface	FC-AL interface (SW-FC)
SAN connectivity	IBM SAN Data Gateway Router 2108-R03 with feature code 2840, part number 2108R3L	1. IBM SAN Data Gateway Router 2108-R03 with feature code 2830, part number 2108R3D 2. IBM SAN Data Gateway 2108-G07	1. directly to a FC-AL adapter to most host system, (for example, see 3466) 2. IBM SAN Fibre Channel switches and hubs

For latest updates on connectivity and supported servers, refer to:

http://www.storage.ibm.com/hardsoft/tape/index.html

4.2.2 IBM TotalStorage Enterprise Tape System 3590 solutions

The IBM 3590 High Performance Tape Subsystem comes in different models and offers several attachment options to meet your needs. Each drive can have an automatic cartridge facility (ACF) with a 10-cartridge magazine. The drives have SCSI attachment or Fibre Channel attachment. Only the 3590 Model E is attachable to Fibre Channel. Each drive can connect to an IBM 3590 tape controller for Enterprise Systems CONnection (ESCON) or Fibre CONnections (FICON) attachment of a 3590. Large scale automation offerings, which include the IBM 3494 Tape Library and StorageTek Silo, support the 3590.

The 3590 Bxx tape drives read and write data on 128-track format on IBM 3590 High Performance Cartridge Tape. This read, and write function results in a 10GB uncompressed data tape capacity. Model Exx tape drives read and write data on the 256-track format on IBM High Performance Cartridge Tape. This read, and write function results in a 20 GB tape. The Extended High Performance Cartridge Tape increases the IBM Magstar 3590 E Model 256-track serpentine format capacity to 40GB. It also increases the IBM Magstar 3590 B Model 128-track serpentine format capacity to 20GB.

At 3 to 1 compression on the High Performance Cartridge Tape, the capacity increases to 60 GB on E models and 30 GB on B models. The Extended High Performance Cartridge Tape doubles the compressed capacities to 120 GB on E models and 60 GB on B models. E models have a 14 MB per second device data rate, and B models have a 9 MB per second transfer rate.

With data compression, the 3590 drives can more effectively utilize the full capability of the Fibre Channel data rate. Data compression also enhances the SCSI Ultra/wide data rate and the ESCON or FICON data rate. The Fibre Channel attachment data rate is an instantaneous 100MB per second. Also, the SCSI Ultra/wide instantaneous data rate is up to 40MB per second. For ESCON, the channel-instantaneous rate is 20MB per second, and for FICON it is 100MB per second.

3590 Tape Drive Models

1. Models B11 and E11 are drives with a 10-cartridge Automated Cartridge Facility (ACF) and five modes of operation. Model B11 has two SCSI-2 differential 2 SCSI Ultra/wide ports. Model E11 has two SCSI-3 differential 2-byte SCSI Ultra/wide SCSI ports or two Fibre Channel ports. In an ESCON-or FICON-attached environment, Models E11 and B11 support the ACF in all modes except random. Models E11 and B11 are supported by ESCON with the A00, A50, and A60 and supported by FICON with the A60 when attached to any of the following:

– **For Model E11:**

  • Model A50 or A60 controller in an IBM 3590 A14 frame
  • Model A50 or A60 controller in a rack

– **For Model B11:**

  • Model A00, A50, or A60 controller in an IBM 3590 A14 frame
  • Model A00, A50, or A60, and 3591 Model A01 controller in a rack

2. Model B1A is a single-cartridge tape drive with two SCSI-2 differential SCSI Ultra/wide ports. Model E1A is a single-cartridge tape drive with two SCSI-3 differential 2-byte SCSI Ultra/wide ports or two Fibre Channel ports. Models E1A and B1A are supported by ESCON with the A00, A50, and A60 and supported by FICON with the A60 when attached to any of the following:

   – **For Model E1A:**

     • Model A50 control unit in an IBM 3590 C14 frame, or an A50 control unit external to a C12
     • Model A60 control unit in an IBM 3590 C10 frame or in a 3494 D14 frame
     • Model A50 controller in an IBM 3494 D14 or L14 frame

   – **For Model B1A**:

     • Model A00 or A50 controller in an IBM 3590 A14 or C14 frame, or an A50 control unit external to a C12
     • 3591 Model A01 external to a C12 frame
     • Model A00 or A50 controller in an IBM 3494 D14 or L14 frame
     • Model A60 control unit in an IBM 3590 C10 frame or in a 3494 D14 frame

3590 Control Units

In the sections that follow we describe the features of the models.

Model A00

Model A00 is a tape control unit that provides ESCON attachment for Models B1A and B11. It is installable in a 3590 Model A14, C14 frame, or 3494 Model L14 or D14 tape library, or external to a 3590 Model C12 or rack. With the selection of IBM Feature Codes, Model A00 provides a single data transfer path with one (FC3311) or two (FC3311 and FC3312) ESA/390 ESCON channel attachment adapters.

It supports up to four 3590 Model B11 or B1A drives. The controller can be at a maximum channel distance of 43 kilometers (27 miles) from the host when using fiber-optic cable between ESCON directors. Model A00 is usable in an automated environment with B1A drives in IBM 3494 or 3495 tape libraries. It is also usable in a stand-alone environment with 3590 Model B11 drives in an A14 frame. Model A00 are installable in an STK Silo with C12 or C14 frames in supported racks. A mixture of A00 and A50 can be in an A14 to provide two controllers in a single enclosure.

Note: A00 does not support Exx Models.

Model A50

Model A50 provides ESCON attachment for Models E11, E1A, B1A, and B11. Model A50 is installable in a 3590 Model A14, or C14 frame, 3494 Model L14, or D14 tape library or rack. It is externally attachable to a 3590 C12. Model A50 provides a single data transfer path with one (FC3311) or two (FC3311 and FC3312) ESA/390 ESCON channel attachment adapters. It supports either up to four Model E1A and E11 drives, or up to four Model B1A and B11 drives. The controller can be at a maximum channel distance of 43 kilometers (27 miles) from the host when using fiber-optic cable between ESCON directors. Model A50 is usable in a stand-alone environment with B11 drives in an A14 frame and supported racks. Model A50 is usable with B1A in an STK Silo with C12 or C14 frames. In a C12, the control unit must be external to the unit. Up to two A50s, or an A50 in combination with an A00, are installable in an A14. The A50 controller is configured to operate in either 3590 native mode or in 3490E emulation mode.

Model A60

Model A60 is installable in a 3494 Model D14 tape library. It is also installable in a 3590 Model A14 Frame, a 3590 Model C10 Silo-Compatible Frame, or a standard 19 inch rack. The controller can be at a maximum channel distance of 43 kilometers (27 miles) from the host when using fiber-optic cable between ESCON directors. FICON attachment is available via either shortwave or longwave. The A60 is directly attachable via a FICON long wavelength attachment to host systems up to a 10km distance or up to 20km with RPQ8P1984. The A60 is also directly attachable up to 100km away with a FICON/Fibre Channel Switch with appropriate repeaters. With FICON short wavelength attachment, the A60 is directly attachable to a host system or FICON/Fibre Channel switch at a distance up to 500m.

In Figure 4-11 we show a connection example.

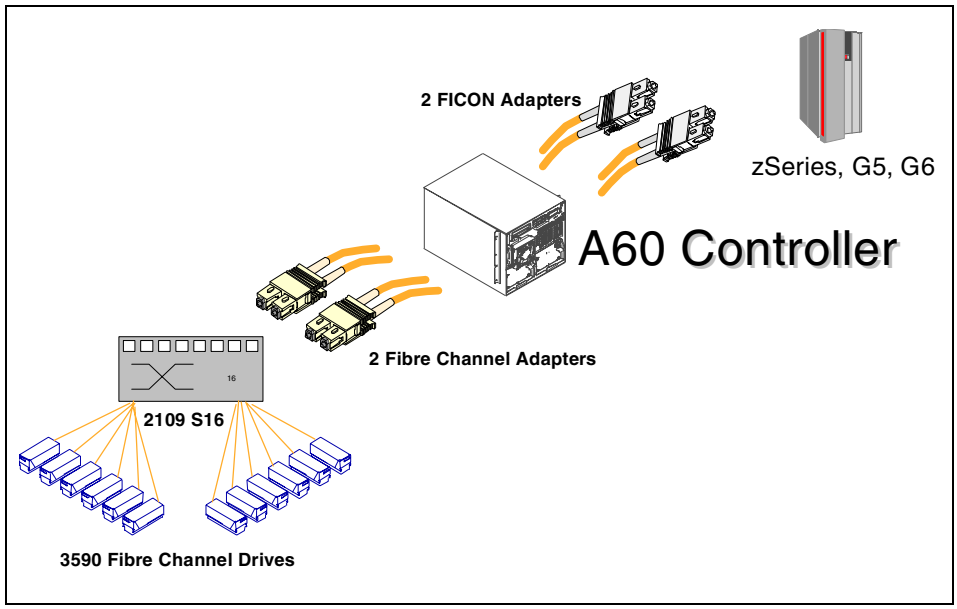


Figure 4-11 Magstar A60 controller FICON and FC connections

Model A60 may contain any of the following combination features:

- One to four dual-ported ESCON attachments
- One to two FICON attachments
- Up to three dual-ported ESCON and one FICON attachments
- Up to two dual-ported ESCON and two FICON attachments

Therefore, an A60 can support either of the following:

- Up to eight ESCON channels
- Up to two FICON channels
- Up to six ESCON channels and one FICON channel
- Up to four ESCON channels and two FICON channels

The A60 has the following attachment capability. More than four drives are attachable when using an adjacent-frame or multi-frame support feature:

- Up to 10 model B1A or E1A drives in a 3494 configuration
- Up to eight SCSI-attached model B1A or E1A drives in an STK Silo solution
- Up to 12 fiber-attached model E1A's in an STK Silo solution
- Up to eight model B11 or E11 SCSI attached drives in an A14 frame or rack solution

- Up to 12 fiber-attached E11 drives in an A14 frame or in either a 1.8m or 2.0m rack solution
- Up to 10 fiber-attached E11 drives in a 1.6m rack solution

Host system attachment

The following sections list the attachments to hosts that are supported.

1. SCSI Attach

   The subsystem attaches to the following host systems:

   - AS/400
   - RS/6000 SP
   - RS/6000
   - HP
   - iSeries
   - pSeries
   - Sun
   - Microsoft Windows NT
   - Microsoft Windows 2000
   - xSeries

2. ESCON Attach

   The following host systems through ESCON channels:

   - ES/3090-J, ES/3090-9000T
   - ES/9000 ™
   - S/390 ™
   - zSeries

3. FICON Attach

   The following host systems through FICON channels:

   - 9672 Enterprise G5 or G6 Servers
   - zSeries

4. Fibre Channel Attach

   The subsystem attaches to the following host systems:

   - RS/6000 SP
   - RS/6000
   - Sun
   - Windows NT
   - Windows 2000

Attaching Magstar 3590 to a SAN environment

There are two ways of attaching 3590s to a SAN, depending on whether they have UltraSCSI interfaces or Fibre Channel interfaces.

UltraSCSI drives can be attached by using the 2108-G07 SAN Data Gateway or the 2108-R03 SAN Data Gateway Router.

Fibre Channel drives can be attached directly to these fabric components:

- IBM 2103 hub (for distance extension only)
- IBM 3524 managed hub
- IBM 2109 switch
- IBM 2031-L00 McDATA ES-1000 loop switch
- IBM 2032-001 and 2032-064 McDATA directors with 2031-L00 Loop switch
- IBM 2042 INRANGE FC/9000 Fibre Channel director

In Figure 4-12 we show the typical SAN distances that are achieved.

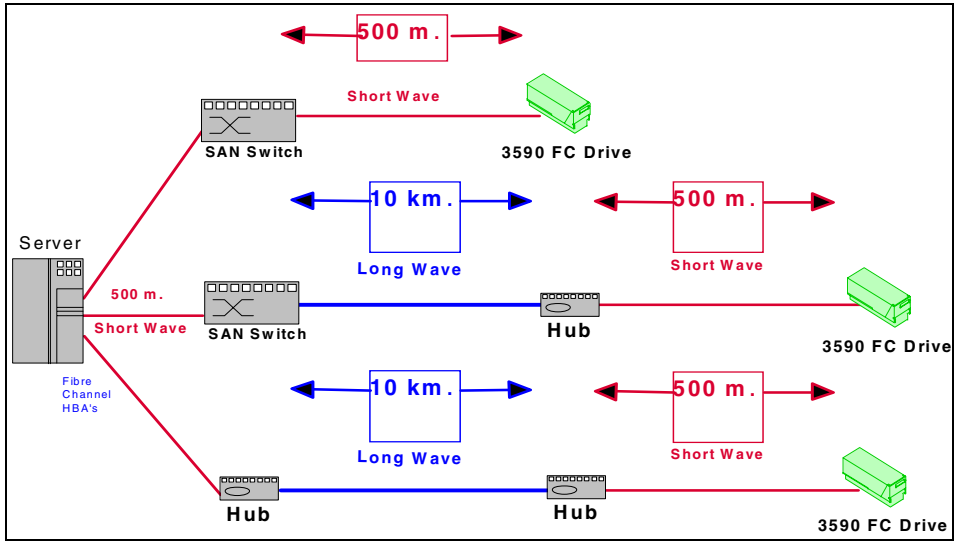


Figure 4-12 SAN distances with Magstar 3590

Usually, tape is installed in SANs using a single path from the tape drive to the switch or host server. The failure of a drive data path, switch, or a host bus adapter causes the immediate tape drive operation to fail. What is needed in the event of a path or component failure is the ability, like with disk drives, to dynamically retry the tape drive command through an alternate set of paths or components.

For example, the 3590 Fibre Channel device driver for AIX provides this fail-over function. During installation, it allows you to configure IBM Magstar 3590 Fibre Channel drives in a SAN environment with redundant paths that can be dynamically repathed, even during a job, in the event of a path or component failure. This repathing is done in a way that is completely transparent to the underlying application, host, or switch. This is possible because of an exclusive fail-over mechanism in the AIX tape device driver that allows the operator to allocate multiple paths to a 3590 drive. In the event of a path or component failure, the fail-over mechanism automatically retries the current tape job using an alternate, pre-configured path. This is accomplished without operator intervention, and in most cases, without aborting the current job in progress.

In Figure 4-13 we show the dual path connectivity of the 3590 in an AIX environment.

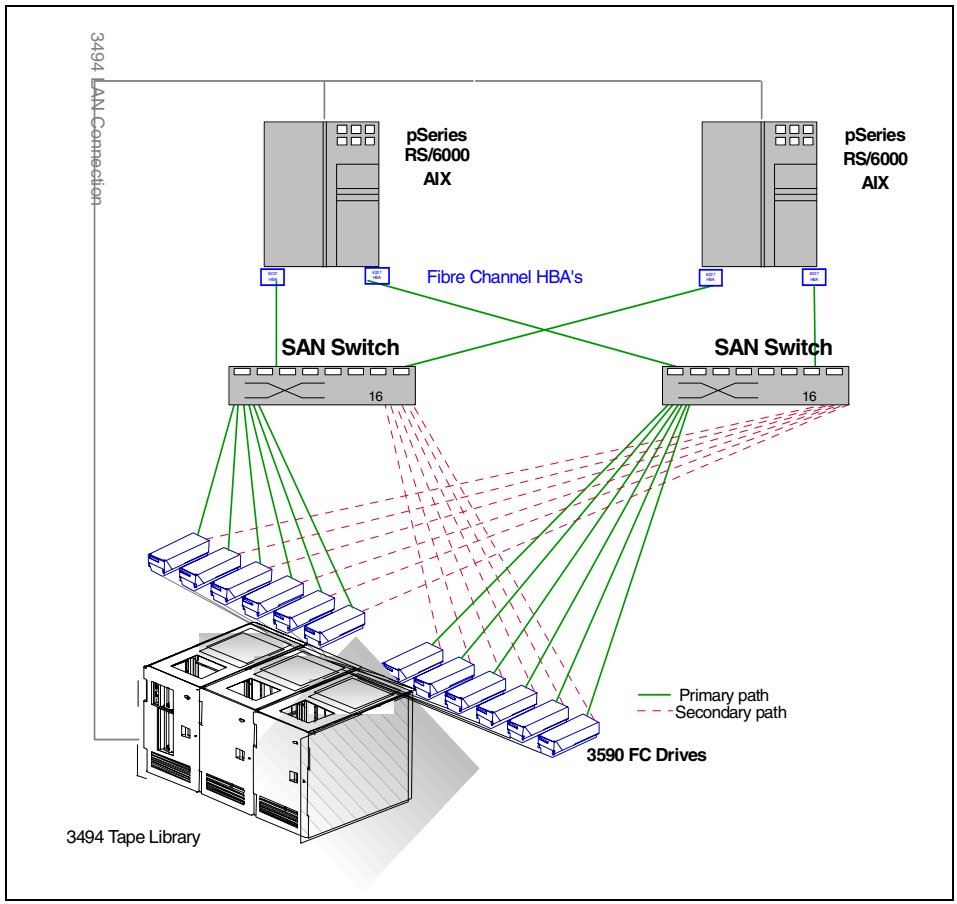


Figure 4-13 Magstar dual path failover configuration for AIX

When properly configured, no single point of failure in the network can cause a permanent failure. Moreover, since failover is managed at the device driver level, any application using the RS/6000 gets this high-availability function.

For latest updates on connectivity and supported servers, refer to:

http://www.storage.ibm.com/hardsoft/tape/3590/3590opn.html

4.2.3 IBM Magstar 3494 tape solutions

The IBM TotalStorage Enterprise Automated Tape Library (3494), meets the storage capacity requirements of 160 to 6240 cartridges. The 3494 supports an intermix of tape drive and 12.7-MM. (0.5-in.) cartridge technologies. It can be configured for a specific number of tape drives and control units.

In Figure 4-14 we show a Magstar 3494 tape library with the L, D and S frames.

Figure 4-14 IBM Magstar 3494

The 3494, with the models currently available, offers the following functions and features:

- Read capability for 3490E tape technology
- Read and write capability for 3490E and 3590 tape technology
- Support for one to 78 tape transports
- L1x Frame attachment to D1x, S10, and HA1 Frames, and the B16 VTS
- Support for 3490E Model C1A, C2A, and F1A tape drives and 3590 Model B1A and E1A tape drives

- Cartridge storage capacity of 160 to 6240 tape cartridges
- Data storage capacity up to 124.8 TB of uncompressed data and up to 561.6 TB of data with 3:1 compression
- Data paths using SCSI-2, Ultra SCSI, Fibre Channel, ESCON ™, FICON ™, and parallel (3490E only)
- Library Manager command paths using RS-232, LAN, ESCON, FICON, and parallel
- Automatic utilization of the full capacity of tape cartridges by stacking logical volumes end-to-end on the cartridge with the Virtual Taper Server (VTS) models
- Hot standby dual Library Manager control units, two service bays, and a second accessor
- Dual active accessors
- L1x upgrades from:
  - L10 to L12 or L14
  - L12 to L14
  - L14 to L12
- D1xTape Drive Expansion Frame model upgrades from:
  - D10 to D12 or D14
  - D12 to D14
- D1x feature upgrades from:
  - FC 5300 to FC 5302
  - FC 5300 to FC 5304
- VTS model upgrades from:
  - B16 to B18
  - B16 to D12 Frame
  - B18 to B20
  - B10 to B20

In Table 4-3 we show the model numbers and their description.

Table 4-3 Model number and description

Models	Description
L10, L12, L14	IBM TotalStorage Enterprise Tape Library Base Frame (L10, L12, L14 Frame) or (Model L10, L12, L14)
D10, D12, D14	IBM TotalStorage Enterprise Tape Drive Expansion Frame (D10, D12, D14 Frame) or (Model D10, D12, D14)
AX0	IBM TotalStorage Virtual Tape Controller (AX0) or (Model AX0)
B16, B18, B10, B20	IBM TotalStorage Virtual Tape Server (B16, B18, B10, B20 VTS), or (Model B16, B18, B10, B20 VTS)
CX0	IBM TotalStorage Virtual Tape Frame (CX0) or (Model CX0)
B16, B18, B10, B20	IBM TotalStorage Virtual Tape Server (B16, B18, B10, B20 VTS), or (Model B16, B18, B10, B20 VTS)
S10	IBM TotalStorage Enterprise Tape Storage Frame (S10 Frame) or (Model S10)
HA1	IBM TotalStorage Enterprise High Availability Tape Frames (HA1 Frames)
Note: Additional storage can be added to a 3494 with a D10, D12, or D14 Frame, with or without tape drives.	

The 3494 includes the following options:

- Convenience I/O station features
- RS-232 or LAN host attachment capability
- A high-capacity I/O facility
- A Dual Gripper feature
- A Remote Library Manager Console feature
- A second disk drive for the Library Manager
- A wide range of attachment capabilities

IBM TotalStorage Virtual Tape Server (VTS)

The VTS models deliver an increased level of storage capability to the traditional storage product hierarchy. To host software, a VTS looks like a 3490E Enhanced Capability Tape Subsystem with associated Cartridge System Tape or Enhanced Capacity Cartridge System Tape. This virtualization of both the tape drive and the storage media to the host allows transparent utilization of the 3590 tape technology capabilities. The IBM TotalStorage Peer-to-Peer Virtual Tape Server (PtP VTS) interconnects models B18, B10, or B20 VTSs for enhancement of data backup and recovery capabilities by providing dual-volume copy, remote functionality, and automatic recovery and switch over.

The HA1 Frames has a left service bay and a right service bay. A second Library Manager and a second, inactive accessor are located in the right service bay as you view the library from the front (I/O station side). The second Library Manager is a “hot standby” unit that assumes control of the tape library automatically if the primary Library Manager fails. The second accessor is also held in “hot standby” status to assume the function if the first accessor fails. With optional feature code FC5050 (Dual Active Accessors), both accessors can be active at the same time.

The 3494 provides an automated tape solution for a variety of system environments. The wide range of configurations and options provides users with the flexibility to most effectively address their unique requirements.

The 3494 automates the retrieval, storage, and control of Cartridge System Tape, Enhanced Capacity Cartridge System Tape, High Performance Cartridge Tape, and Extended High Performance Cartridge Tape. When used with supporting software and appropriate tape subsystems, the 3494 allows cartridges to be mounted and demounted on tape drives without operator involvement.

The 3494 has the following key attributes:

- Automates cartridge tape libraries
- Provides expandable cartridge storage
- Allows additional tape drives to be added
- Selects requested cartridges and accesses tape volumes quickly
- Allows the intermix of Cartridge System Tape, Enhanced Capacity Cartridge System Tape, High Performance Cartridge Tape, and Extended High Performance Cartridge Tape in any frame
- Supports the 3490E Model C1A, C2A, and F1A tape subsystems
- Supports the 3490E Model F1A tape subsystem and the Model F1A FC 3000 and FC 3500 Controllers
- Supports the 3590 Model B1A and E1A tape subsystems and the 3590 Model A00, A50, and A60 Controllers

- Supports B16, B18, B10, and B20 VTSs
- Supports PtP VTSs (B18, B10, and B20 VTSs with FC 4010 [Peer-to-Peer Copy Base] and CX0s with AX0s)
- Supports VTS SCSI Host Attachment
- Supports VTS Export and Import operations
- Cleans tape drives automatically
- Incorporates enhanced tape drive error recovery
- Allows installation on solid or raised floors
- Provides high reliability and availability
- Enables enhanced remote service capability
- Provides unattended operation
- Allows the emulation of an automatic cartridge loader
- Supports stand alone applications and unlabeled cartridges
- Supports dual Library Manager controllers for greater availability and reduced service intervention
- Supports dual accessors for greater availability and reduced service intervention
- Supports dual active accessors for higher performance
- Provides Web-based user interfaces for access to status information

Attaching Magstar 3494 to a SAN environment

In Figure 4-15 we show the typical SAN distances using different SAN fabric components and Magstar 3494.

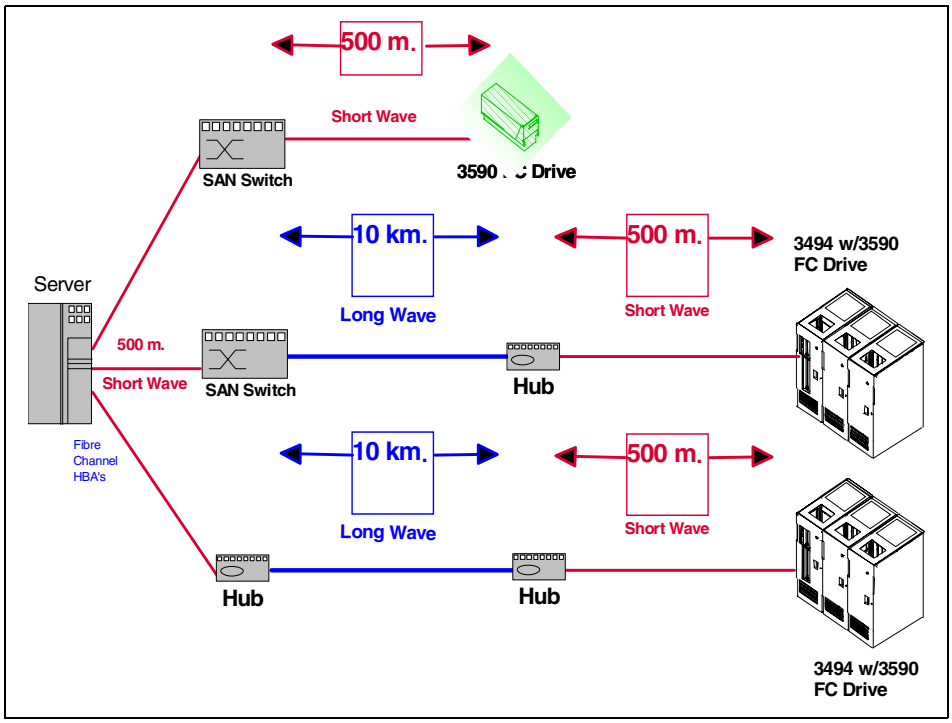


Figure 4-15 SAN distances and Magstar 3494 and 3590

For latest updates on connectivity and supported servers, refer to:

http://www.storage.ibm.com/hardsoft/tape/index.html

4.2.4 IBM TotalStorage Virtual Tape Server (VTS)

The IBM Virtual Tape Server is an enterprise tape solution designed to enhance performance and provide the capacity required for today's backup requirements. Adoption of this solution can help reduce batch processing time, total cost of ownership, and storage management overhead. Automation of a larger portion of a company's data storage may also be facilitated in a shorter period of time. The VTS has been designed to reduce or eliminate the number of bottlenecks that may be present in a given tape environment, depending on the characteristics of installed equipment and processor workloads. For example, if more drives are needed, up to 64 virtual drives may be configured to meet this need. If cartridge capacity is not fully utilized, the VTS can alleviate this problem by filling the physical cartridge.

In Figure 4-16 we show the VTS frame along with Magstar 3494.

Figure 4-16 IBM VTS along with 3494 tape library

The VTS initially creates a virtual volume in a buffer known as the Tape Volume Cache (TVC), a RAID5 disk array. If these virtual volumes are re-referenced, they are accessed in most instances from the TVC, helping to eliminate many of the physical delays associated with tape I/O and improving the performance of the tape process.

The virtual volume is also written to an attached IBM Magstar 3590 Tape Drive as a logical volume, in either a first-in/first-out order, or, if the optional Advanced Function feature is present, under the control of DFSMS.

Virtual volumes may, however, depending on the usable capacity of the TVC and their re-reference pattern, remain in cache for extended periods, which may provide faster access to critical volumes. Virtual drives can be dedicated to a specific processor or shared in supported environments. This flexibility maximizes the efficiency of data transfer operations by ensuring that sufficient drives are available for a specific task.

The Enterprise Tape Library Specialist is available to monitor the VTS. A Call-Home function is also provided to enable the VTS to perform proactive maintenance.

The Virtual Tape Server (VTS) product line has been continuously enhanced since its first availability in May 1997. New performance features have included the Extended High Performance Option (EHPO) which includes data compression, the SCSI Host Attachment feature, the data Import/Export feature,

Extended Performance ESCON Channels, the Performance Accelerator feature (PAF), extended host connectivity with up to eight ESCON channels and enhancements in Tape Volume Cache (TVC) capacity and performance on the IBM TotalStorage Virtual Tape Server.

The current general availability of the VTS includes a further optional enhancement for the Model B18 VTS, support for up to twelve physical tape drives, as well as two new models, the entry level Model B10 VTS and Model B20 VTS.

The Model B10 VTS and Model B20 VTS are based on the IBM server pSeries 660 architecture which features more powerful processors and expanded I/O capability. The Model B10 VTS is configured as a cost effective solution for modest throughput requirements while the Model B20 VTS establishes higher standards for throughput performance and for the number of logical volumes managed. To illustrate the continuous product enhancement, the Model B20 VTS has a host data throughput bandwidth up to twenty times that of the original Model B16 VTS in 1997.

The IBM TotalStorage Enterprise Automated Tape Library (3494) Library Manager allows storage of up to 500,000 logical volumes when two VTSs are included in the 3494 configuration.

Two VTSs can be coupled to participate in an Peer-to-Peer Virtual Tape Server (PtP VTS) environment for installations that require continuous access to data or improved disaster recovery operations.

The PtP VTS design reduces or eliminates single points of failure and improves data availability to address data access requirements during both planned and unplanned outages. PtP architecture allows physical separation of components in two sites to facilitate a more resilient disaster recovery operation.

IBM VTS supported servers

For ESCON attached servers, the VTS can be supported at a distance of up to 26 km using ESCON Directors, or up to 75 km using IBM 2029 Fiber Savers. In addition, the VTS can be supported at up to 10 km on a Storage Area Network (SAN) using the IBM 2109 SAN Fibre Channel switch and the IBM 2108 SAN Data Gateway.

- IBM S/390
- zSeries
- pSeries, RS/6000 servers
- SUN Microsystems
- HP 9000
- Microsoft Windows NT and 2000 servers

IBM VTS supported environments

- z/OS V1 or later
- OS/390® V2 or later
- DFSMS/MVS® V1.2+ or later
- z/VM V3 or later
- VSE/ESA V 2.2 plus PTFs+ is supported as a z/VM guest
- TPF V4.1 plus PTFs or later
- VM/ESA® V2.2 or later
- VSE/ESA V2.2 plus PTFs+ is supported as a VM/ESA guest
- AIX® V4.32+ or later
- Sun Solaris® V2.6+ or later
- Windows NT® / Windows® 2000
- HP-UX R11

IBM VTS SAN support

- 2108 SAN Data Gateway
- 2109 SAN Switch

Both the 2108 and 2109 are supported for AIX, Sun Solaris, Window NT and Windows 2000. VTS can be supported at up to 10 km on a Storage Area Network (SAN). This is shown in Figure 4-17.

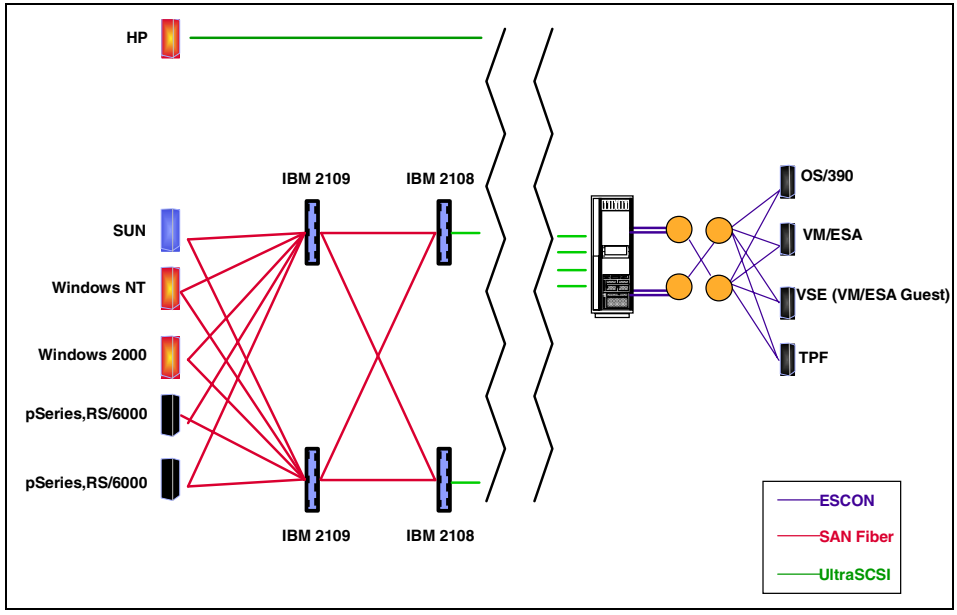


Figure 4-17 VTS SAN support

4.2.5 IBM TotalStorage Peer-to-Peer Virtual Tape Server (PtP VTS)

The Magstar Peer-to-Peer Virtual Tape Server, an extension of the Magstar Virtual Tape Server, builds on a proven base configuration to provide even greater benefits for tape processing and operations. By enhancing data backup and recovery capabilities, the Magstar Peer-to-Peer Virtual Tape Server is specifically designed to enhance data availability. It accomplishes this by providing dual volume copy, remote functionality, and automatic recovery and switch-over capabilities. With a design that reduces single points of failure — including the physical media where logical volumes are stored — the Magstar Peer-to-Peer Virtual Tape Server improves system reliability and availability, as well as data access. To help protect current hardware investments, existing Magstar Virtual Tape Servers can be upgraded for use in this new configuration.

The Magstar Peer-to-Peer Virtual Tape Server consists of new models and features of the Magstar 3494 Tape Library that are used to join two separate Magstar Virtual Tape Servers into a single interconnected system. The two virtual tape systems can be located at the same site or at different sites which are geographically remote. This provides a Remote Copy capability for remote vaulting applications.

The Peer-to-Peer VTS is a configuration of multiple 3494 Virtual Tape Servers with their independent underlying 3494 Tape Libraries and multiple Virtual Tape Controllers which functions as a single subsystem. It addresses data availability, system availability, remote copy and data vaulting concerns or desires for the VTS family. The fundamental design of the Peer-to-Peer VTS alters the current Virtual Tape Server products to a minimal extent; and, through the addition of controllers between MVS hosts and multiple VTSs, provides replication, transparent tracking and synchronization of logical tape volumes. The multiple VTSs and multiple controllers, with their interconnections, form a single Peer-to-Peer VTS.

The Virtual Tape Server was introduced to the market in May 1997. It provides the customer a revolutionary way to efficiently use tape systems. VTS has been well received by the customer and is a very successful product for IBM. VTS has (by design) had a major short coming. The availability of tape data on a VTS subsystem is significantly reduced from traditional tape subsystems. Historically, when a tape subsystem (control unit, drive, or library) fails, data which is critical to the enterprise is still available without the subsystem being repaired first. The customer (although this may be painful) could manually retrieve the tape volume where the critical data resides and move it to an available subsystem. With VTS, when a failure occurs which causes the VTS subsystem to become unavailable, all the data stored on the VTS is unavailable until the repair is complete.

Because of this, many VTS field problems result in critical situations for IBM product engineering to manage. For this reason some customers have been reluctant to adopt the revolutionary benefits VTS offers to the customers tape methodology.

The Peer-to-Peer VTS is a solution to this problem. In addition to providing a tape subsystem for mission critical data, it provides:

- Multiple copies of all tape data
- A solution for disaster recovery and electronic tape vaulting with transparent failover *and* failback

The performance, availability and recovery characteristics will depend on the mode of operation selected. A performance enhancement is gained in the Deferred Copy mode of operation by scheduling the creation of the dual copy as VTS activity permits. This however creates the possibility that access to a successfully written virtual volume is not possible when there is a single point of failure.

In this mode, on single points of failure, the system as a whole will continue to operate and provide access to most of the previously written virtual volumes The amount of data that is inaccessible will depend on the particular workload and how much time for background copying the system has had.

The next step up in availability reduces the performance enhancement while increasing the availability to data. By running the system in Immediate Copy mode, creation of the copy will begin when the Rewind Unload command is received for a virtual volume. Due to the overhead of making the copy, the performance under a single workload will be less than that of a single job on a standalone VTS. Under multiple workloads the system should outperform the same load on a standalone VTS due to the usage of a special data transfer capability that allows the copy to be made much quicker than the time to write the original. In Immediate Copy mode, on single points of failure, access to all virtual volumes that were previously written and closed prior to the failure is possible. However, jobs in progress may fail and the data written during the failing job will not be available until the failure is fixed. Likewise, data that is written to a VTS that has failing hardware not reporting errors may be lost even though it is only a single point of failure.

The Peer-to-Peer VTS has been designed so that:

- All components of the Peer-to-Peer VTS are duplicated
- A dual copy of the data is maintained by the subsystem

The Peer-to-Peer VTS is a classic SeaScape Architecture product which uses a significant amount of existing hardware and software.

New VTS models and features allow new configurations. Peer-to-Peer VTS configurations may be comprised of combinations of four or eight AX0 Virtual Tape Controllers, one, two, three or four CX0 Auxiliary Frames and two VTSs. The allowable combinations are shown in the table below. The number of virtual tape drives in the configurations is feature dependent and the number of tape drives addressable from the host per AX0 Virtual Tape Controller is a function of the number of virtual tape drives and number of AX0s.

In Table 4-4 we show peer to peer configuration possibilities.

Table 4-4 Peer-to-peer configuration

VTS Model	VTS Model	# of AX0	Virtual drives	Drive addresses per AX0
B18	B18	4	64	16
B18 FC5264(1)	B18 FC5264(1)	8	128	16
B18 FC5264(1)	B20	8	128	16
B10	B10	4	64	16
B10	B18	4	64	16
B20	B20	8	128	16
B20 FC5264(2)	B20 FC5264(2)	8	256	32

In Figure 4-18 we show the schematic layout of a Peer-to-Peer VTS with the two Virtual Tape Servers, each with its underlying tape library, and the Model AX0 Virtual Tape Controllers. All connections are ESCON links that may be connected through ESCON directors. The Peer-to-Peer VTS provides a single Virtual Tape Server image to the Host systems that is similar to the image provided by a Model B16, B18, B10 or B20 Virtual Tape Server. To the Hosts, the entire cluster appears as a single Tape Library Partition housed within a single Tape Library that may be shared in any of the same configurations that a VTS Model B16, B18, B10 or B20 may share within a Tape Library.

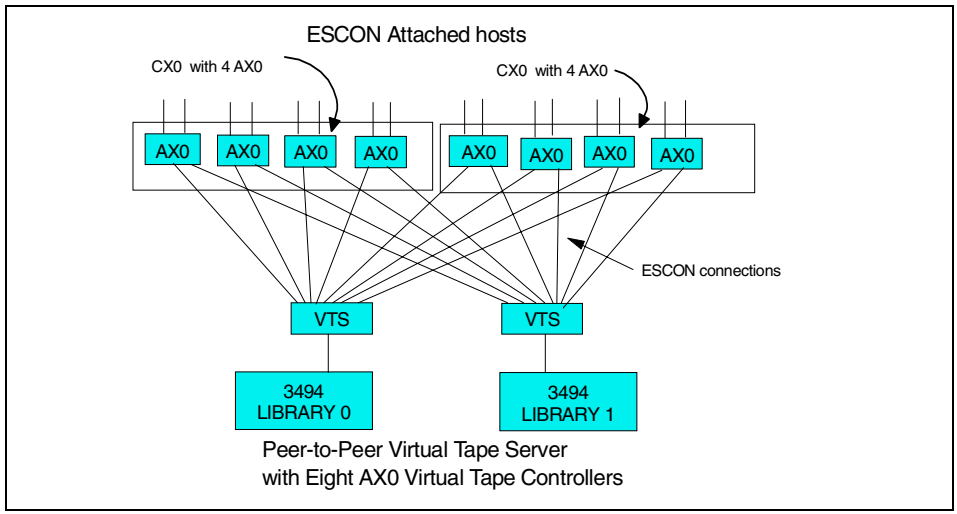


Figure 4-18 Peer-to-Peer VTS with two VTSs

The salient points of the physical configuration of a Peer-to-Peer VTS are:

- There are two Model VTSs whose only ESCON connections are to Model AX0 Virtual Tape Controllers. Each of these Virtual Tape Servers is associated with 3590 tape drives and have the Peer-to-Peer Copy features.
- Each of the VTSs with the Peer-to-Peer Copy features is connected to a different 3494 library. The libraries can have other partitions used by other VTSs with the Peer-to-Peer Copy features as well as any of the other supported configurations for VTS models sharing of a 3494 Tape Library. This feature can also be extended to two separate sites.

An example of this is shown in Figure 4-19.

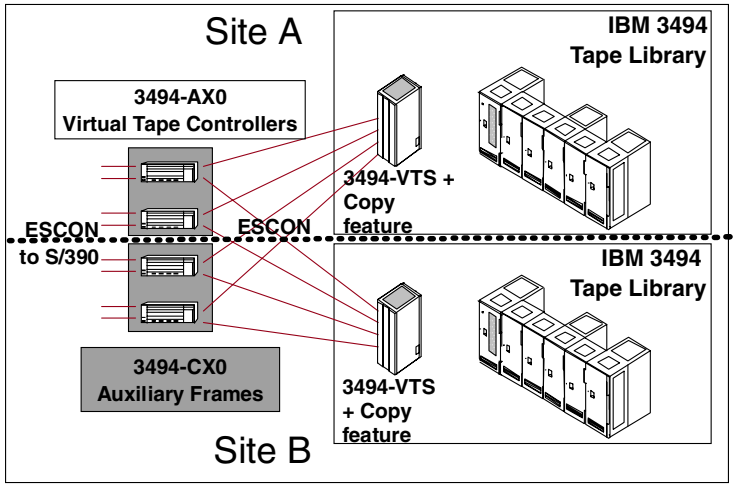


Figure 4-19 Peer-to-Peer VTS two sites

- Four or eight Virtual Tape Controllers (3494 Model AX0) are required for each Peer-to-Peer VTS depending upon number of virtual tape drives.
- Each Model AX0 Virtual Tape Controller has an ESCON connection to each VTS. Each of these connections may use ESCON directors to extend the distance between the controller and the VTS; however, only one dynamic director may be used.
- Each Model AX0 Virtual Tape Controller has two host ESCON connections in addition to the connections to the VTSs. These connections are the external interfaces of the single logical Tape Library image that a Peer-to-Peer VTS provides.
- The 3494 Model AX0 Virtual Tape Controllers are housed in a 3494 Model CX0 Auxiliary Frame. Each frame may contain two or four Model AX0s. Two AX0 controllers in a CX0 frame must be within the same Peer-to-Peer VTS. An additional two AX0s (for a maximum of four) in the CX0 frame may be used for another Peer-to-Peer VTS, or the four AX0s in a CX0 frame may be in the same Peer-to-Peer VTS. A Peer-to-Peer VTS may use one, two, three or four CX0 frames with each frame having two or four AX0s; however, the total number of AX0s must be four or eight.

In Figure 4-20 we show the AX0 and CX0 frames on a Local PtP and remote PtP configurations.

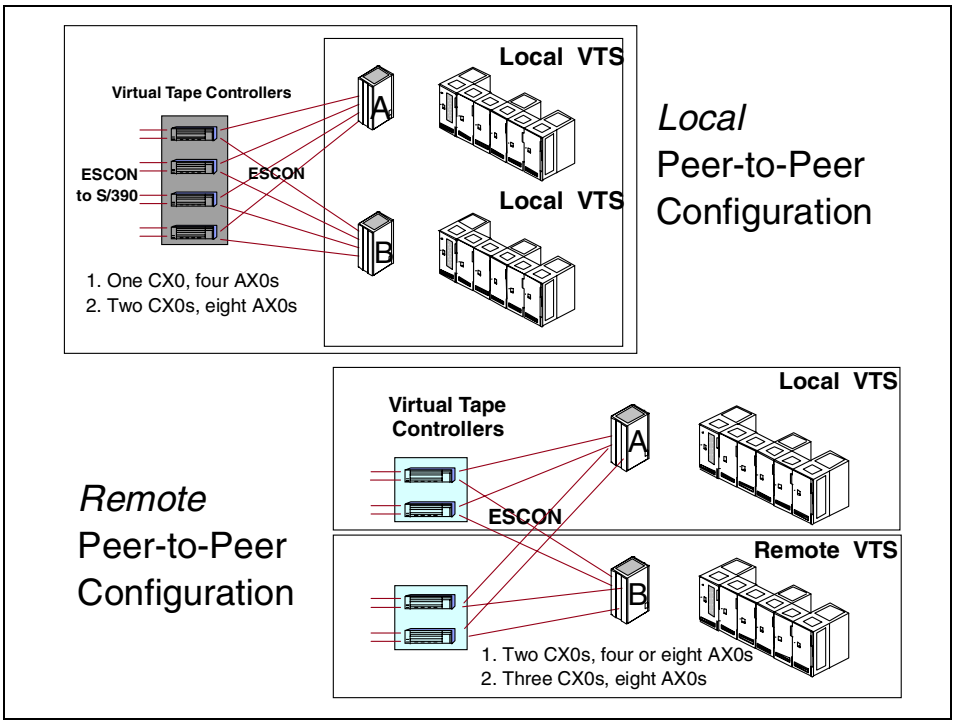


Figure 4-20 Local and remote peer-to-peer configurations

- The Model CX0 Auxiliary Frames may be placed up to 26 km (16 miles) away from the Virtual Tape Servers with appropriate ESCON directors and cables. Through use of the 2029 Fiber Saver, CX0 frames and VTSs may be separated by 25 km (15.6 miles). They may also be placed 43 km (27 miles) away from the hosts with appropriate ESCON directors and cables and 75 km (46.8 miles) away from the hosts with 2029 Fiber Savers.

In Figure 4-21 we show an example of this with ESCON directors.

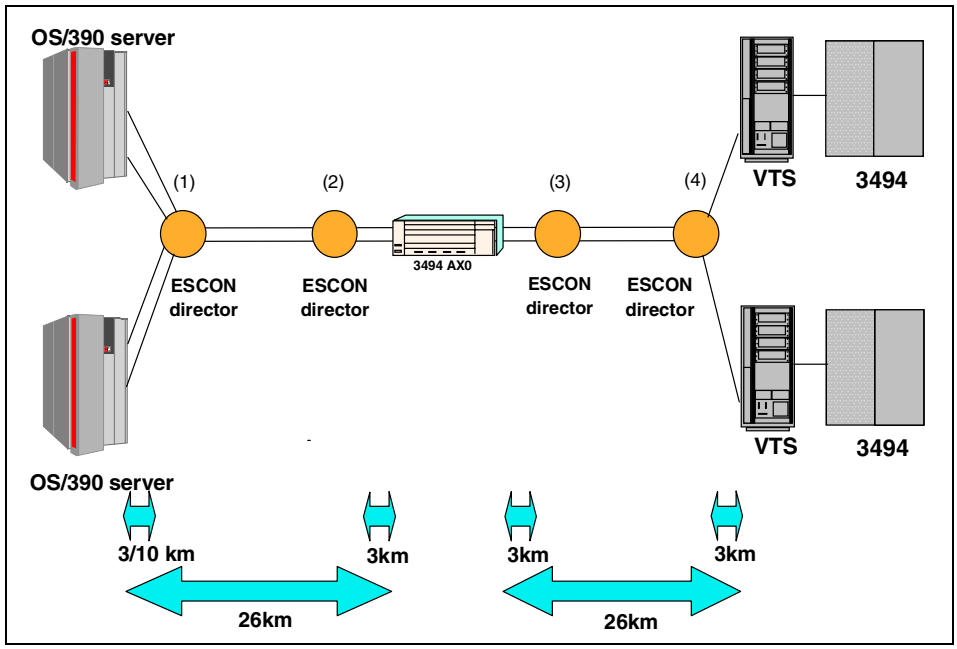


Figure 4-21 Peer-to-Peer VTS with ESCON directors

In Figure 4-22 we show an example using channel extenders.

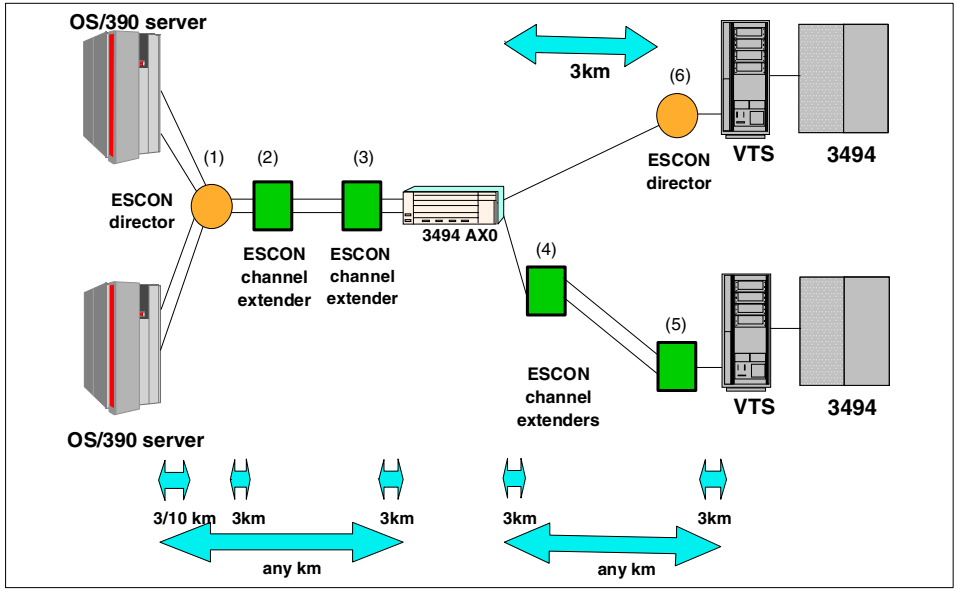


Figure 4-22 Peer-to-Peer VTS with channel extenders

- Each Model AX0 provides a connection for a LAN with Internet protocols. When this connection is used, each of the Model AX0 Virtual Tape Controllers provide a Web Page that may be used to access status of the Peer-to-Peer VTS and view statistics. The LAN must be provided and managed separately from the Peer-to-Peer VTS and is not required; however, it is recommended.
- Each AX0 Virtual Tape Controller or VTS model may attach to the TotalStorage Service Facility to enable enhanced remote service.

Although a Peer-to-Peer VTS consists of Virtual Tape Controllers and two Virtual Tape Servers, it appears logically and is used as a single Virtual Tape Server. This logical VTS provides four, eight or sixteen logical control units each handling sixteen logical tape drives. The installation is provided with two physical ESCON connections into each logical Virtual Tape Controller. For up to 128 virtual tape drives, each Virtual Tape Controller is a separate logical control unit with two ESCON attachments. When features are installed to provide 256 virtual tape drives, each of the ESCON attachments to the Virtual Tape Controller becomes a control unit image with sixteen tape drives: that is, two control unit images and 32 virtual tape drive addresses per Virtual Tape Controller.

VTS functional modes

There are two modes of operation concerning data replication:

- **Deferred Copy mode**: In this mode the copy is created in the background as VTS activity permits after a Rewind Unload command is received for a virtual volume and the unload is complete.
- **Immediate Copy Mode**: In this mode, when the job writing a virtual volume sends the Rewind Unload command, the Peer-to-Peer VTS begins making the copy. When there are no failures, the system does not respond successfully to the Rewind Unload until the copy is made. Therefore, the system may see additional time-outs on Rewind Unloads. This increases the Batch window time when compared to running the same batch job sequence to a single VTS Model B18 (with the Performance Accelerator feature), Model B10 or Model B20. If multiple batch jobs are run simultaneously to different drive addresses of the Peer-to-Peer VTS, total time to complete all the jobs will be improved from that of running the same multiple batch jobs to a single VTS.

In this redbook we will show how these components can be used in a SAN distance solution.

5

SAN fabric solutions at a distance

The future of Storage Area Networks is characterized by three trends that highlight the importance of extending storage over the Wide Area Network (WAN).

Firstly, the vast majority, approximately 80% according to sources, of SANs are isolated islands of data where there is only local access to the stored data. This is inconsistent with your customers' new paradigm of a digital business model where everybody wants access to all data all the time.

Secondly, customers are demanding a cost-effective way of consolidating storage to alleviate the growing discrepancy between the need for endless additional Terabytes of capacity and IT budgets that are just growing marginally. An analysis suggests that storage consolidation can cut storage costs by as much as 30 percent.

Finally, there is a strong trend towards outsourcing storage. IBM's customers are expecting to bring them not just the technology to allow for local storage, but also the technology that will allow them to move storage into managed data centers. Outsourcing storage will allow your customers to save an additional 20 to 30 percent on storage costs.

In this chapter we review some of the important SAN fabric solution characteristics and terminologies as applicable for distance solutions. We also visit some examples of typical SAN fabric distance solutions.

For additional details in designing a SAN fabric and detailed component characteristics, refer to these redbooks:

- *Designing an IBM Storage Area Network*, SG24-5758
- *IBM SAN Survival Guide*, SG24-6143
- *Introduction to Storage Area Network, SAN*, SG24-5470
- *Implementing an Open IBM SAN,* SG24-6116

5.1 SAN topologies

Fibre Channel provides three distinct interconnection topologies. By having more than one interconnection option available, a particular application can choose the topology that is best suited to its requirements. The three Fibre Channel topologies are shown in Figure 5-1.

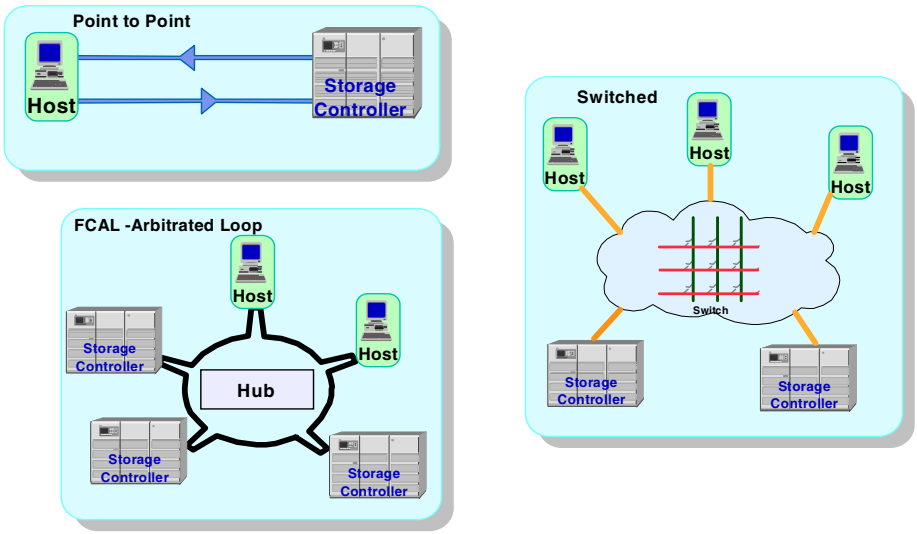


Figure 5-1 SAN topologies

- **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.

- **Arbitrated loop**

  Also called as Fibre Channel Arbitrated Loop (FC-AL). FC-AL is more useful for storage applications. It is a loop of up to 126 nodes (NL_Ports) that is managed as a shared bus. Traffic flows in one direction, carrying data frames and primitives around the loop with a total bandwidth of 100 MB/s. Using arbitration protocol, a single connection is established between a sender and a receiver, and a data frame is transferred around the loop. Loops can be configured with hubs to make connection management easier. A distance of up to 10 km is supported by the Fibre Channel standard for both of these configurations. However, latency on the arbitrated loop configuration is affected by the loop size.

- **Switched fabric**

  The third topology used in SAN implementations is Fibre Channel Switched Fabric (FC-SW). A Fibre Channel fabric is one or more fabric switches in a single, sometimes extended, configuration. Switched fabrics provide full 100MB/s bandwidth per port, compared to the shared bandwidth per port in Arbitrated Loop implementations.

5.2 Fiber optic interconnects

In Fibre Channel technology, frames are moved from source to destination using gigabit transport, which is a requirement to achieve fast transfer rates. To communicate with gigabit transport, both sides have to support this type of communication. This can be accomplished by installing this feature into the device or by using specially designed interfaces which can convert other communication transport into gigabit transport. Gigabit transport can be used in a copper or fiber-optic infrastructure. We recommend that you consider using a fiber-optic implementation if you need to avoid the distance limitation of 30 meters with copper, or are likely to in the future.

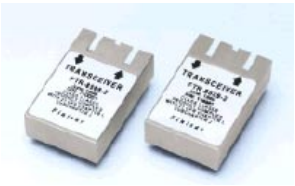
The interfaces that are used to convert the internal communication transport of gigabit transport are shown in Figure 5-2.

Small form factor transreciever

Gigabit interface convertor

Gigabit link module GLM

Media Inteface adpater MIA



1*9 Transcevier

Figure 5-2 Fiber optic interconnects

- **Small Form Factor Transceivers (SFF)**

  The IBM 1063 Mb/s up to 2125 b/s Small Form Factor Transceivers (SFF) 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. The SFF optical transceivers use short wavelength and long wavelength lasers and are available in pin through hole (PTH) or hot-plugged versions. The Small Form Factor Hot-Plugable module is also known as SFP.

- **Gigabit Interface Converters (GBIC)**

  The IBM 1063 Mb/s and 1250 Mb/s Gigabit Interface Converters (GBICs) are laser-based, hot-plugged, data communications transceivers for a wide range of networking applications requiring high data rates.

The transceivers, which are designed for ease of configuration and replacement, are well-suited for Gigabit Ethernet, Fibre Channel, and 1394b applications. The GBICs are available in both short wavelength and long wavelength versions, providing configuration flexibility. Users can easily add a GBIC in the field to accommodate a new configuration requirement or replace an existing device to allow for increased availability. The GBICs use lasers that enable cost-effective data transmission over optical fibers at distances of up to 10 km. These compact, hot-pluggable, field-replaceable modules are designed to connect easily to a system card through an industry-standard connector.

- **Gigabit Link Modules (GLM)**

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

- **Media Interface Adapters (MIA)**

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

- **1x9 transceivers**

  Some of the switch manufacturers prefer to use 1x9 transceivers for providing SC connection to their devices. 1x9 transceivers have some advantages over GBICs, which are the most widely used in switch implementations.

5.3 SAN fabric distances

Fibre Channel allows for much longer distances than the 25 meters limit of SCSI links. The longer distances are made possible with different GBICs, and currently supported distances are with:

- Short wave GBIC, using 850nm laser: 500 meters using 50 micron multimode fiber cable and 175 meters using 62.5 micron multi mode fibre cable
- Long wave GBIC, using 1310nm laser: 10 kilometers using 9 micron single mode fiber cable.

- Extended or EL GBIC, using 1550 nm laser: 100 km using 9 micron single mode fiber cable. The Extended GBICs are made by Finisar. Refer to http://www.finisar.com for additional details.

When longer distances are required there are different options, for example, extenders, protocol converters, or Dense Wave Division Multiplexors (DWDM), and selection will depend on the available links between the two locations, distance and budget.

Some distance solutions convert Fibre Channel protocol, FCP, to several OC3 or ATM channels, route the signals through telecommunications (telco) lines and reconvert the signals at the other end. These can reach hundreds or thousand of miles. By using repeaters and dedicated fibers we can get distances of about 100 km. DWDM allows us to send several channels over the same fiber.

5.3.1 Cable types

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. Every data communications fiber falls into one of two categories:

- Single-mode
- Multi-mode

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

In most cases, it is impossible to distinguish between single-mode and multi-mode fiber with the naked eye. Most manufacturers now follow the color coding schemes specified by the *Fibre Channel physical layer working subcommittee*, which is orange for multi-mode and yellow for single-mode.

IBM supports the following distances for fibre optic cables, the supported distances are based on GBIC technologies:

- 50 Micron Multimode Shortwave <= 500 meters
- 62.5 Micron Multimode Shortwave <= 175 meters
- 9 Micron Single mode Longwave =< 10 km

5.3.2 Dark fiber

Dark fiber is a dedicated end-to-end fiber that can be used without additional equipment up to 10 kilometers for long wave transceivers, or may require the use of extenders or repeaters, either external or internal in some directors, for longer distances. By using dark fiber, we get the most direct connection and full bandwidth, but the down side is the cost of the dedicated fiber links.

5.3.3 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 additional details refer to Chapter 2, “Extending fiber over distance with DWDM” on page 13.

5.4 Port types

The types of Fibre Channel port that are likely to be encountered are:

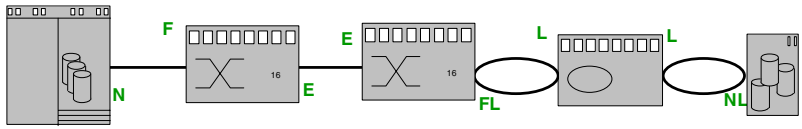
- E_Port is an expansion port. A port is designated an E_Port when it is used as an interswitch expansion port to connect to the E_Port of another switch, to build a larger switched fabric. These ports are found in Fibre Channel switched fabrics and are used to interconnect the individual switch or routing elements. They are not the source or destination of information units (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.
- 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 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 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.
- 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.
- 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 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 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 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 is an inter switch link (ISL) port more commonly known as an E_Port.
- TL_Port is a private to public bridging of switches or directors.

The Figure 5-3, shows the various commonly used ports and their connectivity.



FL_Port
Fabric Loop Port

NL_Port
Node Loop Port

L_Port
Loop Port

F_Port
Fabric Port

E_Port
Expansion Port

N_Port
Node Port

Figure 5-3 SAN ports

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

EE_Credit

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

During data transmission a port should not send more frames than the buffer of the receiving port can handle before getting and indication from the receiving port that it has processed a previously sent frame.

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

This is particularly true with distance. At 1 Gb/s a frame occupies 4 km of fiber. In a 100 km link we can send 25 frames before the first one reaches destination. We need an ACK back to start replenishing EE_Credit. We will be able to send another 25 before we receive the first ACK. We need at least 50 buffers to allow for non stop transmission at 100 km distance.

Most fabric vendors offer over 60 buffer credits per port on their equipment. Some manufacturers however offer the added number of buffers as an optional feature. The IBM TotalStorage SAN Switch S08 and S16 FC switch (Brocade), offers feature code 7303 for extended fabric activation. IBM 2042 directors (INRANGE FC/9000) offers 64 buffer credits on all its ports by default. IBM 2032 (McDATA) director offers 60 buffer credits per port by default.

5.6 Optical link budgets for SAN fabric

As mentioned earlier, multiple connectors are recognized by the standard giving the physical installation design flexibility and some assurances of repeatability. The performance of the different fiber-optic cable types is also identified in the standard. Therefore, flexibility is provided to the physical infrastructure designer. The achievable link distances for the different cable types is identified in Table 5-1.

Table 5-1 Optical link budgets

	Singlemode			**Multimode - 50**			**Multimode - 62.5**		
Data rate (MB/second)	400	200	100[a]	400	200	100[a]	400	200	100[a]
Modal bandwidth (MHz km)	N/A	N/A	N/A	500 / 500	500 / 500	500 / 500	200 / 500	200 / 500	200 / 500
Cable plant dispersion (Ps/nm km)	12	12	12	N/A	N/A	N/A	N/A	N/A	N/A
Transmitter spectral center wavelength (nm)	1310	1310	1310	770 -860	830 -860	830 -860	770 -860	830 -860	830 -860
Operating range (m)	2-10,000	2-10,000	2-10,000	2-150	2-300	2-500	2-70	2-150	2-300
Loss budget (dB)	7.8	7.8	7.8	2.06	2.62	3.85	1.78	2.1	3.01

a. 100 MB/s products were available at time of writing. Other baud rate information is included for growth planning purposes.

The overriding trend is decreased distance and loss budgets as the data rate increases for the same type of cable. Looking at the multi-mode cables, 50 micron fiber enables significantly longer links than the 62.5 micron fiber at the same data rate. Lastly, the single mode distances are generally not affected by the increased data rates. It is also noted in the standard that lower performance multi-mode fibers have been installed in the past and performance will be affected. It should also be noted that higher performance multi-mode cables have been available since the beginning of 1999 which allows increased distances. The actual performance will need to be obtained from the manufacturer.

The bandwidths, distances, and losses identified in Table 5-1, are those specified in the ANSI Fibre Channel Physical Interface document.

It is interesting to note the different drive distances and link budgets for the many protocols over the same fiber types. Consideration should be given to these different options as more network protocols make their way onto the same fiber cable infrastructure.

Also, it should be pointed out that fibers with higher modal bandwidth than specified in the various standards are now available, and, therefore, will allow longer distance links. Exact performance details will have to be obtained from the manufacturers. When this increased capability is used in planning a link ensure that appropriate documentation is provided for future reference.

The link budget for the different bit rates in Fibre Channel makes designing systems that have room for growth significantly difficult. Therefore, a thorough assessment and understanding of current and future goals must be considered and all designed capabilities be clearly communicated. Furthermore, the decision between multi-mode and single-mode solutions may require cost analyses in more cases to offer solutions that meet current needs and provide for future growth.

One other consideration to assess with an installed fiber-optic system is the issue of polarity. Many systems designed with the idea of a duplex connector system enforce the correct polarity to ensure the transmitter of one device connects to the receiver of the other. IBM Fiber Transport Services (FTS), for example, provides this feature throughout. Other generic fiber-optic cabling systems that provide connectivity within a building, on a campus, or distance connections such as dark fiber from a local exchange carrier, may not provide this feature automatically. So any design needs to identify to the installers that polarity be maintained throughout the system.

To summarize the situation with existing infrastructure, there essentially is no easy answer. Each currently installed system would need to be assessed for applicability with Fibre Channel. If the data for the fiber-optic cables and connectors is known, then the decision can be made relatively easily. If the information is not known, bandwidth and loss measurements may be required of each link in question.

Since there are no simple methods to measure bandwidth in the field, as of yet, loss measurements may be the only information that will be useful with 100MBs multi-mode solutions. When the cost of testing is weighed against the cost of installing a new cabling, a new system may actually cost less than measuring and replacing the currently installed system.

In conclusion, each situation will need to be decided on its own merits between the client and the connectivity specialist.

5.7 Hierarchical design

What we have seen is that a SAN can take numerous shapes. When you start thinking about SAN design for your own organization you can learn from the experience gained in the design of other, mature networks such as LAN, and the Internet. In these, a hierarchical network structure has generally been adopted, to facilitate change, allow easy replication as the structure grows, and minimize costs. This hierarchy comprises three layers as shown in Figure 5-4.

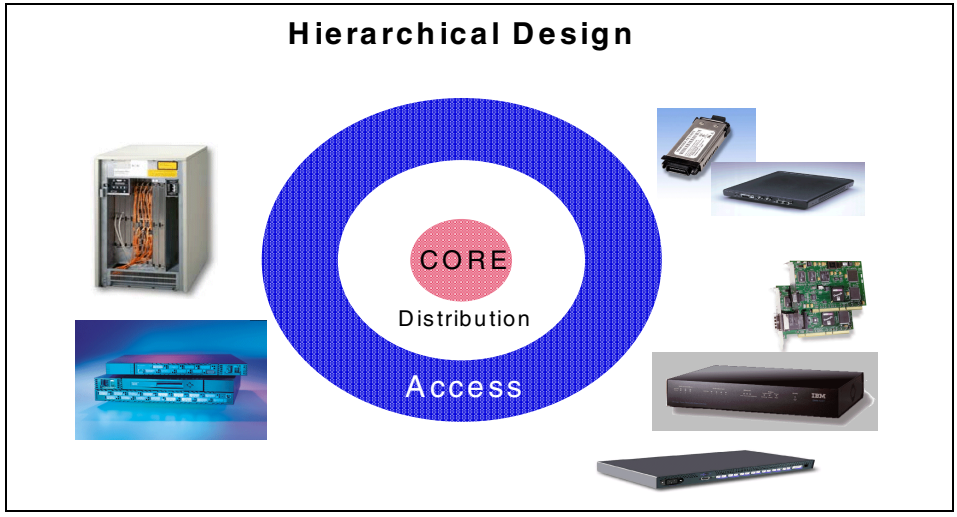


Figure 5-4 SAN hierarchical design

- **The core**

  At the center is a high speed, fault tolerant backbone, which is designed to provide very high reliability. This is designed to minimize latency within the fabric, and to optimize performance. This core would normally be built around fault tolerant Fibre Channel directors or a fully redundant, meshed topology of switches, like the IBM TotalStorage SAN Switch S08 and S16.

- **The distribution layer**

  The distribution layer of the hierarchy would comprise fault resistant fabric components. Good connectivity and performance would be prime considerations.

- **The access layer**

  Here are the entry point nodes to the fabric, comprising host bus adapters, routers, gateways, hubs, and switches appropriate to service the number of servers and storage devices supported on the fabric.

This hierarchy is analogous to the telephone switching system. Each user has access to the network using an individual node (the telephone); these link to local area switches, which in turn link to central core switches which serve a large national and international network with very high bandwidth. A similar hierarchy has been built to serve the Internet, with end users linked to local Web servers, which in turn communicate with large scale, high performance.

Planning considerations and recommendations

Many miscellaneous considerations are needed to successfully install fiber-optic links for any protocol. However, the higher data rate and lower optical link budgets of Fibre Channel lends itself to more conservative approaches to link design. Some of the key elements to consider are:

- All links must use the currently predominant physical contact connectors for smaller losses, better back reflectance, and more repeatable performance.
- The use of either fusion or mechanical splices is left to the designer to determine the desired losses weighed against the cost of installation.
- Multi-mode links cannot contain mixed fiber diameters (62.5 and 50 micron) in the same link. The losses due to the mismatch may be as much as 4.8 dB with a variance of 0.12 dB. This would more than exceed the small power budgets available by this standard.
- The use of high quality factory terminated jumper cables is also recommended to ensure consistent performance and loss characteristics throughout the installation.
- The use of a structured cabling system is strongly recommended even for small installations.

- A structured cabling system provides a protected solution that serves current requirements as well as allows for easy expansion.
- The designer of a structured system should consider component variance affects on the link if applicable.

Much of the discussion so far has been centered around single floor or single room installation. Unlike earlier FDDI or ESCON installations that had sufficient multi-mode link budgets to span significant distances, Fibre Channel multi-mode solutions for the most part do not. Though the Fibre Channel standard allows for extend distance links and handles distance timing issues in the protocol the link budgets are the limiting factor.

Therefore, installations that need to span between floors or buildings will need any proposed link to be evaluated for its link budget closely. Degradation over time, environmental effects on cables run in unconditioned spaces, as well as variations introduced by multiple installers need to be closely scrutinized. The choice between single-mode and multi-mode devices may need to be made for many more links. Repeating the signal may also provide a cost effective solution if intermediary conditioned space can be found.

In addition to its 10 km link distances using single-mode fiber, there will be more consideration for off-campus or across city links. In these cases, right-of-way issues, leasing of dark fiber (no powered devices provided by the lessors) issues, service level agreements, and other factors associated with leaving the client owned premises needs to be planned for and negotiated with local providers. The industry has also announced interest in providing wide area network (WAN) interfaces similar to those employed in the networking world of today. When these devices are made available, then connections to these devices will need to be included in the designs as well.

5.8 Defining the infrastructure requirements

If you are starting from scratch with a totally new network in a green field site then you can go straight ahead with selection of the optimum SAN topology to meet your needs. But in most situations it is likely that you are replacing an existing infrastructure for storage. You may even be planning to change or upgrade an existing SAN implementation. So, before selecting a design for the new SAN, it makes good sense to fully understand what it is that is being replaced. The current storage configuration, LAN or SAN network structure, application uses, traffic loads, peak periods and performance, as well as current

constraints, are all relevant information in determining realistic goals for the SAN. This information will also help you to determine what, if any, of the existing components can be used in a new topology; and what will be involved in migrating from today's environment to the new one.

5.8.1 Use of existing fiber

In many cases you may already have fiber-optic cables laid in your organization. IT budget holders will want to know if you can use the existing cabling. If the existing cabling has been laid for some time the answer may well be that the high speeds and accuracy required of Fibre Channel requires new cable investments. It is possible to test if installed fiber meets the necessary quality, but this can also be a costly exercise. If recent fiber cable has been laid you may need to decide what extensions need to be added to the configuration.

5.8.2 Application traffic characteristics

Before selecting a SAN topology 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?

5.8.3 Platforms and storage

How many servers and what are the operating platforms which 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 (combining AIX, UNIX, Windows NT and so on). This has implications for security.

Fibre Channel capable servers require Fibre Channel HBA 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. Sun SBus operates at 50 MB/s, and 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 over run 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.

5.8.4 Service requirements

An important criterion for selection of SAN components relates to the level of service required from the SAN. This includes all aspects of the technology (hub, switch or director), the topology (loop or fabric), and the degree of redundancy, including fault tolerance. This is particularly relevant for organizations serving the global marketplace 24 hours per day, seven days per week over the Internet. In the e-business economy of today, continuous availability is not optional. If you are not online, you are not open for business, and widely reported incidents of system outages in well known e-business companies show that loss of revenue can be immense.

The term “system availability” is commonly used when defining service levels. These are normally described in terms of percentage systems availability.

A 99.999% (five 9s) up time refers to achievement of less than five minutes systems downtime in one year. A one 9 measure refers to a 90% availability (less than 36.5 days systems downtime), and a three 9s level is 99.9% uptime (less than 8 hours 45 minutes systems downtime annually). Downtime can be defined as any complete interruption of service for any reason, whether planned or unplanned.

To meet the very high levels of uptime required by planners and administrators, it is essential to design the correct network architecture. It needs built-in fault tolerance, fail-over capabilities, and available bandwidth to handle unplanned outages in a transparent manner.

High availability can be built in to the fabric by eliminating single points of failure. This is achieved by deploying hardware components in redundant pairs, and configuring redundant paths. Redundant paths will be routed through different switches to provide availability of connection. In the event of a path failure (for instance due to HBA, port card, fiber-optic cable, or storage adapter) software running in the host servers initiates failover to a secondary path. If the path failover malfunctions the application will fail. Then the only choice is to repair the failed path, or replace the failed device. Both these actions potentially lead to outages of other applications on multiple heterogeneous servers if the device affected is the switch.

Switches, like the IBM TotalStorage SAN Switch S08, S16 and F16, have redundant, hot-plugged components (including fans, power supplies, ASICs and GBICs), which can be replaced during normal operation. These hardware failures cause little or no noticeable loss of service. However, in the case of some failed components (such as the mother board) the switch itself will be treated as the field replaceable unit (FRU). Then all the ports and data paths are taken down. Automatic path failover will occur to another switch, so the network continues to operate, but in degraded mode.

Here there is a distinction between a switch and a director. Using the analogy of disk arrays, an individual switch could be likened to a JBOD in that it is just a bunch of ports. That is to say, although it has redundant components, in the event of certain component failures the total switch can fail, or must be replaced as the FRU. Cascading of multiple switches can achieve a higher level of fault tolerance. A single director could be viewed more like a RAID subsystem, in that it is designed to be highly fault tolerant. Only the failure of the mother board would result in total failure of the director. All other components are redundant, with automatic failover. Redundant field replaceable units are hot swappable, and microcode updates can be made non-disruptively. Maintenance capabilities, such as call-home are supported.

5.8.5 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 “connection less”. 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 connection less” service. It optimizes the use of fabric resources, but it is now up to the upper layer protocol to insure 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.

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

  IBM TotalStorage SAN Switch S08, S16 and F16, INRANGE and McDATA directors support Class 2 and Class 3 service. The IBM TotalStorage Enterprise Storage Server (ESS) also supports Class 2 and 3 service.

- **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.

5.9 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. Zoning 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.

Figure 5-5 shows and example of two zones, Zone A and Zone B, each serving a different host.

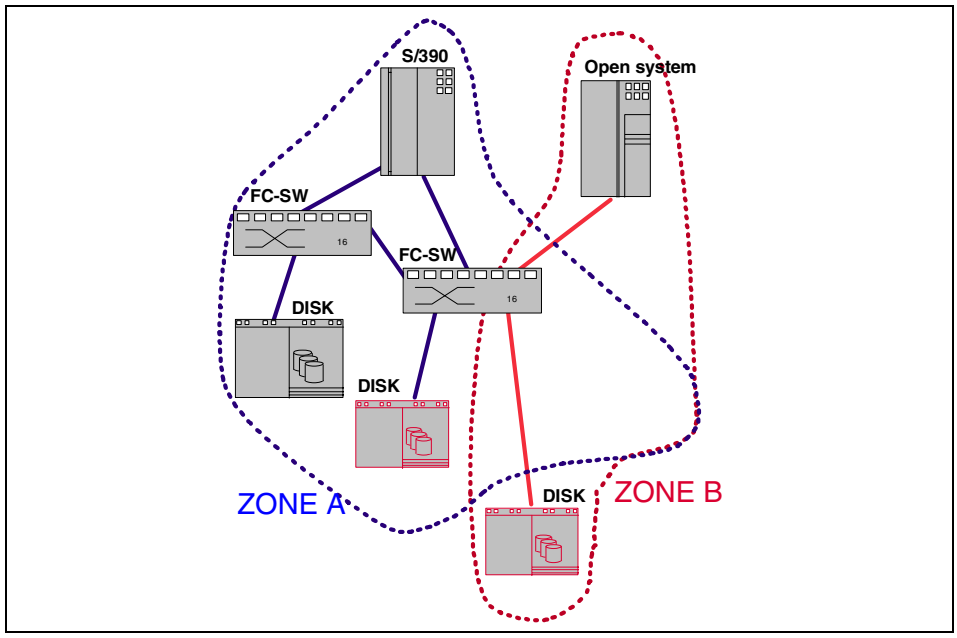


Figure 5-5 Zoning

5.10 SAN software management standards

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.

The elements that make up the SAN infrastructure include intelligent disk subsystems, intelligent removable media subsystems, Fibre Channel switches, hubs and bridges, metadata 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.

Fabric monitoring and management is an area where a great deal of standards work is being focused. Two management techniques in use are:

- In-band management

- Out-band management

In-band management as the name suggests means that 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 in-band management. It is simple to implement and requires no LAN connections.

Out-band 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).

5.11 SAN distance solution examples

In this section we briefly describe some examples of SAN distance solutions using the SAN fabric switches and directors along with additional equipment like DWDM and channel extenders.

5.11.1 Remote disk

In this solution the basic objective is to physically separate the disk subsystem from the servers as shown in Figure 5-6.

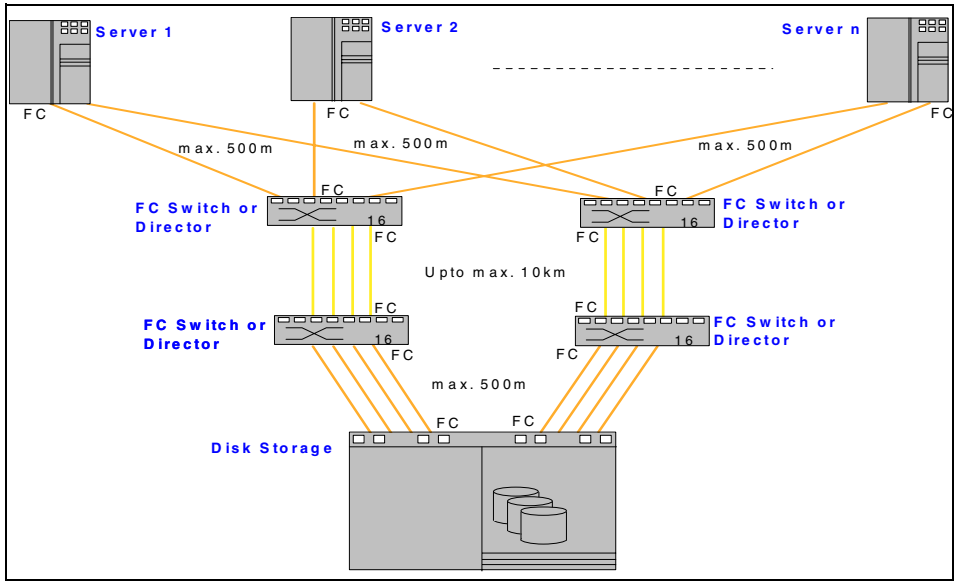


Figure 5-6 Typical remote disk solution

In this the Longwave GBICs along with Single-mode fibre cables are between the 2 switches or directors. In the case of switches, the extended fabric feature code is required to be installed to make use of the all the buffer credits for maximum throughput. This solution may be implemented with remote tape vaulting, where a tape library and a backup server are also put in at the remote site.

5.11.2 Mirroring and disaster tolerance solution

As shown in Figure 5-7 we describe a basic mirroring and disaster tolerance solution by protecting primary data using remote mirror and “hot stand-by” disaster recovery site.

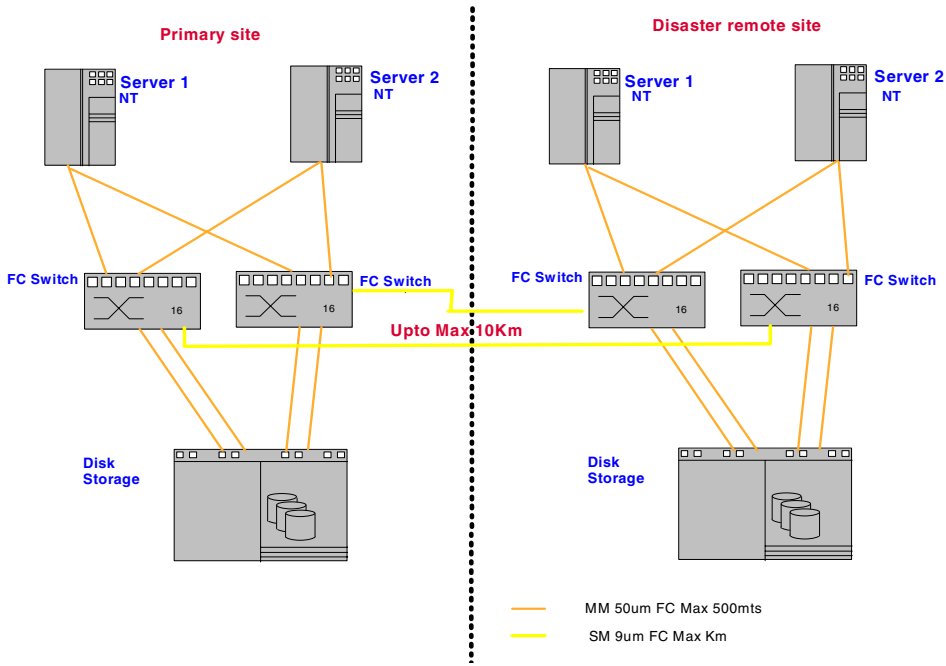


Figure 5-7 Typical disk remote mirroring and disaster tolerance

If the primary site fails, a remote system takes over (“imports”) the storage volumes. The services can be manually started on the remote system and have access to the mirrored data.

IBM SAN Fibre Channel switch’s solution provides for Veritas Volume Manager synchronous mirroring at full Fibre Channel speed at distances of up to 10 km. Achieving full Fibre Channel speed at distances of up to 10 km is possible due to the switch’s large number of buffers (credits) available at the E_Ports or ISLs.

As this is a highly available solution, an alternate redundant path is connected from the primary site to the disaster recovery site.

Veritas Volume Manager is used to mirror the content from disk storage in the primary location to the disaster recovery location. If the primary site fails, the disaster recovery site will be activated.

This primary site may be shielded from the disaster recovery site using zoning. As the disaster recovery site is used to mirror data from the primary site, and have access to the primary site. However, users from primary site will not be able to access to the disaster recovery site. In the event that the primary site's storage fails, zoning information from the disaster recovery site may be propagated to the primary site to allow access to the disaster recovery site's storage.

However, should the entire primary site fail, you will need to shift the disaster recovery site.

Extended distances (20 to 120 km) are possible, using optical extender devices such as CNT gateway, as shown in Figure 5-8.

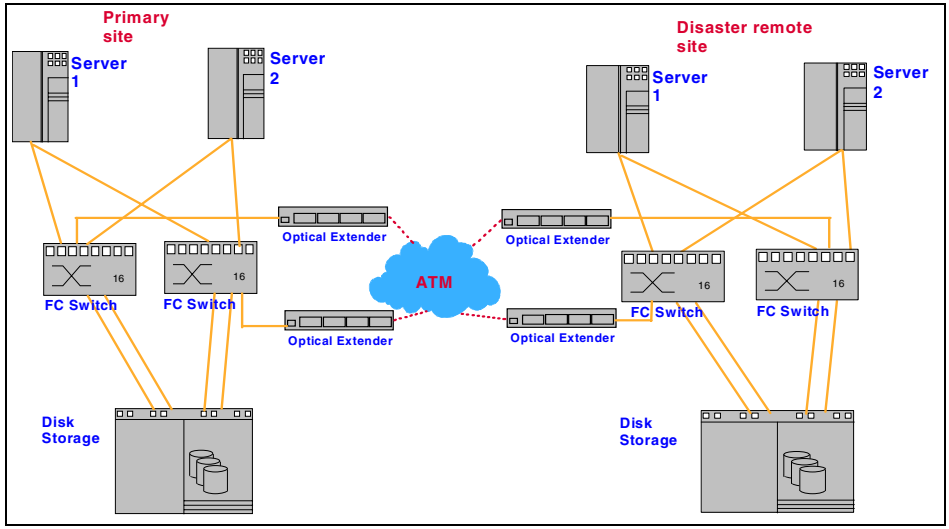


Figure 5-8 Disaster tolerance using ATM extenders

We provide more comprehensive solutions in Part 2, "Solutions" on page 303.

6

Cisco ONS 15540 ESP

The Cisco ONS 15540 Extended Services Platform (ESP) is a highly modular and scalable next generation DWDM platform that delivers the integration of data, storage and metro networking over an ultra high bandwidth.

Enterprise users of the Cisco ONS 15540 ESP will benefit from a single optical infrastructure that scales to meet the explosive demand of storage area networking (SAN) and metropolitan area networking (MAN) and supports 10 Gigabit Ethernet connectivity for streamlined operational efficiency, flexibility and cost savings.

With the Cisco ONS 15540, Service Providers will be able to rapidly create and provision new service offerings while realizing greater revenue and profitability per wavelength for a competitive advantage.

The Cisco ONS 15540 is shown in Figure 6-1.

Figure 6-1 Picture of Cisco ONS 15540

6.1 Componentry

The Cisco ONS 15540 has a modular architecture that allows flexibility in configuration and permits incremental upgrades of the system. These components are described in the following sections.

The chassis has a total of 12 slots. Each chassis supports up to 32 lambdas. Two slots are dedicated to the optical multiplexer/demultiplexer, eight slots are dedicated to line cards, and the remaining two slots are dedicated to two redundant CPU's. Each line card can house up to four hot-pluggable transponder modules that operate from 16 Mb/s to 2.5 Gb/s. The redundant CPUs provide management and switch fabric for the services supported in the Cisco ONS 15540.

6.1.1 Transponder modules

The transponder modules populate the line card motherboards and have two interfaces: an external interface that connects to client equipment and an internal interface that connects to the line card motherboard.

Table 6-1 shows the available modules and the associated bands.

Table 6-1 Modules and associated bands

Cisco ONS 15540 Channel	4 Channel Add/drop Mux/demux module	8 Channel Add/drop Mux/demux module	16 Channel Mux/demux module
1-4	Band A	Band AB	Band AD
5-8	Band B		
9-12	Band C	Band CD	
13-16	Band D		
17-20	Band E	Band EF	Band EH
21-24	Band F		
25-28	Band G	Band GH	
29-32	Band H		

We show a picture of a transponder module in Figure 6-2.

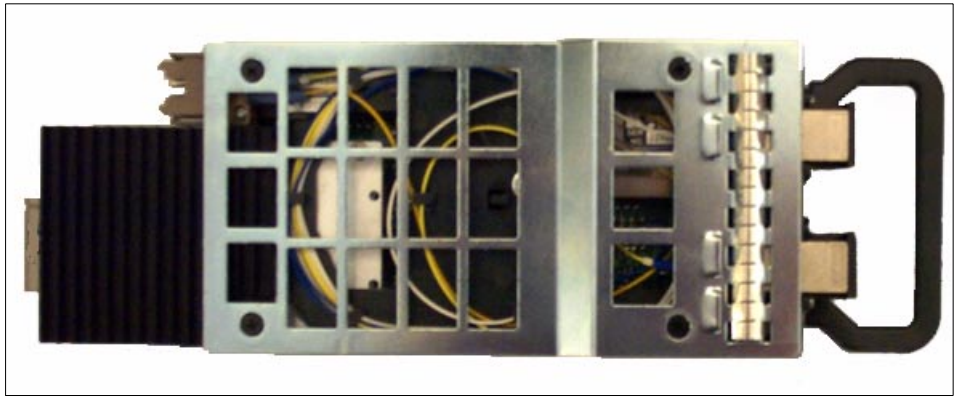


Figure 6-2 Transponder module

In Figure 6-3 we show a full motherboard and an empty motherboard.

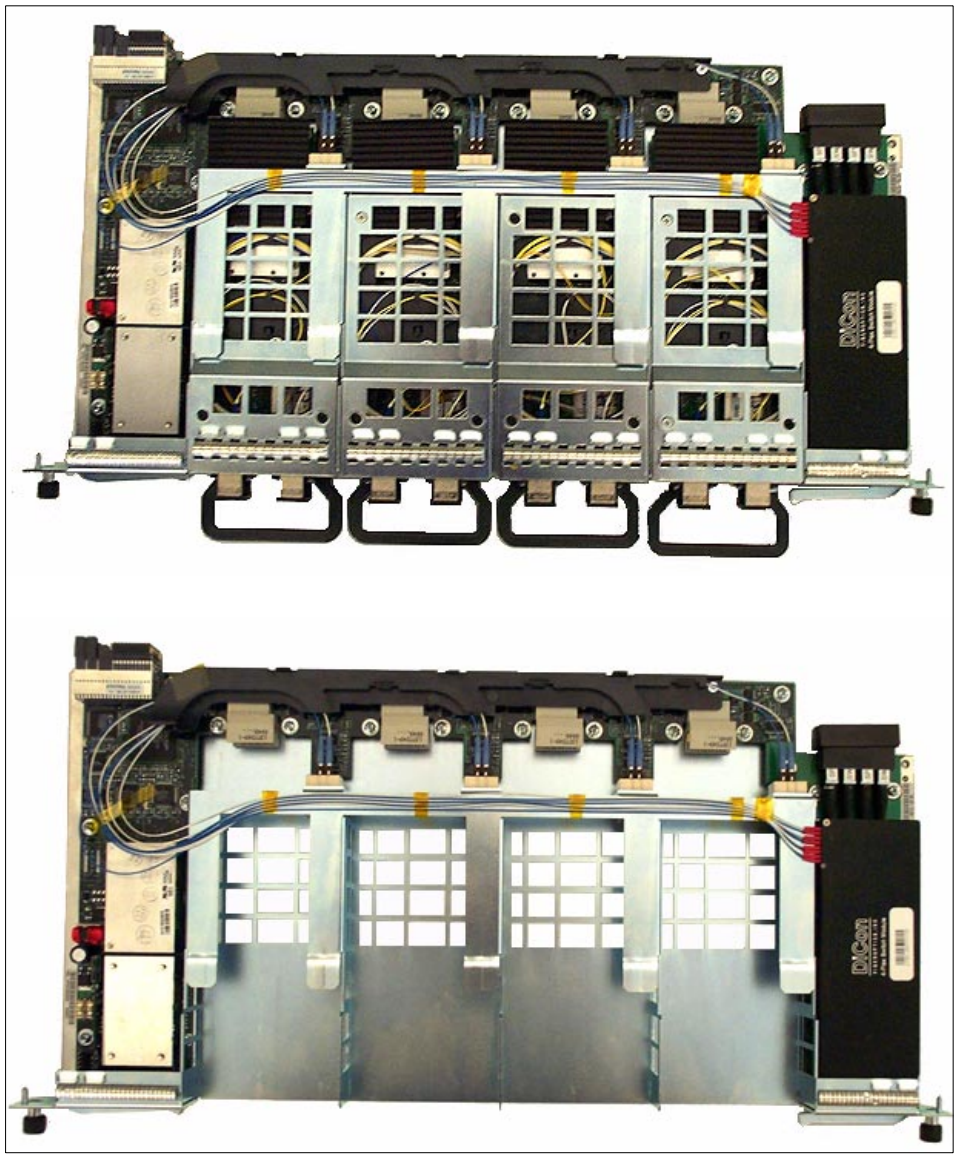


Figure 6-3 Full motherboard (top), empty motherboard (bottom)

Note: A 16 channel mux/demux module occupies two subslots in a mux/demux slot.

6.1.2 Client-side interfaces

The client interface on the transponder module is protocol transparent and bit-rate transparent and accepts a client signal on the 1310 nm wavelength through SC connectors. Both multimode (MM) and single-mode (SM) fiber are supported for client connections.

6.1.3 Optical backplane

We have highlighted the optical backplane ports within the Cisco ONS 15540. These are shown in red in Figure 6-4.

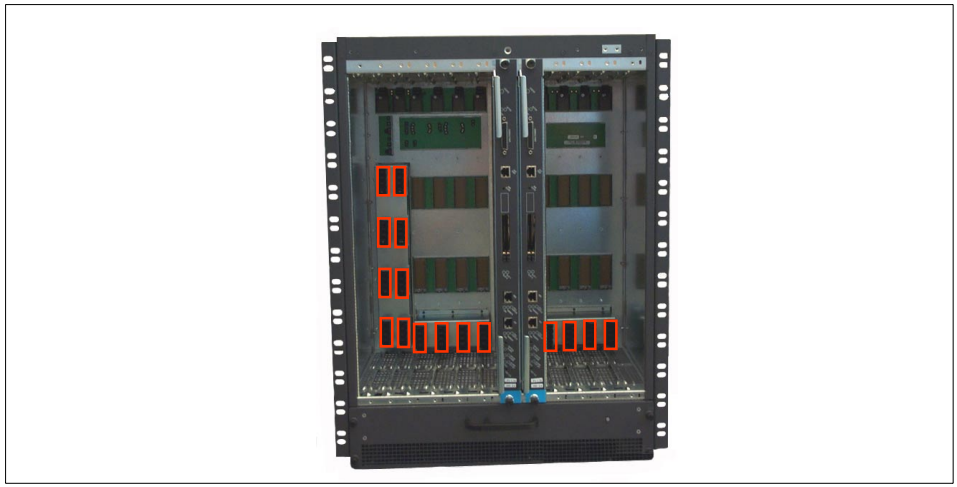


Figure 6-4 Cisco ONS 15540 optical backplane

The general mapping between the internal interfaces on the line card motherboards and on the optical mux/demux modules is shown in Table 6-2.

Table 6-2 Interface mapping

Transponder module in slot/subslot	Connect to optical mux/demux module in slot/subslot
2/0 through 2/3	0/0 and 1/0
3/0 through 3/3	0/0 and 1/0
4/0 through4/3	0/1 and 1/1

Transponder module in slot/subslot	Connect to optical mux/demux module in slot/subslot
5/0 through 5/3	0/1 and 1/1
8/0 through 8/3	0/2 and 1/2
9/0 through 9/3	0/2 and 1/2
10/0 through 10/3	0/3 and 1/3
11/0 through 11/3	0/3 and 1/3

When the splitter line card motherboards are used, these cross connections, which are fixed and nonconfigurable, couple the signal from each transponder module to a specific position on the optical mux/demux modules in both slot 0 (west) and slots 1 (east). If west line card motherboards are used, the transponder modules are connected only to the optical mux/demux modules in slot 0; if the east line card motherboards are used, the transponder modules are connected only to the optical mux/demux modules in slot.

In Figure 6-5 we show the concept of how the OSC add DeMux concept works.

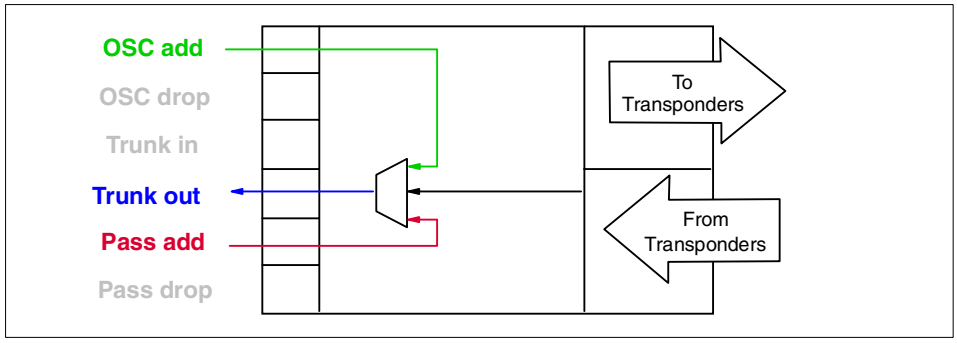


Figure 6-5 Optical DeMux OSC add

In Figure 6-6 we show the concept of how the OSC add DeMux concept works.

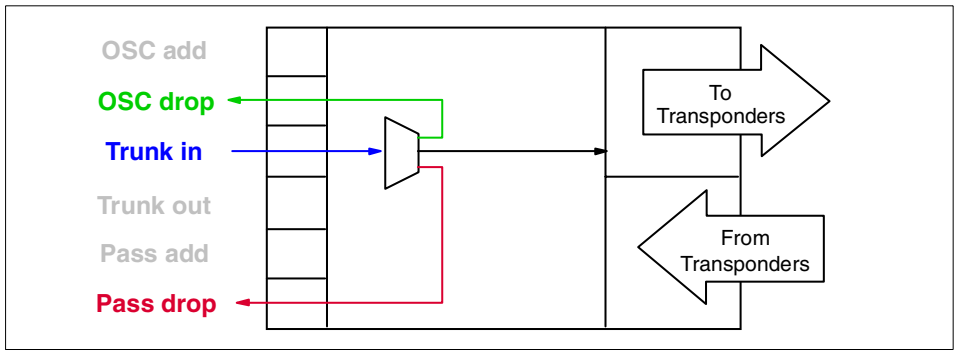


Figure 6-6 Optical DeMux module OSC drop

6.1.4 Line card motherboards

The line card motherboards hold the transponder modules and provide the optical connections from the transponder modules to the optical backplane. The line card motherboards are modular and can be populated according to user needs. A single system can hold up to eight line card motherboards, each of which accepts four transponder modules.

There are three types of line card motherboards:

- Splitter
- East
- West

The splitter motherboard supports protection against fiber failure by delivering the ITU wavelengths emitted from their associated transponders over the optical backplane to the optical mux/demux modules in both the west and east slots (slots 0 and 1, respectively).

The east and west line card motherboards deliver the ITU wavelengths from their associated transponder modules over the optical backplane to the optical mux/demux modules in either the east or west slot.

6.1.5 Hot-swappable

The Cisco ONS 15540 backplane has no active components and uses a cable of single mode fibers. The power connectors on the modules connect to the backplane allowing modules to draw up to 100W of power. The backplane used on the chassis provides the optical connections between the line card motherboards and their attached transponder modules on the client side and the mux/demux motherboards and modules on the network side.

The alarm signals from the processor card are sent to the alarm card attached to the bottom of the backplane.

6.1.6 Mux/demux motherboards

The mux/demux motherboards hold the optical mux/demux modules. Either slot 0 or slot 1 can be populated with a single mux/demux motherboard for unprotected operation, or both slots can be populated for protected operation. Each motherboard can accept up to four optical mux/demux modules, depending upon the type of module used, and can be populated according to user needs.

This mux/demux motherboard is shown in Figure 6-7.

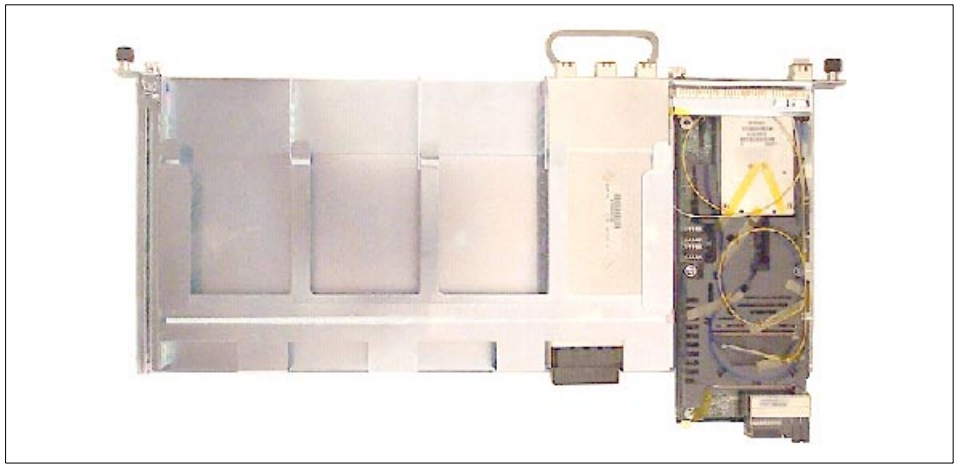


Figure 6-7 Cisco ONS 15540 mux/demux motherboard module

6.1.7 Processor cards

Slots 6 and 7 of the Cisco ONS 15540 chassis hold processor cards. The processor cards support redundancy and online insertion and removal. In a redundant system, the processor cards monitor each other using the backplane Ethernet and signals. The processor card monitors the fan assembly operation and airflow temperature. During a fan failure or an out-of-temperature range condition the processor card activates an alarm.

We show a processor card and its interfaces in Figure 6-8.

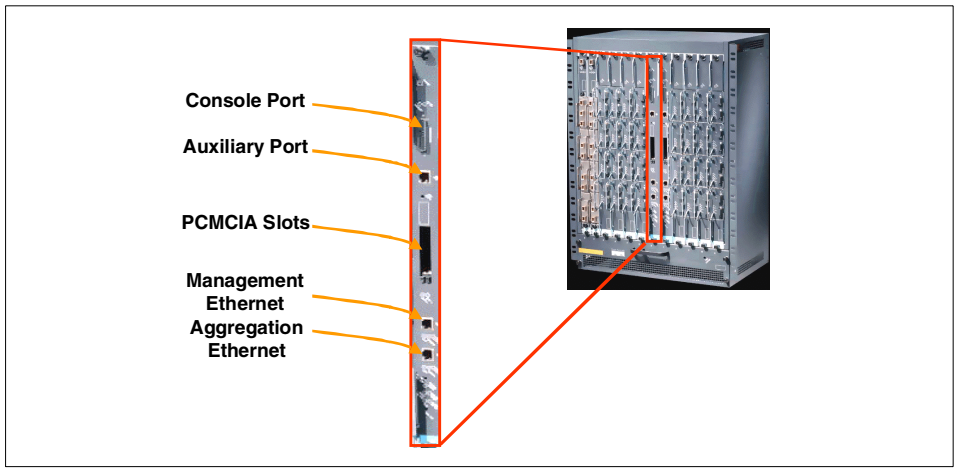


Figure 6-8 Management and administration interfaces

Some of the highlights are:

- Two redundant CPUs per chassis
- IOS
- DB-25 DCE RS-232 console port
- RJ-45 DTE RS-232 auxiliary port
- Two Fast Ethernet ports (network management and aggregation)
- Two PCMCIA slots
- Status and alarm LEDs
- 64 – 512 MB DRAM

Processor cards manage communication functions for the system. The cards monitor all modules in the chassis and determine the state of the system. Each module determines its state from feedback at various system monitoring points. The processor generates clocking to all the modules and some additional components in the system.

6.2 Management

The Cisco ONS 15540 supports Simple Network Management Protocol (SNMP) compliant network management systems. Communication, management, and provisioning between nodes is supported by an optional Optical Supervisory Channel (OSC). Connection to the 10/100 Ethernet port or serial console port on any node allows for the full management of all nodes in a ring or between nodes in point-to-point topologies.

The processor card is equipped with a console port, a Fast Ethernet interface for Telnet access and network management, and an auxiliary port. The NME (network management Ethernet) interface supports 10-Mbps or 100-Mbps UTP (unshielded twisted-pair) ports. This RJ-45 interface supports full-duplex or half-duplex connections. The NME port on the processor card is a management port that allows multiple simultaneous Telnet or SNMP network management sessions. The Ethernet port interface on the processor card does not route or bridge traffic to other Ethernet ports on the Cisco ONS 15540. This Ethernet port is a management port only and cannot be configured as a routing port. On the processor card front panel are LEDs that display the status of critical, major, and minor signals, as well as the status of alarm cutoff and history conditions. The alarm signals from the processor go to an alarm daughter board on the backplane, which has a connector for central office alarm facilities. The system processors run Cisco IOS software and support the following features:

- Automatic configuration at startup
- Automatic discovery of network neighbors
- Online self-diagnostics and tests
- Arbitration of processor status (active/standby) and switchover in case of failure without loss of connections
- Automatic synchronization of startup and running configurations
- In-service software upgrades
- Per-channel APS (Automatic Protection Switching) in linear and ring topologies using redundant subsystems that monitor link integrity and signal quality
- System configuration and management through the CLI (command-line interface) and SNMP
- Optical power monitoring on the transport side, digital monitoring on the client side, and per-channel transponder in-service and out-of-service loopback (client and transport sides)
- Optional out-of-band management of other Cisco ONS 15540 systems on the network through the OSC (optical supervisory channel).

The Cisco ONS 15540 supports an SNMP interface for remote network management. The systems can also be operated through a local command-line interface (CLI). In the future, the Cisco ONS 15540 will be supported by the Cisco Transport Manager and the CiscoWorks 2000 Enterprise Network Manager.

Cisco Transport Manager (CTM) is the carrier-class element management system (EMS) for the Cisco ONS 15000 Series product line. Cisco Transport Manager provides advanced capabilities in the functional management areas of configuration, faults, performance, and security for Cisco optical network elements, subnetworks, and networks.

Cisco Transport Manager is based on a client/server architecture that scales to support up to 1000 network elements and 100 simultaneous clients. Cisco Transport Manager is a key enabler for automation in the Internet OSS through the northbound interfaces to a network management system (NMS) or operations support system (OSS).

The CiscoWorks 2000 product line provides a set of solutions designed to assist in managing an enterprise network. Three major solutions include WAN, LAN, and service-level management. These solutions are built on a CiscoWorks 2000 Web-based management server to provide integrated management including data sharing and system process integration to improve overall system administration.

6.3 Serviceability

The systems support hot-swappable modules for in-service upgrade and repair. The systems also support hot software upgradability providing hitless software download and rebooting capability.

6.3.1 Protection

The design of the Cisco ONS 15540 provides for two levels of network protection:

- Facility protection
- Line card protection

Facility protection provides protection against failures due to fiber cuts or unacceptable signal degradation on the transport side.

Line card protection provides protection against failures both on the fiber and in the transponders, which contain the light emitting and light detecting devices as well as the 3R (regenerate, reshape and retime) electronics. Line card protection can also be implemented using redundant client signals. This provides protection against the failure of the client, the transponder, or the fiber.

In Figure 6-9 we show the component layout within the Cisco ONS 15540.

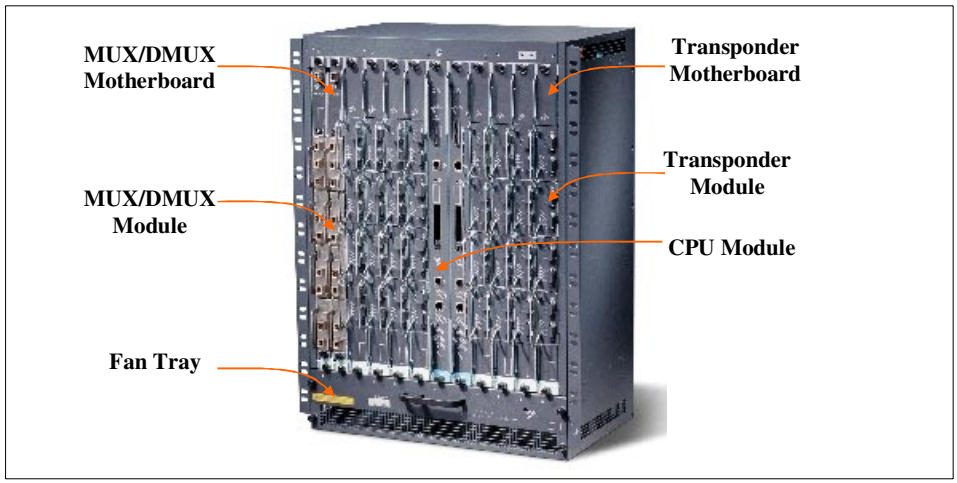


Figure 6-9 Cisco ONS 15540 components view

The Cisco ONS 15540 includes two processor cards for redundancy. Each processor is composed of a number of subsystems, including a CPU, a system clock, Ethernet switch for communicating between processors and with the LRC (line card redundancy controller) on the mux/demux and line card motherboards, and a processor redundancy controller.

The active processor controls the node, and all cards in the system make use of the system clock and synchronization signals from the active processor.

When the Cisco ONS 15540 is powered up, the two processors engage in an arbitration process to determine which will be the active and which will be the standby. All things being equal, the processor in the lower numbered slot assumes the active role. During operation, the two processors remain synchronized (application states, running and startup configurations, system images), and the two clocks are maintained in phase alignment. The operational status of each processor is monitored by the processor redundancy controller of the other processor through the backplane Ethernet.

In the event of a failure or removal of an active processor, the standby processor immediately takes over and assumes the active role. Once the problem on the faulty card has been resolved, it can be manually restored to the active function. In addition to providing protection against hardware or software failure, the redundant processor arrangement also permits installing a new Cisco IOS system image without system downtime.

For more information about processor redundancy operation, as well as other software features, refer to the *Cisco ONS 15540 ESP Configuration Guide and Command Reference*, 78-12669-01.

6.3.2 Redundancy and availability

The Cisco ONS 15540 supports path switching based under automated software control. The triggers to cause a switch are bit error rate and loss of signal. The bit error rate trigger value is provisionable.

To survive a fiber failure, fiber optic networks are designed with both working and protection fibers. In the event of a fiber cut or other facility failure, working traffic is switched to the protection fiber. The Cisco ONS 15540 supports such facility protection using a splitter scheme to send the output of the DWDM transmitter on two transport side interfaces.

Line card protection on the Cisco ONS 15540 provides protection against both facility failure and transponder failures. With line card protection, the signal from the client equipment is duplicated on two transponder interfaces, one active and one standby. The WDM signal from one of the transponder modules is sent across the optical backplane to a mux/demux module in slot 0; the signal from the other transponder module is connected to a mux/demux module in slot 1. At any given time, one of the transmitters at the client interface is turned on and is generating the required optical signal, and the second transmitter is off. Using a y-cable enables full protection on the Cisco ONS 15540 and offers protection against both facility failures and transponder card failures.

For detailed information about the hardware rules, refer to the *Cisco ONS 15540 ESP Planning and Design Guide*, 78-13102-01.

6.3.3 Splitter protection

In Figure 6-10 we show that on the trunk side, a fiber pair, with one receive fiber and one transmit fiber, connects to one mux/demux module in each slot. The client signal is transmitted through both mux/demux modules. A 2 x 2 switch on the line card motherboard receives both signals and selects one as the active signal. When a failure is detected, the 2 x 2 switch switches over to the standby signal. The standby signal then becomes the active signal.

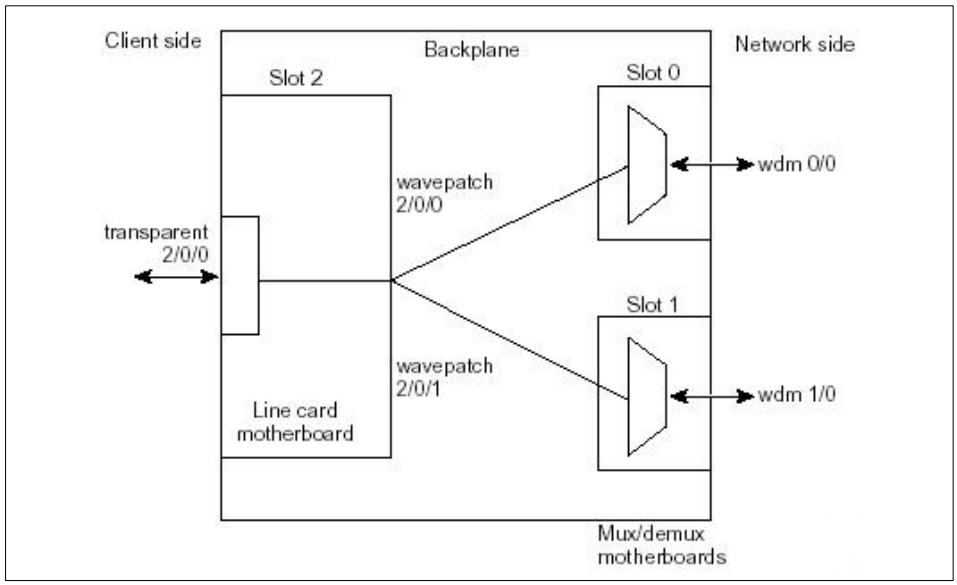


Figure 6-10 Internal splitter protection

Considerations for using splitter protection

The following considerations apply when considering the use of splitter protection:

- Transponder modules come in wavelength-specific versions. Therefore, the type of transponder module you insert in a transponder slot determines the wavelength on which the client signal is transported.
- Each subcard position in the splitter-protected line card motherboard corresponds to a specific subcard position in both of the mux/demux motherboards in slots 0 and 1.

For detailed information on backplane connectivity, refer to the *Cisco ONS 15540 ESP Planning and Design Guide*, 78-13102-01.

Splitter protection does not protect against failure in the transponder module where the lasers are located. Splitter protection also does not protect against failure of the client equipment.

To protect against transponder module failure, use y-cable protection as described in “Line card protection” on page 162. To protect against both transponder module failure and client failure, implement protection on the client equipment instead.

- A fully configured system can support 32 channels in splitter protection mode.

- Splitter protection is nonrevertive. After correcting the problem that caused the failure and verifying the signal quality, you must manually switch over to begin using the former working signal. Use optical testing equipment to verify the signal quality. For more information on troubleshooting signal failures and restoring connectivity, refer to the *Cisco ONS 15540 ESP Troubleshooting Guide*, 78-13022-01.

For detailed information on shelf configuration rules, refer to the *Cisco ONS 15540 ESP Planning and Design Guide*, 78-13102-01.

A splitter on each line card motherboard couples the transponder's DWDM interface across the optical backplane to the internal interfaces on the optical mux/demux modules in the east and west mux/demux slots.

On the transport side, one fiber pair serves as the working connection, while the other pair provides protection. The signal is transmitted on both connections, but in the receive direction, an optical switch selects one signal to be the active one. If a loss of light is detected on the working fiber, a switch to the standby signal is made under control of the LRC (line card redundancy controller). Assuming, for example, that the working signal is on the east interface, a failure of the signal on that fiber would result in a switchover, and the signal on the west interface would be selected for the receive signal.

The following considerations apply when using splitter protection:

- The splitter protected line card motherboard supports splitter protection. Because the signal splitter introduces 4.6 dB of loss in the transmit direction, Cisco recommend using the nonsplitter protected line card motherboards (east or west version) for configurations where splitter protection is not required.
- Switchover after a failure under splitter protection is nonrevertive. After a switchover, manual intervention is required to revert to using the previously failed fiber for the working traffic once the fault has been remedied.
- The OSC plays a crucial role in splitter based protection by allowing the protection fiber to be monitored for a cut or other interruption of service.
- LSC (laser safety control) is not available when splitter protection is used.

Line card protection is implemented on the Cisco ONS 15540 using a y-cable scheme. Y-cable protection protects against both facility failures and failure of the transponder module. Using an external 2:1 combiner (the y-cable), connections between the client equipment and the transponder interfaces are duplicated so that each input and output client signal is connected to two transponder interfaces.

6.3.4 Line card protection

Line card protection on the Cisco ONS 15540 provides protection against both facility failures and transponder module failures. With line card protection, the client equipment sends a duplicated signal to two separate transponder modules on the Cisco ONS 15540. The client equipment duplicates the signal itself on two separate connects or uses a y-cable. On the Cisco ONS 15540, the signal from one of the transponder modules crosses the optical backplane to a mux/demux module in slot 0; the signal from the other transponder module crosses the optical backplane to a mux/demux module in slot 1. This is shown in Figure 6-11.

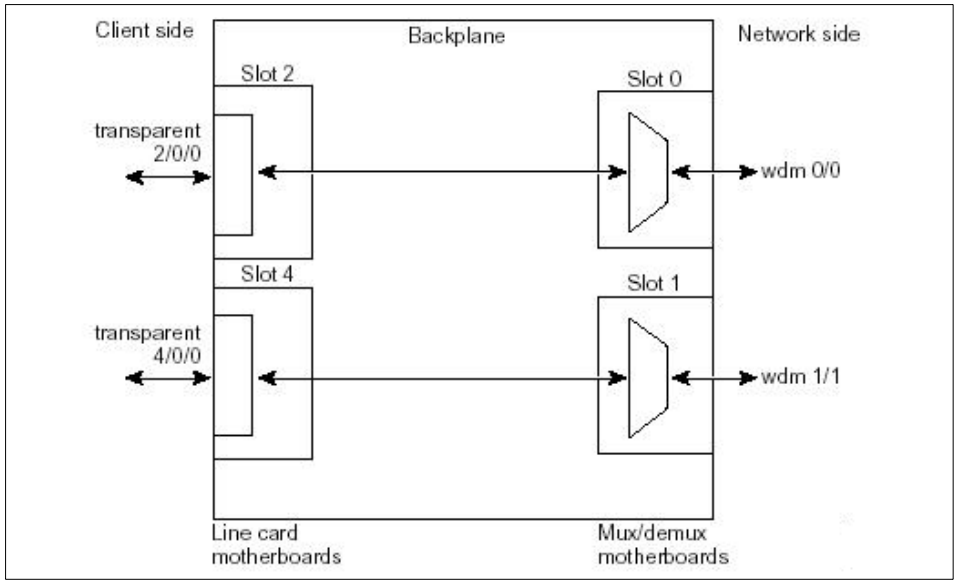


Figure 6-11 Line card protection

The Cisco ONS 15540 supports two types of line card protection:

- Client protection
- Y-cable protection

In client protection mode, both signals are transmitted to the client system. The client system decides which signal to use and when to switch over.

With y-cable protection, the signal from only one of the transparent interfaces is transmitted to the client. The Cisco ONS 15540 turns on the laser at the active transparent interface, and turns off the laser on the standby transparent interface. At each receiver on the trunk side of the transponder module, the

system monitors the optical signal power level. If the system detects a failure of the active signal when an acceptable signal exists on the standby transponder module, a switchover to the standby signal occurs by turning off the active transmitter at the client interface and turning on the standby transmitter.

Considerations for using line card protection

The following considerations apply when considering the use of line card protection:

- Transponder modules come in wavelength-specific versions. Therefore, the type of transponder module you insert in a transponder motherboard slot determines the wavelength on which the client signal is transported.
- Each subcard position in an unprotected line card motherboard corresponds to a specific subcard position in one of mux/demux motherboards. If the line card motherboard supports the “west direction”, the signal is transmitted to slot 0. If the line card motherboard supports the “east direction”, the signal is transmitted to slot 1.
- Y-cable line card protection does not protect against failures of the client equipment. To protect against client failures, ensure that protection is implemented on the client equipment itself.
- A fully provisioned single shelf configuration can support 16 channels in line card protection mode. A fully provisioned dual shelf configuration can support 32 channels in line card protection mode.
- Y-cable line card protection supports revertive behavior. With revertive behavior, the signal automatically switches back to the working path after the signal failure has been corrected. The default behavior is nonrevertive.
- Proper physical configuration of the system is critical to the operation of line card protection. The following rules apply:
  - To simplify system management, terminate the client signal on two transponder modules of the same channel type. In this way the client signal maps to the same WDM wavelength on both the working and protection channels.
  - The transponder modules connected to a given client must be in different chassis slots. For example, client equipment connecting to a transparent interface in slot 2 would also connect to a transparent interface in slot 4.

For detailed information on shelf configuration rules, refer to the *Cisco ONS 15540 ESP Planning and Design Guide*, 78-13102-01.

6.3.5 Path switching

The Cisco ONS 15540 supports per-channel unidirectional and bidirectional 1+1 path switching. When a signal is protected and the signal fails, or in some cases degrades, on the active path, the system automatically switches from the active network path to the standby network path.

Signal failures can be total loss of light caused by laser failures, or by fiber cuts between the Cisco ONS 15540 and the client equipment. Loss of light failures cause switchovers for both splitter protected and y-cable protected signals.

For y-cable protected signals, you can also configure alarm thresholds to cause a switchover when the signal error rate reaches an unacceptable level.

Figure 6-12 shows a simple point-to-point configuration with splitter protection. The configured working path carries the active signal, and the configured protection path carries the standby signal.

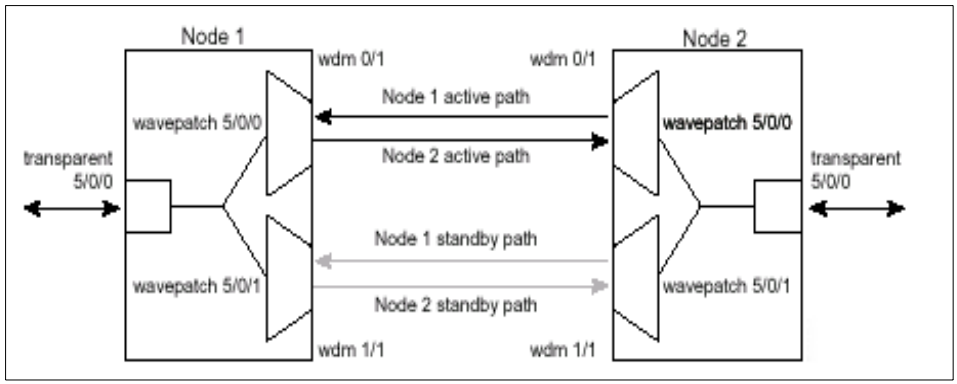


Figure 6-12 Path configuration with splitter

Figure 6-13 shows the behavior of unidirectional path switching when a loss of signal occurs.

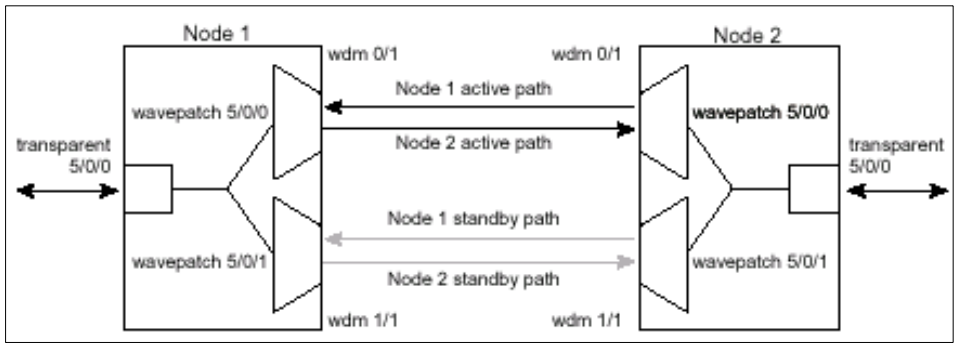


Figure 6-13 Unidirectional path switching overview

For the two node example network, unidirectional path switching operates as follows:

- Node 2 sends the channel signal in both the active and standby paths.
- Node 1 receives both signals and selects the signal on the active path.
- Node 1 detects a loss of signal light on its active path and switches over to the standby path.
- Node 2 does not switch over and continues to use its original active path.

In Figure 6-14 we show the behavior of bidirectional path switching when a loss of signal occurs.

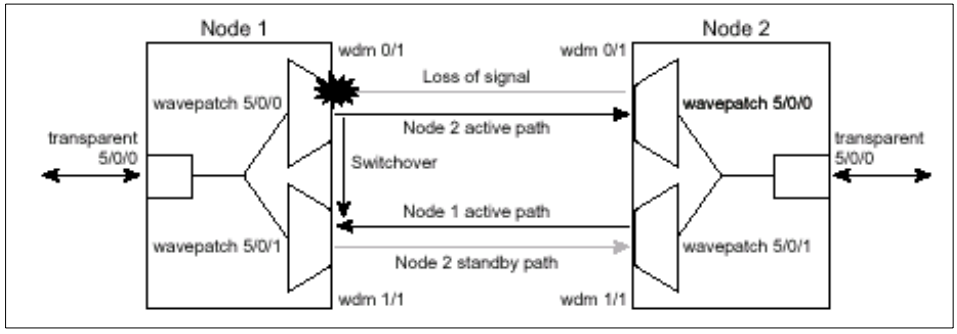


Figure 6-14 Bidirectional path switching overview

For the two node example network, bidirectional path switching operates as follows:

- Node 2 sends the channel signal in both the active and standby paths.
- Node 1 receives both signals and selects the signal on the active path.

- Node 1 detects a loss of signal light on its active path and switches over to the standby path.
- Node 2 does not switch over and continues to use its original active path.

Note: Both nodes in the network that add and drop the channel must have the same APS configuration.

Specifically, both must have the same path switching behavior, and working and protection paths.

6.4 Footprint

The Cisco ONS 15540 uses a 12-slot modular vertical chassis. The system receives power through redundant –48 VDC inputs. A redundant external AC power supply is available, or DC power can be provided directly. As you face the chassis, the two leftmost slots (slots 0–1) hold the mux/demux motherboards. These slots, which are populated with optical mux/demux modules, correspond to the west and east directions, respectively.

Slots 2–5 and 8–11 hold the line card motherboards, which are populated with transponder modules. Slots 6–7 hold the processor cards. Air inlet, fan tray, and cable management are located beneath the modular slots. The system has an optical backplane for carrying signals between the transponders and the optical mux/demux modules and an electrical backplane for system control.

The chassis is NEBS Level 3 compliant and is shown in Figure 6-15.

.

Figure 6-15 Cisco ONS 15440 chassis

6.5 Connectivity

The Cisco ONS 15540 connects to client equipment on one side and to the DWDM transport network on the other side. Simply described, the Cisco ONS 15540 takes a client signal and converts it to an ITU G.692 compliant wavelength, and then optically multiplexes it with the other client signals for transmission over a fiber link.

The Cisco ONS 15540 supports 1+1 path protection using a scheme based on the Automatic Protection Switching (APS) standard. In a single-shelf configuration, the Cisco ONS 15540 can support up to 32 channels with facility (fiber) protection or 16 channels with line card protection. The Cisco ONS 15540 can be deployed in point-to-point, hubbed ring, and meshed ring topologies.

In Figure 6-16 we illustrate the principal functions involved in transmission of the signal between the client and transport networks within the Cisco ONS 15540.

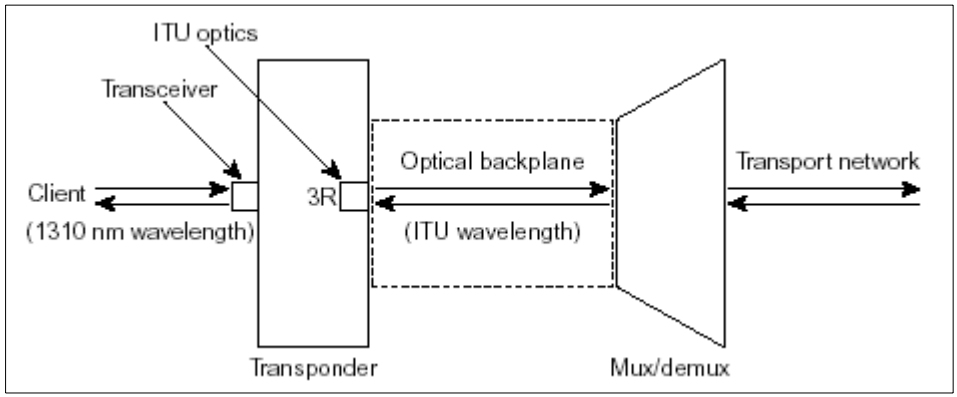


Figure 6-16 Transmission of signal

In the transmit direction, these functions include receiving the client signal by a transceiver, converting the client signal in the transponder, transmitting the signal over the optical backplane, and multiplexing the signal with other client signals before putting it on the fiber. The opposite functions are performed in the receive direction.

The client signal is received through a transceiver attached to the transponder module's external port. Inside the transponder module the 1310 nm input optical signal is converted to an electrical signal and the 3R (reshape, retime, retransmit) function is performed.

A modulated laser diode then converts the electrical signal back to an optical one with a specific wavelength that complies with the ITU laser grid. The optical signal leaves the transponder and travels across the optical backplane, which is an optical fiber array circuit comprised of fiber ribbon cables. This backplane serves as a fixed optical cross connection between the transponder modules and the optical mux/demux modules.

Inside the optical mux/demux module the input signals are multiplexed into a single DWDM signal and launched into the fiber on the DWDM network side. Therefore, there is a one-to-one relationship between each client signal and each wavelength on the transport side.

The Cisco ONS 15540 is a duplex system; therefore where there are light emitters there are also light detectors. For example, the client side interfaces on the transponders both transmit and receive light. The same is true of the transponder's DWDM interface. Also, the optical mux/demux modules both multiplex the transmit signal and demultiplex the receive signal.

6.5.1 OADM

The Cisco ONS 15540 provides both service termination and optical add/drop multiplexing (OADM), 4 or 8 channel OADMs, 16 or 32 channel terminal MUX/DMUX.

This can be ordered with or without OSC connections. An example of one is shown in Figure 6-17.

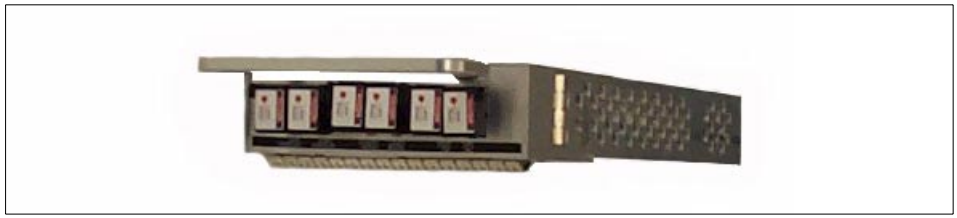


Figure 6-17 Mux/demux add/drop

6.5.2 Specification

The 2.5-Gbps transparent line interface can interface to synchronous data streams from 16 Mbps up to 2.5 Gbps. This module can carry SONET/SDH, OC-3/STM-1, OC-12/STM-4, or OC-48/STM-16 traffic. It can also be used to carry Gigabit Ethernet, Fast Ethernet, ESCON, FICON, and Fibre Channel.

It also features:

- Transport side mux/demux ports
- OSC (optical supervisory channel) 1562.23 nm
- Wavelengths 32, ITU G.692
- Connector type MU
- System performance Bit error ratio 10 -15 for ESCON,10 -12 for all other applications

6.5.3 Bandwidth

The DWDM optical multiplexer/demultiplexer module provides a 32 wavelength inter-nodal optical connection. These modules interconnect the channel transmit/receive modules to the individual wavelengths. Redundant modules provide interconnection to eastbound and westbound fibers. Cisco ONS 15540 modules are available to add/drop 4, and 8 wavelengths. Terminal mux/demux modules are available in 16 lambda growable to 32 lambda. The number of wavelengths supported is 32 per fibre pair, these have a channel spacing of 100Ghz. These wavelength are all C band.

6.6 Protocols

The Cisco ONS 15540 is protocol independent. The transponder interface monitors the incoming bit streams and reports the following digital performance parameters:

- 8B10B coding violations
- Out of bit synch
- Sequence error (ESCON)
- Fibre Channel (FC)
- SONET/SDH bit interleaved parity (BIP)
- Section monitoring
- Out of frame (OOF)
- Severely errored frame (SEF)

The transponder interface also monitors the received optical power level.

6.6.1 Client side

Encapsulation of client signals is supported on the transponder interfaces in either 3R enhanced mode, which allows some client protocol monitoring (such as code violations and data errors) or 3R mode, where the transponder is transparent to the client data stream.

In either case, the format of the contents of the client data stream remains unmodified. Configurable failure and degrade thresholds for monitored protocols are also supported.

The following encapsulation types are supported in 3R mode with protocol monitoring:

- Gigabit Ethernet
- SONET (OC-3, OC-12, OC-48)
- SDH (STM-1, STM-4, STM-16)

- Fibre Channel (1 Gbps)
- ESCON
- FICON

The following additional encapsulation types are supported in regular 3R mode:

- Fast Ethernet
- FDDI
- Fibre Channel (2 Gbps)

These services are transparently transported across Cisco ONS 15540 network. They can be added or dropped at the wavelength level as they traverse across the network.

An additional set of discrete rates between 16 Mbps and 2.5 Gbps is also supported in regular 3R mode.

Open fiber control (OFC) is supported for Fibre Channel. Alternatively, forward laser shutdown (FLS) can be enabled to shut down the client side laser if a trunk fiber cut is detected.

For detailed information about client interface configuration, refer to the *Cisco ONS 15540 ESP Configuration Guide and Command Reference*, 78-12669-01.

6.6.2 Transport (dark fiber) side

In the transponder module, the client signal is regenerated and retransmitted on an ITU-compliant wavelength across the optical backplane. The laser on each transponder module is capable of generating one of two wavelengths on the transport side.

Therefore, there are 16 different transponder modules (for channels 1–2, 3–4,..., 31–32) to support the 32 channels. The software determines which wavelength each module should generate based on whether it is inserted in the upper (subslot 0 or 2) or lower (subslot 1 or 2) of a line card motherboard. A safety protocol, laser safety control (LSC), is provided to shut the transmit laser down on the transport side when a fiber break or removed connector is detected. The transponder modules are hot pluggable, permitting in-service upgrades and replacement.

Remote links use single mode (SM) fiber as defined in the ITU-T G.652 standard. This fiber is commonly referred to as SMF-28.

Certain brands of nonzero dispersion (NZDSF) shifted fiber will also be qualified for remote link inter connectivity including Lucent Tru-Wave NZ-DSF and Corning "All-Wave" NZ-DSF.

Multimode fiber does not operate between Cisco ONS 15540 systems.

6.6.3 Optical power budget and attenuation

InTable 6-3 we show the client power budget and attenuation requirements for the IBM storage protocols and IBM's implementation of other common protocols.

These data pertain to the Cisco ONS 15540 transponder interfaces, which have a receive sensitivity of –23 dBm to –1.5 dBm and a transmit power of –2 dBm. It is important to verify that the client equipment specification fall within these ranges; otherwise, attenuation may be required.

Table 6-3 Optical power budget

Protocol	Transmit (dBm)	Receive(dBm)	ONS 15540 minimum attenuation @ 0 km
ESCON (MM)	-15 to -20.5	-14 to -29	Rx: 2.5 to 8 dB/none Tx: 27 dB / -12 dB
ESCON (SM)	-3 to -8	-3 to -28	Rx: 15 to 20 dB/none Tx: 26 dB / -1dB
FICON (SM/LX)	-4 to -8.5	-3 to -22	Rx: 12.5 to 19 dB/none Tx: 20 dB / -1 dB
ATM 155 (MM)	-14 to -19	-14 to -30	Rx: 4 to 9 dB/none Tx: 12 dB to 28 dB/-12dB
ATM 155 (SM)	-8 to -15	-8 to -32.5	Rx: 15 to 8 dB/none Tx: 30.5 dB / -6 dB
FDDI (MM)	-14 to -19	-14 to -31.8	Rx: 9 to 4 dB/none Tx: 29.8 dB / -12 dB
GE (SM/LX) (MM via MCP)	-14 to -20	-17 to -31	Rx: 9 to 3 dB/none Tx: 29.8 dB / -15 dB
GE (MM/SX) (850nm)	-4 to -10	-17 to -31	Rx: 19 to 13 dB/none Tx: 29 dB / -15 dB

6.7 Power requirements

The system supports redundant -48 VDC power which draws 15 amps. Redundant external 110/220 VAC power supplies are also available. The external power supply is a single-phase, AC-DC, 1050W, -48V output power supply that connects to the chassis through terminal blocks. The external power supply is installed in an external power shelf that fits into a standard equipment rack. Up to

three external power supplies can be installed in the external power shelf. When the chassis is used in the telco environment, DC-input power is directly powered to the chassis through the terminal blocks. The NEBS 3-compliant Cisco ONS 15540 supports direct -40.5 to -72V DC power.

An optional external and redundant 3RU rectifier shelf supplies:

- 110V/240V AC power to the system
- Heat dissipation 3840 BTU/hr

6.8 Microcode and firmware

The processor card Flash SIMM is a 16 MB, 80-pin SIMM that contains a compressed Cisco IOS image that is loaded and executed automatically by ROMMON upon powerup.

The processor card has two Flash PC Card slots that are accessible from the front panel. Either slot can be a memory or an I/O device. The Flash PC Cards are typically used to copy system images and save standard configurations. Flash PC Cards are a type of Flash memory that provide expanded file storage for your system. Flash PC Cards, unlike the onboard Flash SIMM (bootflash), are not required for the operation of the system.

6.9 Standards compliance

The Cisco ONS 15540 supports up to 32 wavelengths. All wavelengths are as specified in the ITU C Band specifications.

The Cisco ONS 15540 complies with NEBS III requirements for use in service provider central office applications.

6.10 Clocking or bit racing

All transponders in the Cisco ONS 15540 support clock recovery and each can support the following clock speeds:

- 16Mbit
- 100Mbit
- OC3/STM1
- OC12/STM4
- OC48/STM16
- 1.062Gbit

- 2.124GBit
- 1.25Gbit

There is no option to synchronize an entire chassis to a network clock at this time, but a network clock module will be available in later phases.

6.11 Supported topologies

The Cisco ONS 15540 supports the following topologies:

- Point-to-point
- Optical ring
- Logical mesh

In the simplest configuration, two systems can be connected in a point-to-point configuration. For higher line density requirements, multiple systems can be coupled together supporting up to 32 wavelengths per fiber pair. The point-to-point configuration can be extended into an add/drop, ring, or mesh ring topology with multiple systems interconnected as we describe in the following topics.

6.11.1 Point-to-point topologies

In a point-to-point topology, two Cisco ONS 15540 systems are connected to each other in the network.

The client equipment connects to one of the systems. Each of the systems originates and terminates all configured wavelengths. You can use splitter protection to protect against fiber failure, or line card protection to protect both the fiber and transponders. To also protect against client failure, you can implement protection on the client equipment itself.

Up to 32 client signals in splitter protection mode, or 16 client signals in line card protection mode, are optically multiplexed at each end and are multiplexed onto a single fiber pair for transport over a maximum distance of 100 km. This distance will be dependent on link budget and devices attached.

In Figure 6-18 we show an example of this topology using two DWDM fiber links, one working and one for protection.

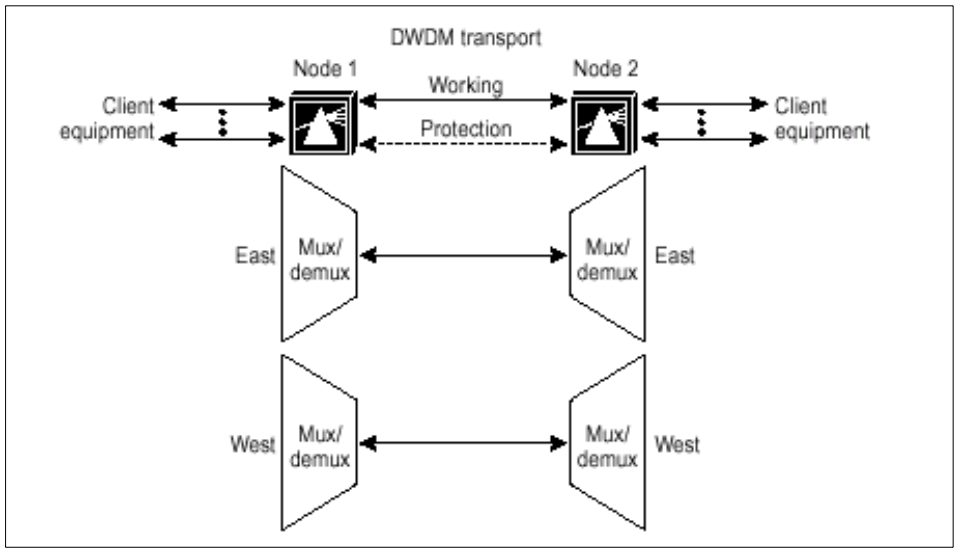


Figure 6-18 Protected point-to-point topology example

In Figure 6-19 we show an example of an unprotected point-to-point topology using one unprotected DWDM fiber link.

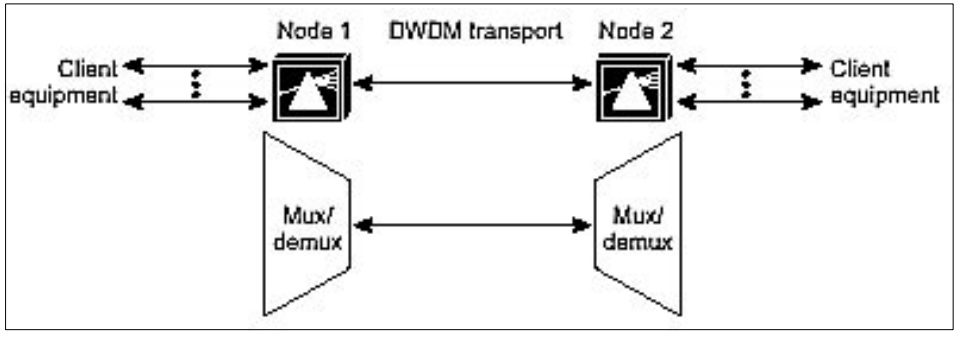


Figure 6-19 Unprotected point-to-point topology example

Point-to-point topologies have many common applications, including extending the reach of Gigabit Ethernet or SONET in long-haul transport.

Note: Point-to-point topologies support an intermediate add/drop node only in an unprotected configuration. If add/drop with protection is required, a ring configuration should be used.

The following criteria should be used in determining the equipment needed for a point-to-point topology:

- Number of channels at deployment and in the future
- Distance between nodes
- Potential topology changes in the future (such as migration to ring)
- Presence of OSC

There are many optical mux/demux module combinations that can satisfy the requirements of a network design. For example, a shelf can support 32 channels using eight 4-channel mux/demux modules, four 8-channel mux/demux modules, or two 16-channel mux/demux modules. However, certain configurations can prove costly as network requirements change.

The 16-channel mux/demux modules are ideally suited for a point-to-point topology. They impose less optical link loss than cascading the 4-channel and 8-channel modules, thereby maximizing the distance between nodes. Price per channel is also less if the current or future channel requirement is near 16 or 32. However, if future plans include migrating to a ring environment, the 16-channel mux/demux module is not ideal.

If, for example, a point-to-point topology using 16-channel mux/demux modules at each end were migrated to a hubbed ring, the node that became an add/drop node could not use the 16-channel module (though the hub node could use that module). If the migration were to a meshed ring, neither node could use the 16-channel module.

For more information on point-to-point topologies, refer to the *Cisco ONS 15540 ESP Planning and Design Guide*, 78-13102-01.

6.11.2 Ring topologies

A ring topology is a network of three or more nodes each of which connects to two other nodes to form a ring. Protection options are the same as in a point-to-point configuration, and the two fiber pairs are often geographically diverse so that a break in one fiber does not necessarily also result in a break of the other fiber. On a ring, traffic is transmitted in both directions from each node; traffic is received from only one direction. In case of a fiber failure, the node switches over to receive traffic from the other direction.

Meshed and hubbed ring topologies.

The Cisco ONS 15540 supports hubbed ring and meshed ring topologies. The following sections give a brief overview of these topologies.

Hubbed ring topologies

In a hubbed ring topology, all channels originate and terminate on the hub node (node1 in Figure 6-20). The other nodes on the ring, sometimes called satellite nodes, add and drop one or more channels. The added and dropped channels terminate at the node, while the channels that are not being dropped (express channels) are passed through optically, without being electrically terminated.

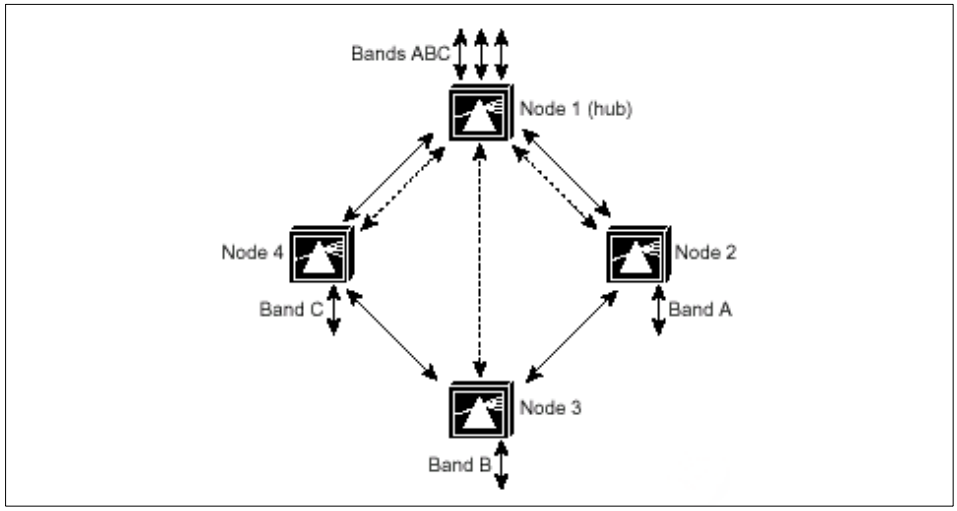


Figure 6-20 Hubbed ring topology example

Meshed ring topologies

A meshed ring is a physical ring that has the characteristics of a mesh. In Figure 6-21 we show an example of this type of configuration, which is sometimes called a logical mesh.

While all traffic travels around the physical ring, nodes 1 and node 3 share band B (channels 5-8), and nodes 3 and node 4 share band D (channels 29-32). Therefore, there is a logical mesh overlay on the ring.

Protection options and optical link loss budget considerations are the same as in a hubbed ring configuration.

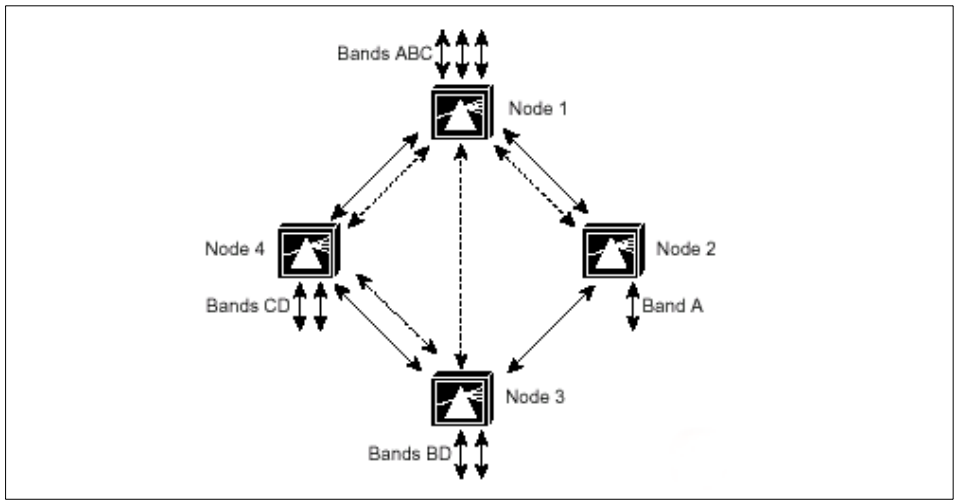


Figure 6-21 Meshed ring topology example

The following network rules apply to ring topologies:

- When using splitter protection, a channel must be present on only two nodes in the ring.
- All channels added by a node on an east add/drop mux/demux modules must be dropped on a west add/drop mux/demux module of one or more other nodes on the ring. All channels added by a node on a west add/drop mux/demux module must be dropped by an east add/drop mux/demux module of one or more other nodes on the ring. This rule may be violated during migration.
- A node cannot add a channel that is already present in the same direction until it has dropped that channel.

In addition, Cisco recommends that if there are plans to migrate from a hubbed ring to a logical mesh, the 4-channel or 8-channel add/drop mux/demux modules should be considered for deployment at the terminal node. This strategy avoids the necessity of discarding a terminal mux/demux module when migrating.

Hubbed ring with splitter protection and OSC

In Figure 6-22 we show an example topology of a three-node hubbed ring.

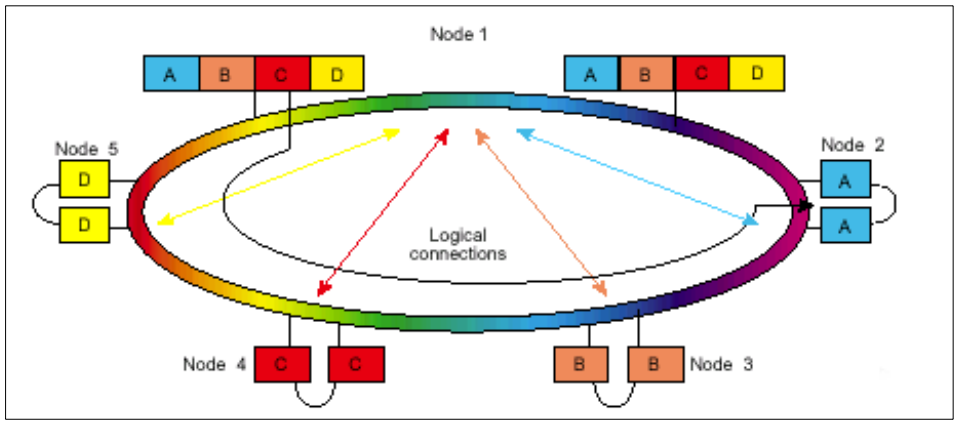


Figure 6-22 Hubbed ring channel plan

Node 1 is the hub configured with band A through band D (channels 1-16). Node 2 adds and drops band A (channels 1-4), node 3 adds and drops band B (channels 5-8), node 4 adds and drops band C (channels 9-12), and node 5 adds and drops band D (channels 13-16). The transponders carry Gigabit Ethernet traffic.

Meshed ring with splitter protection and OSC

In Figure 6-23 we show an example topology of a four-node meshed ring with splitter protection and OSC support.

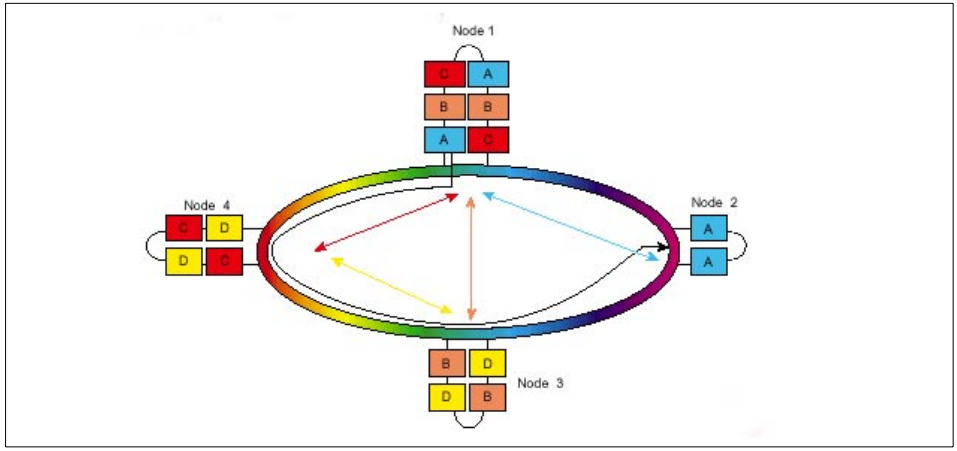


Figure 6-23 Channel plan for meshed ring node

Node 1 supports bands A, B, and C (channels 1 through 12). Node 2 adds and drops band A, node 3 adds and drops band B and band D, and node 4 adds and drops band C and band D. The transponders carry Gigabit Ethernet traffic.

Splitter protected meshed ring unprotected

In Figure 6-24 we show an example topology of a four-node meshed ring with splitter protection and OSC support.

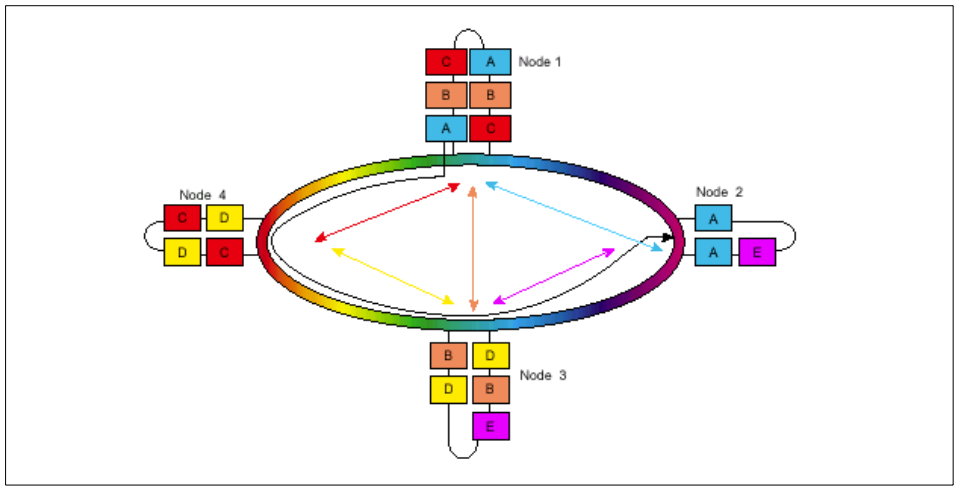


Figure 6-24 Meshed ring topology with splitter protection

Node 1 supports bands A, B, and C (channels 1-12). Node 2 adds and drops band A (channels 1-4) and unprotected band E (channels 17-20). Node 3 adds and drops band B (channels 5-8), band D (channels 13-16), and unprotected band E (channels 17-20). Node 4 adds and drops band C (channels 9-12) and band D (channels 13-16). The transponders carry Gigabit Ethernet traffic.

6.12 Resilience

The Cisco ONS 15540 can support 32 protected wavelengths. There are three high availability options. These include a splitter-based protection High Availability option, which provides economical fiber break protection.

The second High Availability option is automated dual line card protection, which offers enhanced capabilities by protecting against client and ITU port failures. The third High Availability option includes "Y" cable interconnection to dual client line cards on routers, switching, and other client equipment.

6.12.1 Y-cables

Line card protection on the Cisco ONS 15540 provides protection against both facility failure and transponder failures. With line card protection, the signal from the client equipment is duplicated on two transponder interfaces, one active and one standby. The WDM signal from one of the transponder modules is sent across the optical backplane to a mux/demux module in slot 0; the signal from the other transponder module is connected to a mux/demux module in slot 1. At any given time, one of the transmitters at the client interface is turned on and is generating the required optical signal, and the second transmitter is off. Using a y-cable enables full protection on the Cisco ONS 15540 and offers protection against both facility failures and transponder card failures. UserSide Tx-Enable Green Client port transmit laser is enabled.

6.12.2 Line card protected 32-channel dual shelf configuration

By cascading two Cisco ONS 15540 shelves, 32 channels can be supported in a line card protected point-to-point configuration. Shelf 1 is configured for channels 1–16 with OSC, while shelf 2 is configured for channels 17–32 without OSC. The terminal mux/demux modules are patched between the two shelves as if they were in the same shelf. In this configuration, shelf 2 cannot support the OSC, which means that a separate Ethernet connection to that shelf is required for management purposes.

6.12.3 Mux/demux

The optical mux/demux modules are passive devices that optically multiplex and demultiplex a specific band of ITU wavelengths. A mux/demux module is shown in Figure 6-25.

Figure 6-25 Mux/demux module

In the transmit direction, the optical mux/demux modules multiplex signals transmitted by the transponder modules over the optical backplane and provide the interfaces to connect the multiplexed signal to the transport (DWDM) network. In the receive direction, the optical mux/demux modules demultiplex the signals from the transport network side before passing them over the optical backplane to the transponders.

The optical mux/demux motherboards occupy slots 0 and 1 of the Cisco ONS 15540 chassis. The chassis uses one optical mux/demux motherboard for unprotected operation or two per system for protected operation. Each mux/demux motherboard can accept up to four mux/demux modules. The modular mux/demux motherboards are available with or without OSC (optical supervisory channel) and can be populated according to user needs. There are three types of mux/demux modules available:

- 4-channel
- 8-channel
- 16-channel

Up to four 4-channel or 8-channel optical add/drop mux/demux modules or two 16-channel optical mux/demux modules can be installed in a mux/demux motherboard. Each module can multiplex and demultiplex a band of 4 or 8 channels, for a maximum of 32 channels. Channels not filtered are passed on to the next mux/demux module. The add/drop mux/demux modules interface to the network transport side and to the transponder modules over the optical backplane.

Two optical terminal mux/demux modules can be installed in a mux/demux motherboard. Each terminal mux/demux module can multiplex and demultiplex a band of 16 channels, for a maximum of 32 channels. All of the channels received by the module are terminated, none are passed through. The terminal mux/demux modules interface to the network transport side and to the transponder modules over the backplane.

6.12.4 Channel spacing and band allocation

The Cisco ONS 15540 band and channel allocation adheres to the ITU-T grid. In Table 6-4 we detail the allocations of bands and channels and their separation.

Table 6-4 Channel spacing and bands

Band	Channel	ITU-T channel	Wavelength	Frequency
OSC		19	1562.23	191.900
A	1	21	1560.61	192.100
	2	22	1559.79	192.200
	3	23	1558.98	192.300
	4	24	1558.17	192.400

Band	Channel	ITU-T channel	Wavelength	Frequency
B	5	26	1556.55	192.600
	6	27	1555.75	192.700
	7	28	1554.94	192.800
	8	29	1554.13	192.900
C	9	31	1552.52	193.100
	10	32	1551.72	193.200
	11	33	1550.92	193.300
	12	34	1550.12	193.400
D	13	36	1548.51	193.600
	14	37	1547.72	193.700
	15	38	1546.92	193.800
	16	39	1546.12	193.900
E	17	41	1544.53	194.100
	18	42	1543.73	194.200
	19	43	1542.94	194.300
	20	44	1542.14	194.400
F	21	46	1540.56	194.600
	22	47	1539.77	194.700
	23	48	1538.98	194.800
	24	49	1538.19	194.900
G	25	51	1536.61	195.100
	26	52	1535.82	195.200
	27	53	1535.04	195.300
	28	54	1534.25	195.400

Band	Channel	ITU-T channel	Wavelength	Frequency
H	29	56	1532.68	195.600
	30	57	1531.90	195.700
	31	58	1531.12	195.800
	32	59	1530.33	195.900

6.13 Distance

The Cisco ONS 15540 extended services platform supports a wide variety of topologies and architectures including point-to-point, ring, hubbed, and mesh. Fiber distance is very flexible and is dependent on many variables including fiber quality, and installation variables such as splice quality, connectors, patch panels, and so forth.

The budget also depends on how reliable the fiber plant is, the network topology, the number of nodes in the network and the multiplexer and demultiplexer types. High-quality fiber plant has a loss of approximately 0.25 dB/km. Average fiber plant loss is approximately 0.33 dB/km and bad fiber plant loss can be as high as 0.5 dB/km. Additionally, a loss of approximately 0.3 dB to 0.5 dB per connector occurs when going through patch panels, depending on connector type and quality. A design tool for calculation of fiber optic budgets and network topologies is available on Cisco.com. A maximum unamplified distance of 80 km is achievable on the platform.

The maximium distance between nodes is 92 km (based upon 0.25 dB/km) with a link budget of 23 dB.

On the transport side, the Cisco ONS 15540 has an output (laser) power in the range of 6 to 8 dBm and a receive (detector) sensitivity of -32 dBm. Cisco ONS 15540 systems can be deployed in a point-to-point link over a maximum distance of 100 km.

6.13.1 Loss/link/light budget

An optical signal degrades as it propagates through a network. Network elements, such as optical mux/demux modules, fiber, fiber connectors, splitters, and switches, introduce attenuation.

Ultimately, the maximum allowable distance between the nodes of a network is based upon the optical link budget that remains after subtracting the power losses experienced by the channels with the worst path as they traverse the components at each node.

In Table 6-5 we show the laser transmit power and receiver sensitivity range.

Table 6-5 Laser transmit power and receiver sensitivity range

Channel type	Transmit power (dBm)	Receiver sensitivity (dBm)
Data channels	6	-32 to -8
OSC	6	-24 to 0

The goal in calculating optical link loss is to ensure that the total loss does not exceed the overall optical link (or span) budget.

For the Cisco ONS 15540, this is 38 dB for data channels. This value is arrived at by subtracting the lower limit of the receiver sensitivity (–32 dBm) from the laser launch power (6 dBm) on the transponders. The OSC has an optical link budget of 30 dB, which is equal to the OSC receiver sensitivity (–24 dBm) subtracted from the OSC laser launch power (6 dBm) on the mux/demux motherboard.

Typically, in point-to-point topologies, the OSC optical power budget is the distance limiting factor, while in ring topologies, the data channel optical power budget is the distance limiting factor.

The following general rules apply to the optical link loss budget for data channels:

- The power loss between the laser and receiver must not exceed 38 (6 – (–32)) dB or the signal will not be detected accurately.
- There must be at least 14 (6 – (–8)) dB of attenuation between neighboring nodes to avoid saturating the receiver.

The following general rules apply to the optical link loss budget for the OSC:

- The power loss between the laser and receiver must not exceed 30 (6 – (–24)) dB or the signal will not be detected accurately.
- There must be least 6 (6 – (0)) dB of attenuation between neighboring nodes to avoid saturating the receiver.

To validate a network design, the optical power loss must be calculated for each band of channels. This calculation must be done for both directions if protection is implemented, and for the OSC between each pair of nodes. The optical power loss is calculated by summing the losses introduced by each component in the signal path. At a minimum, any data channel path calculation must include line card motherboard transmit loss, channel add loss (transponder to OUT), fiber

loss, channel drop loss (IN to transponder), and line card motherboard receive loss. In ring topologies, pass-through add losses (UA to OUT) and pass-through drop losses (IN to UD) must be considered. Losses due to external devices such as patch panels also need to be included.

For client equipment to interoperate with the Cisco ONS 15540, the transmit and receive specifications of the client equipment interfaces must fall within the range of the transponder interfaces on the Cisco ONS 15540.

The transponder client interfaces (multimode and single-mode) have the following optical power characteristics:

- Receiver sensitivity: –23 dBm to –1.5dBm
- Transmitter launch power: –2 dBm

If the specifications of the client equipment interfaces do not fall within these ranges, attenuators may be required.

In the transmit direction, the splitter protected line card motherboard attenuates the ITU signal emitted from its associated transponders significantly more than does the east or west motherboard. In the transmit direction, the splitter protected line card motherboard attenuates the signal destined for its associated transponder significantly more than does the east or west motherboard.

Optical mux/demux modules attenuate the signals as they are multiplexed, demultiplexed, and passed through. The amount of attenuation depends upon the type of optical mux/demux module and the path the optical signal takes through the modules.

6.13.2 Latency

Internal switching time is < 50 ms and this will make up part of the total latency.

6.13.3 33rd lambda (wavelength) optical supervisory channel

Cisco ONS 15540 offers Optical Supervisory Channel (OSC) support for each trunk fiber. These OSCs give the ability to ensure fast (less than 50 ms) protection switching, an embedded communications link between each node in the optical topology at the fiber and lambda (λ) levels, and sophisticated operation, administration, and management (OAM) capabilities for per-service performance monitoring.

The Cisco ONS 15540 dedicates a separate channel (channel 0 or 1562.23 nm) for the OSC, which is used for network control and management information between Cisco ONS 15540 systems on the network. The OSC is carried on the same fiber as the data channels (channels 1 through 32), but it carries no client data traffic.

In Figure 6-26 we show the path of the OSC in a protected ring configuration.

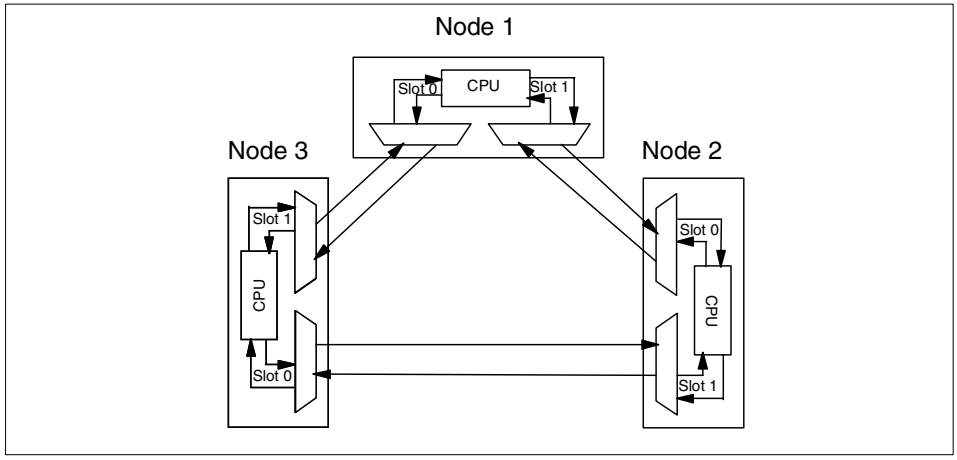


Figure 6-26 OSC signal path in a ring configuration

The OSC signal is generated by a laser on each mux/demux motherboard and is sent in both directions from the node; both receive signals are monitored to maintain communication with the neighboring nodes. The OSC signal terminates at each node.

There are two versions of the mux/demux motherboard, with and without the OSC (optical supervisory channel). Implemented with a dedicated laser and detector for a 33rd wavelength (channel 0) on the mux/demux motherboard, the OSC is a per-fiber duplex management channel for communicating between Cisco ONS 15540 systems.

The OSC allows control and management traffic to be carried without the necessity of a separate Ethernet connection to each Cisco ONS 15540 in the network. The OSC is established over a point-to-point connection and is always terminated on a neighboring node. By contrast, data channels may or may not be terminated on a given node, depending on whether the channels are express (pass-through) or add/drop.

The OSC carries the following types of information:

- **Cisco Discovery Protocol (CDP) packets:** Used to discover neighboring devices
- **IP packets:** Used for SNMP and Telnet sessions between nodes
- **OSC Protocol (OSCP):** Used to determine whether the OSC link is up

  The OSC is on ITU channel 19 at wavelength 1562.23 191.9000 frequency in THz (on 100 Ghz grid).

Note: A Cisco ONS 15540 system on which the OSC is not present is not known to other systems in the network and cannot be managed by any NMS. Without the OSC, a Cisco ONS 15540 system must be managed individually by separate Ethernet or serial connections. Therefore, it is important when adding a node to an existing network of Cisco ONS 15540 systems that the added node have OSC support.

6.13.4 Mounting

In a chassis for 32 wavelengths, the chassis dimensions are 24.5 x 17.3 x 12 in (62.2 x 43.9 x 30.4 cm). Three chassis can be installed in a seven-foot rack. If you are using an external AC-input power supply, only two chassis can be installed in the rack with the power supply. The chassis shelf is pictured in Figure 6-27.

Figure 6-27 Cisco ONS 15540 chassis

The Cisco ONS 15540 fan assembly is located at the bottom of the chassis and is horizontally oriented. Air is drawn in by the fans from the bottom front of the chassis and is exhausted through the vertically mounted modules and through the top back of the chassis. The fan assembly is hot-swappable.

6.13.5 Wavefill

Wavefill is one of many similar terms that refers to the ability to multiplex various data streams onto a single wavelength. This allows a wavelength to be used most efficiently, maximizing the amount of traffic that can be transported between locations. This feature is present in the Cisco ONS 15540.

6.13.6 3R

All transponder modules in the Cisco ONS 15540 are clocked with full 3R (regenerate, reshape and retime), which supports the following bit rates:

- 16 Mb
- 100 Mb
- 200 MbOC3/STM1
- OC12/STM4
- OC48/STM16
- 1.062 Gb
- 1.25 Gb
- 2.124 Gb

This allows the cascading of these services between systems without the requirement for any equipment to regenerate the signal.

6.14 Summary

The Cisco ONS 15540 Extended Services Platform (ESP) is a highly modular and scalable next- generation dense wave division multiplexing (DWDM) platform that integrates data networking, storage, and information streaming over an ultra-high bandwidth-intelligent optical infrastructure that can support any packet, over any wavelength, on any platform.

Key product benefits

The following are the key product benefits:

- **Scalable, incremental bandwidth**

  The Cisco ONS 15540 offers a transparent type 1 transponder line card that supports line rate connectivity of different protocols at speeds ranging from 16 Mbps to 2.5 Gbps. Each line card consists of a motherboard that can be equipped with up to four type 1 transponder modules.

- **Connectivity**

  Each transparent type 1 transponder module supports one TX, RX pair SC connectors for interconnecting the customer premise equipment (CPE).

Key product features

- Compact modular design
- Fault tolerant
- Hot-swappable modules
- Fully integrated end-to-end connectivity
- Multiprotocol transparent support
- Ethernet
- Fast Ethernet
- Gigabit Ethernet
- Enterprise System Connection (ESCON)
- Fiber Distributed Data Interface (FDDI)
- Token Ring
- ATM OC3/STM1, OC12/STM4, and OC48/STM16
- Synchronous Optical Networking (SONET)/Synchronous Digital Hierarchy (SDH)
- Packet over SONET (POS)
- Fiber connectivity (FICON)
- Fibre Channel
- Other proprietary protocols
- Standards-based Management SNMP, CiscoView, and Cisco Transport Manager
- Wavelength conversion (transponder) CPE wavelength: 1310 nm to 1550 nm
- Bit rate and format transparency from 16 Mbps to 2.5 Gbps
- 3R (signal retime, regenerate and reshape) capability
- 100-GHz (0.8 nm) channel spacing based on ITU-T G.692
- Up to 32 channels per fiber, (64 in a fiber pair) in a point-to-point, hubbed, and mesh ring topology
- Optical add/drop multiplexing design
- Optical pass-through capabilities for flexible design options
- Wavelength path switching (1+1) for each fiber; < 50ms

Hardware interface specification

- Power
  - System supply voltage: -48V DC +/- 10% rated
  - System fully populated power consumption: 900 W
- Environment
  - Temperature: (0-40° C)
  - Humidity: 5-95% noncondensing
- Physical
  - 153 lbs
  - Dimensions (inches): 24 x 17.3 x 12 (H x W x D) (14RU)
- Client Fiber Specification
  - Fiber type:
  - Single-mode Fiber: ITU-T G.652
  - Multimode Fiber: 50 m and 62.5 m
  - Trunk Fiber Specification
  - Fiber type
  - Single-mode Fiber: ITU-T G.652
  - Cladding diameter: 125 +- 1.0 m
  - Cutoff wavelength: < 1260 nm
- Transparent Type 1 Transponder Specifications
  - Client-side ports
  - Data rate: 16 Mbps to 2.5 Gbps
  - Client wavelength: 1260 to 1310 nm
  - ITU wavelength: 1530.33 to 1560.61
  - Receiver dynamic range: -8 to -32.0 dBm
  - Transmit output power: +6.0 to +8.0 dBm
  - Fiber type: single mode and multimode
  - Connector type: SC
- Transport Side Multiplexer/Demultiplexer Ports
  - Optical Supervisory Channel (OSC): 1562.23
  - 32 wavelengths ITU-T G.692
  - MU connectors
- System performance
  - Bit error rate 10^{-15} for ESCON and 10^{-12} for all other applications
  - Supported applications: SONET, ATM, 1310 nm Gigabit Ethernet, ESCON, Fiber Channel, and other proprietary protocols within 16Mbps - 2.5 Mbps.

- Remote Protection
  - Fiber route diversity: 1:1 path protection
  - Switch-over time: < 50ms
- Network Element Management Interface
  - RS-232 (DB-25 connector)
  - RS-232 auxiliary port (RJ-45 connector)
  - 10/100 Ethernet (RJ-45 connector)
- Maximum System Configuration 32-Channel System
  - Expandable eight-wavelength optical add-drop multiplexer/demultiplexer with or without Optical Supervisory C Band (OSC)
  - Expandable four-wavelength optical add-drop multiplexer/demultiplexer with or without OSC
  - Two CPU
  - Dual AC power supplies
  - Eight transparent line cards with splitter motherboards populated with four modules each
  - Expandable 16-wavelength terminal optical multiplexer/demultiplexer with OSC; second 16-wavelength terminal optical multiplexer/demultiplexer without OSC
- Regulatory Standards Compliance - Cisco ONS 15540 complies with the following standards:
  - Products bear CE Marking indicating compliance with the 89/366/EEC, 73/23/EEC directive, which includes the following safety and EMC standards
- Safety
  - UL 60950
  - CAN/CSA-C22.2 Number 60950-00
  - EN 60950
  - IEC 60950
  - TS 001
  - AS/NZS 3260
  - IEC 60825-1
  - EN 60825-2
  - 21CFR 1040
  - EMC
  - FCC Part 15 (CFR 47) Class A
  - ICES-003 Class A
  - EN 55022 Class A
  - CISPR22 Class A

  - AS/NZS 3548 Class A
  - VCCI Class A
  - EN 55024
  - ETS 300 386
  - EN 50082-1
- Industry EMC, Safety and Environmental Standards
  - GR-63-Core NEBS Level 3 Requirements
  - GR-1089-Core NEBS Level 3 Requirements
  - ETSI 300 019 Storage Class 1.1
  - ETSI 300 019 Transportation Class 2.3
  - ETSI 300 019 Stationary Use Class 3.1

7

CNT USD and UWM

In this chapter we present an overview of CNT's distance products.

The products that will be covered are:

- CNT UltraNet Storage Director
- CNT UltraNet Wave Multiplexer (DWDM)

7.1 CNT UltraNet Storage Director

The CNT UltraNet Storage Director (USD) is a multi-gigabit switching platform that interconnects host to storage and storage to storage systems across your enterprise. It works at the storage network infrastructure in local, campus and wide area environments to create a high performance, scalable solution that integrates channel technology and high performance network technology. This is used for mission critical storage networking applications, such as disk mirroring, backup/restore, archive/retrieve, data migration, content distribution, and shared storage.

Although the USD is not DWDM technology, CNT does have a DWDM which will be described in 7.2, “CNT UltraNet Wave Multiplexer” on page 204.

7.1.1 UltraNet high capacity architecture

In Figure 7-1 we show an illustration of the USD architecture.

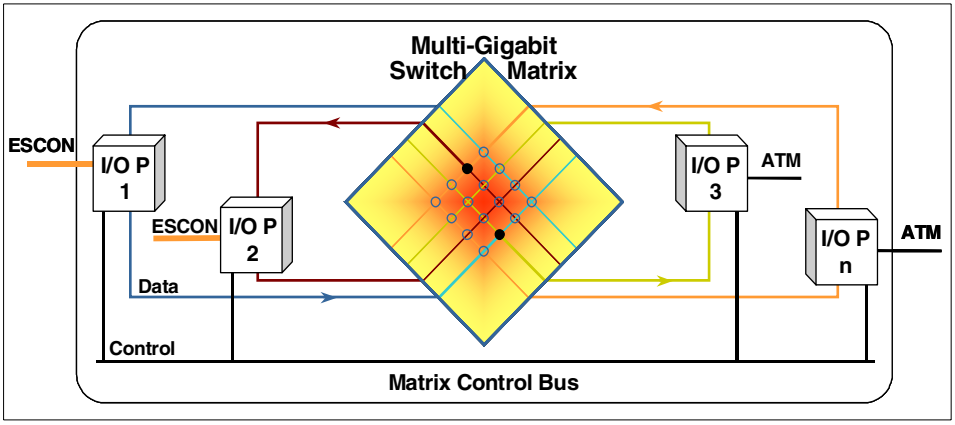


Figure 7-1 USD architecture

7.1.2 Network interface support

The USD supports multi protocols with the following number of ports:

- ESCON: 1-2 ports
- T3/E3: 1-2 ports
- T1/E1: 4 or 8 ports (currently in development)

- ATM
  - OC-3c Single Port
  - OC-12 Single Port (currently in development)
- Fibre Channel: Single Port
- SCSI: 2 or 4 Ports
- Fast Ethernet (10/100): Single Port with IP protocol

The USD will support 12 ESCON channels per node.

In Figure 7-2 we show the network interface support in the USD.

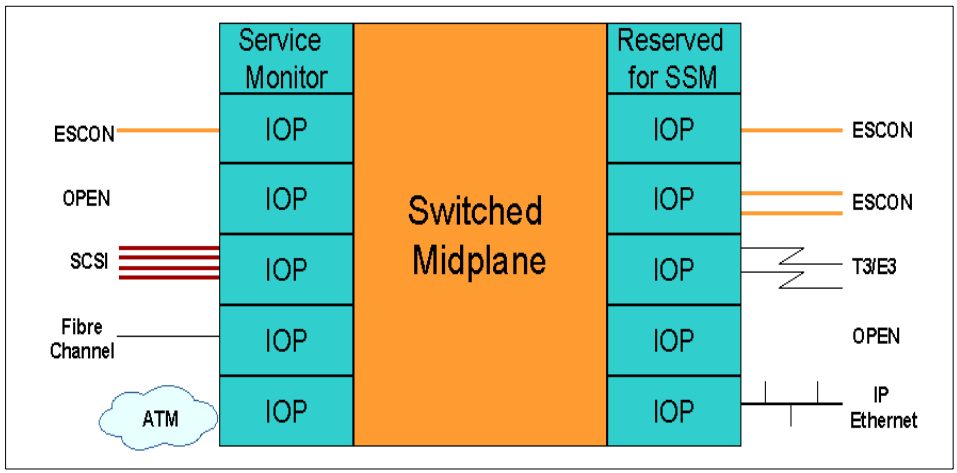


Figure 7-2 Network interface support

7.1.3 Scalability

There are two types of directors. One is UltraNet 9006 Director which has a six slot chassis with five available. The other is the UltraNet 9012 Storage Director which has a 12 slot chassis with 11 available.

In Figure 7-3 we show the illustration of UltraNet 9012 Storage Director.

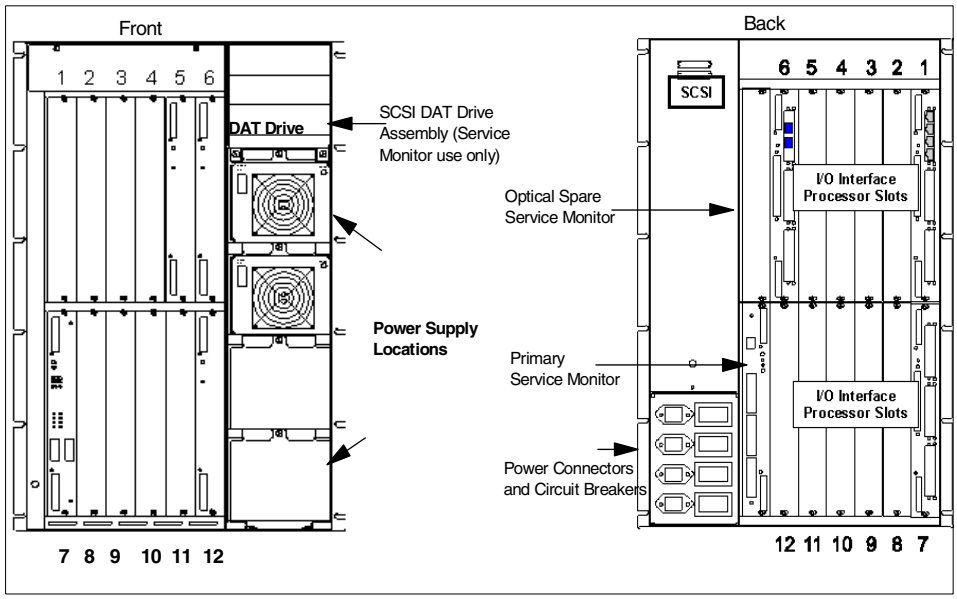


Figure 7-3 UltraNet 9012 storage director

It also has one to eight WAN ports available.

7.1.4 Reliability

The USD brings with it a number of elements, functions and features which all add to the reliability of it.

These include:

- Independent data paths through the mid-plane switch
- Non-blocking mid-plane switch technology
- Multi-processor architecture
- Redundant, load-sharing, hot-swappable power supplies
- Hot-swappable processors
- Optional Hot Spare system service monitor

7.1.5 Availability

To ensure the highest levels of availability, the USD has the following features included in its architecture:

- Any-to-any data pathing with intelligent I/O interface processors
- Traffic management
- Pipelining

- Dynamic load leveling
- Automatic alternate path routing
- Traffic prioritization
- 32 bit Cyclical Redundancy Check (CRC)

7.1.6 Performance

The UltraNet Storage Director's intelligent software incorporates features to improve network throughput and maximize the efficiency of application execution whether across town or around the globe. These features include dynamic load leveling across all available bandwidth, automatic alternate path selection, network-based error detection and recovery, traffic prioritization, data compression, and pipe-lining.

It supports up to one gigabit per path and provides dedicated I/O path processor support.

7.1.7 Management

Management is an important feature of any solution, and the USD management features are described in the sections that follow.

System Management

The system is managed by a Windows based GUI interface. Configurations are stored on the service monitor and DAT drive.

In Figure 7-4 we show an example of the system management panel.

.

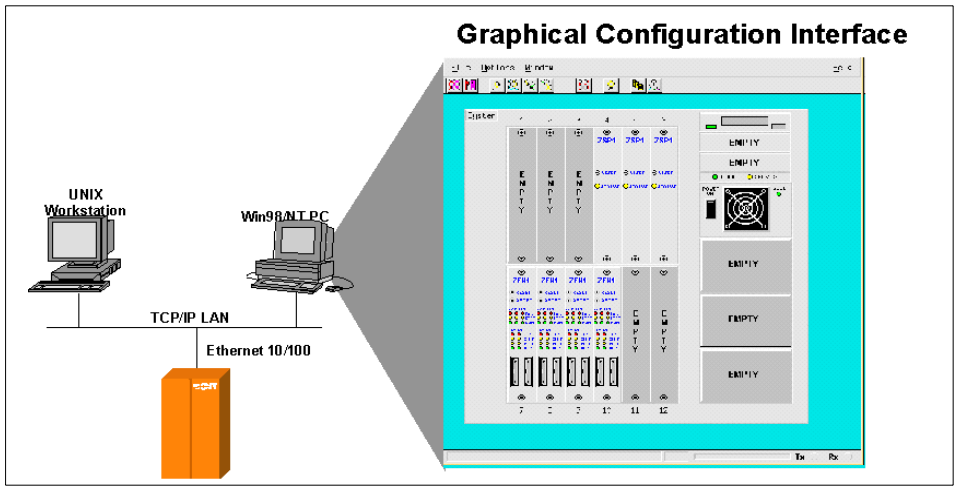


Figure 7-4 System management

Network Management

CNT Management Facility (CMF) management software is an SNMP manager optimized for the unique characteristics of the CNT product line.

The CMF management software provides an easy point-and-click interface representing CNT nodes, with drill-down views of individual boards and power supplies. If specified thresholds are exceeded, CMF management software generates alerts or alarms and sends them to a designated SNMP manager.

In Figure 7-5 we show an example of network management configuration using CMF.

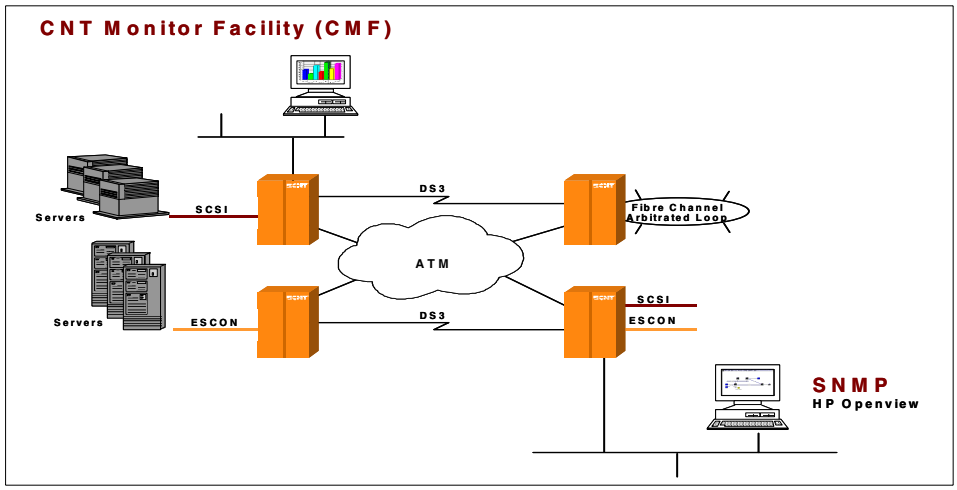


Figure 7-5 Network management

7.1.8 Server support

The USD supports diverse server platforms including OS/390 and UNIX, NT systems.

7.1.9 Specifications

The specifications related to its dimensions, power usage and temperature environment are described in Table 7-1.

Table 7-1 Specifications of USD

	12-slot	**6-slot**
Height	43 in (1092.2 mm)	19.25 in (489 mm)
Width	23.06 in (599.4 mm)	23.06 in (599.4 mm)
Depth	36.2 in (918.8 mm)	27.75 in (705 mm)
Weight Fully Configured	400 lbs (181 kg)	200 lbs (90.7 kg)
AC Voltage	90-240 VAC	90-240 VAC
Hertz	47-63 Hz	47-63 Hz
Phases	1	1
Max. Amps	110V/21A, 220V/10.5A, 208V/11.5A	110V/21A, 220V/10.5A, 208V/11.5A

	12-slot	6-slot
Max. Operating VA	2090A	2090A
Input Power Requirements	90-240VAC, 47-63Hz	90-240VAC, 47-63Hz
Operating Temperature	40F to 104F (5C to 40C)	40F to 104F (5C to 40C)
Operational Relative Humidity	5% to 80% (non-condensing)	5% to 80% (non-condensing)

7.1.10 Serviceability

From a serviceability point of view the USD can be fully SNMP-managed by CNT or enterprise manager client products. Configuration wizards simplify USD installation and configuration.

7.1.11 Compression

The UltraNet Storage Director's switched mid-plane architecture supports multi-ported I/O interface processors, making the full bandwidth of each switch port available for each I/O processor. In addition, compression ratios between 2:1 and 20:1 can be achieved depending on the compressibility of the data. This can more than double the capacity of existing bandwidth and delay investments in additional infrastructure.

7.1.12 Distance

There is no limitation with respect to distance. However, you must consider the performance of applications which depend on network speed, copy solutions (synchronous or asynchronous copy), and the size of data sent to an alternate site at a particular period of time.

7.2 CNT UltraNet Wave Multiplexer

The CNT UltraNet Wave Multiplexer (UWM), a DWDM product, can dramatically increase available fiber bandwidth, expanding the amount of data that can be handled by a SAN. It eliminates the need to lease or install multiple fiber optic cables, which allows companies to dramatically cut costs.

7.2.1 Componentry

The UltraNet Wave Multiplexer (UWM) system consists of the following hardware components:

- A 19 inch 6U chassis which includes a section for fans and power supplies.
- A Dense Wavelength Division Multiplexer (DWDM) module.
- Eight channel card capacity supporting data services up to speeds of 2.5 Gb/s.
- An SNMP Agent/CMS Master card or an SNMP/CMS expansion card.
- An optional fiber optic switch (main chassis only).

The UWM can hold up to eight channel cards. Each channel card has a specific wavelength that corresponds to a wavelength on the ITU-T grid. The eight channel cards are positioned below the DWDM unit. Channel card 1 is located at the far left of the system. The channel cards are placed consecutively (1-8) in the UWM. The optional fiber optic switch is located to the right of the eighth channel card and the SNMP Agent/CMS Master card is located at the far right of the system.

The DWDM module is positioned above these cards and this is shown in Figure 7-6.

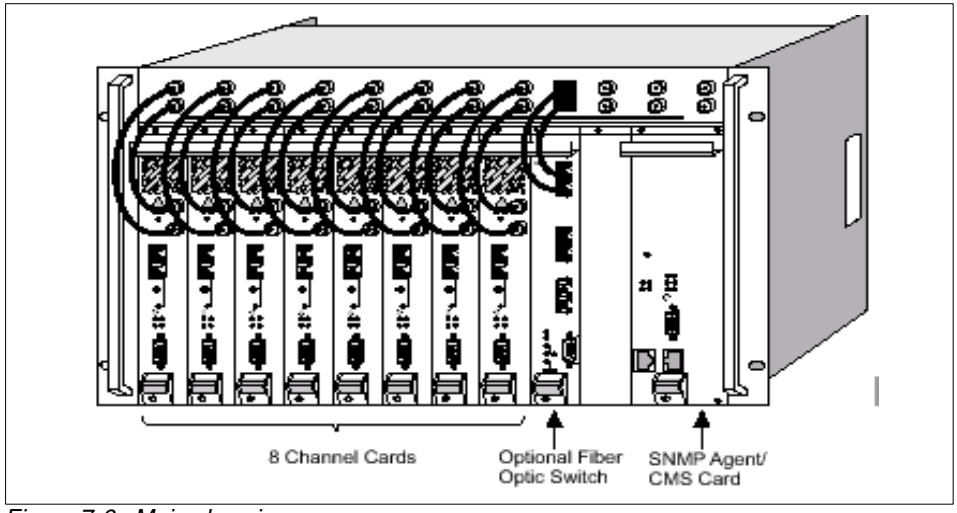


Figure 7-6 Main chassis

Note: Each channel card has a unique wavelength. The wavelengths for the channel cards are provided. The wavelength (nm) is displayed on the channel card and beside the corresponding connectors on the DWDM. The wavelength of the channel card installed in the slot should correspond to the wavelength displayed on the DWDM connectors above the slot.

An expansion chassis can be connected to the main system and provides an additional eight channel cards. This system has a similar layout but does not have a fiber optic switch (the fiber optic switch is only used by the main system) and features an SNMP/CMS expansion card.

Although the DWDMs of both systems look similar, they are technically different and cannot be exchanged. The DWDM on the main chassis offers connections for monitoring and management, the DWDM on the expansion chassis does not require these. In Figure 7-7, we show the expansion chassis of the UWM.

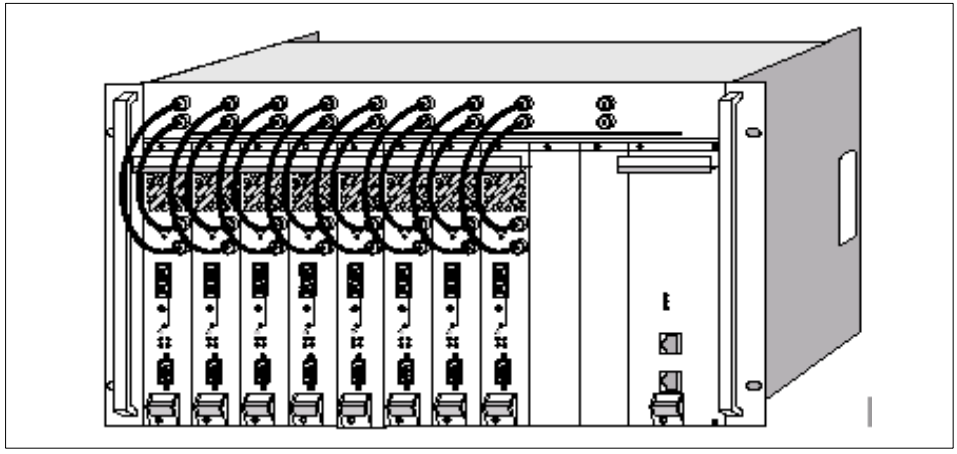


Figure 7-7 Expansion chassis

Channel card

The channel cards convert standard interface fiber optic signals in the regions of 850 nm and 1310 nm into ITU conform signals in the region of 1550 nm for transmission over the line and also perform backward conversion.

Each of these signals conforms to a wavelength on the ITU-T G.692 grid. Each channel card in the UWM system transmits data over the line at a specific wavelength. Each wavelength is designated for only one channel card and cannot be used by any of the other channel cards in the system. For example, the channel card 1 uses a wavelength of 1558.98 nm, regardless of whether there is a card with a data transmission rate of 622 Mb/s, 1.25 Gb/s, or 2.5 Gb/s.

There are three types of channel cards:

- One with a data transmission rate of 622 Mb/s
  - Line: single-mode (SM) DFB laser with FC connector
  - Interface: 1310 nm multi-mode (MM) LED with SC connector
- One with a data transmission rate of 1.25 Gb/s
  - Line: SM DFB laser with FC connector
  - Option 1: 1310 nm SM/MM LED with SC connector
  - Option 2: 850 nm MM LED with SC connector
- One with a data transmission rate of 2.5 Gb/s
  - Line: SM DFB laser with FC connector
  - Interface: 1310 nm SM LED with SC connector

Each channel card has a unique product identification number that is assigned based on the data transmission rate of the card and the channel number. Channel cards one through eight are data transmission rate of the card and the channel number. Channel cards one through eight are installed in the main chassis. Channel cards nine through sixteen are installed in the expansion chassis.

7.2.2 Scalability

The UWM concentrates up to sixteen input channels on either a single fiber optic strand or a fibre optic cable pair.

Up to eight channel cards, each a specific wavelength, can be installed on a single chassis. Two chassis can be interconnected. Up to 32 channels are supported over fiber pair; 128 when combined with UltraNet Optical Multiplexer.

Each input channel is capable of data rates up to 2.5 Gb/s, for a combined system data rate of 40 Gb/s.

The following is the maximum capacity of the DWDM system:

- Up to 16 (32) x 2.5 Gbps per single fiber (fiber pair)
  - 16 channel DWDM with 2.5 Gbps cards per single fiber)
- Up to 32 (64) x FC/FICON/GbE per single fiber (fiber pair)
  - 16 channel DWDM with Gigabit TDM cards per single fiber
- Up to 128 (256) x ESCON per single fiber (fiber pair)
  - 16 channel DWDM with Gigabit TDM cards and ESCON TDM in front per single fiber

7.2.3 UltraNet Wave Optimizer

The UltraNet Wave Optimizer (UWO) multiplexes up to four independent 200 Mb/s ESCON channels onto a fiber pair. The signal is transmitted over the fiber pair at a data rate of 1.062 Gb/s. UWO also transports up to two FC / GbE (1.0625/1.25 Gbs) channels on one 2.488 Gbps link (OC48 / STM16 speed).

The UltraNet Wave Optimizer is available in three chassis configurations: a one slot, a three slot, and an eight slot chassis. If you use ESCON only, a one slot chassis provides a basic configuration of four ESCON channels. A two slot chassis provides two options. Either a dual four ESCON or when combined with an optional WDM module, it provides an eight ESCON channel WDM system. An eight slot chassis can either be used as a front-end to the UltraNet Wave Multiplexer or on its own to provide up to eight point-to-point connections. If the eight slot chassis is used as a front end to the UltraNet Wave Multiplexer, it can provide up to 64 ESCON channels on a single fiber optic line or fiber pair.

The ESCON multiplexer (EMX) cards of the UltraNet Wave Optimizer each provide four ESCON channels. The signals from the four ESCON channels are combined onto a fiber pair using time division multiplexing (TDM) technology. Using TDM the input data streams from the four channels are combined by assigning each input data stream a different time slot in a set. Using this technology, the EMX cards repeatedly transmit a fixed sequence of time slots over a single transmission channel. The EMX cards provide a time slot for each of the four ESCON channels featured on the EMX card, as well as two additional time slots for control and management.

7.2.4 Connectivity

Each DWDM module offers eight ITU channel card mux/demux ports for connection to the channel cards. The two DWDMs available for the main chassis offer connections for channel cards one through eight. The DWDM of the expansion chassis offers connections for channel cards nine through sixteen.

Fiber pair DWDM with expansion port

This module provides eight ITU channel card mux/demux ports (for connections to channel cards one to eight), a line port (for a fiber pair), an expansion port for connection to an expansion chassis, a monitor port, and a separate management port.

In Figure 7-8 we show a Fibre Pair DWDM with expansion port.

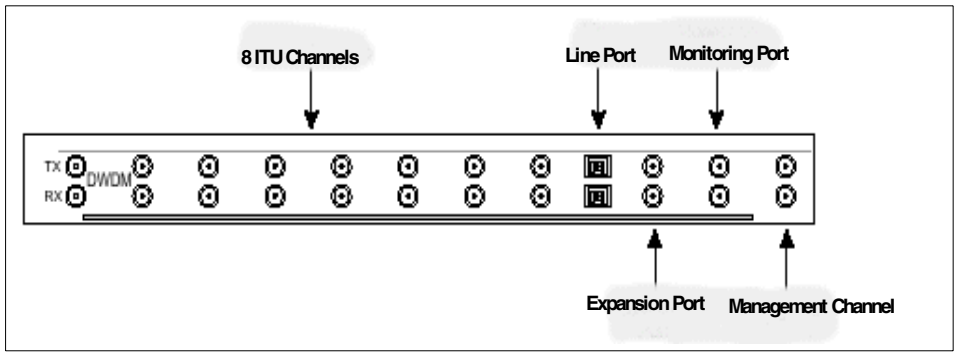


Figure 7-8 Fibre pair DWDM with expansion port

Using the expansion port of this module, you can connect your UWM main chassis to the line port on the DWDM of the expansion chassis. By adding an expansion chassis, eight additional ITU channels (channels nine to sixteen) are available.

The monitoring port allows for optical spectrum and power measurements during normal operation without disturbing data transmission. Connecting a fiber optic spectrum analyzer (OSA) allows you to see a display of the spectral distribution of the ITU signals on the line. The management port offers access to a separate 1310 nm channel on the line.

Single fiber DWDM with expansion port

This module provides eight ITU channel card mux/demux ports (for connections to channel cards one to eight), a line port (for a single fiber), an expansion port for connection to an expansion chassis offering an additional eight ITU channels, a monitor port, and a separate 1310 nm management port.

Up to sixteen ITU channels can be multiplexed onto a single fiber optic line when this DWDM is connected to the DWDM on the expansion chassis.

We show the illustration of single fiber DWDM in Figure 7-9.

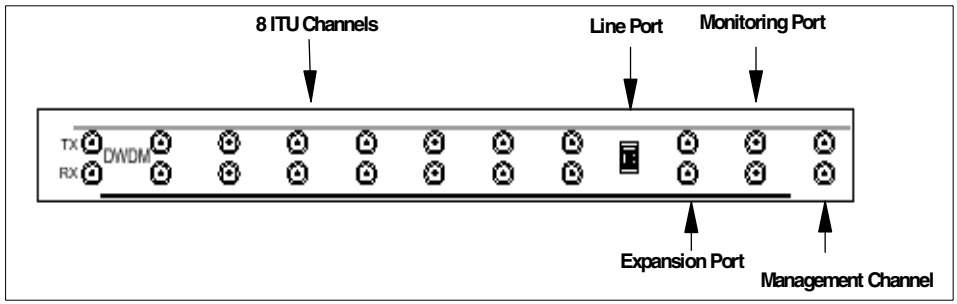


Figure 7-9 Single fiber DWDM with expansion port

7.2.5 Protocols

The UWM supports Fibre Channel, ESCON, Gigabit Ethernet, SONET/SDH and ATM, Fast Ethernet, and handles bit rates between 10Mb/s and 2.5 Gb/s.

The supported protocols are as follows:

- SONET/SDH (ATM)
  - OC-1 / STM-0 (51.84 Mb/s)
  - OC-3 / STM-1 (155.52 Mb/s)
  - OC-12 / STM-4 (622.08 Mb/s)
  - OC-24 / STM-8 (1244.16 Mb/s)
  - OC-48 / STM-16 (2488.32 Mb/s)
- Ethernet
  - Fast Ethernet 100BaseFX (125 Mb/s)
  - Gigabit Ethernet 1000BaseFX (1.25 Gb/s)
  - Gigabit Ethernet Double (2.5 Gb/s)
- FDDI (125 Mb/s)
- ESCON (200 Mb/s)
- FICON (1.062 Gb/s)
- Fibre Channel
  - Fibre Channel (1062.5 Mb/s)
  - Double Fibre Channel (2125 Mb/s)
- Coupling Link
  - Coupling Link (1062.5 Mb/s)
  - Double Coupling Link (2125 Mb/s)
- Digital Video SDI (270 Mb/s)

- Proprietary protocols (digital only) within the data rates of:
  - 10 Mb/s and 622.08 Mb/s if channel cards type FX-IA are installed
  - 100 Mb/s and 622.08 Mb/s if channel cards type FX-IG, FX-IH are installed

7.2.6 Cables and equipment

The following sections list the cables, connectors, and other equipment required for installation of UWM.

CNT-supplied cables and equipment

In Table 7-2 we show the CNT-Supplied cables and equipment required for installation of the UWM.

Table 7-2 CNT-supplied cables and equipment

Type	Description
Modem	CNT supplies a modem for remote support of the UWM (domestic orders only).
Power Plug	CNT supplies the power cords for the UWM system. The UWM chassis requires two power cords and the modem requires 1. The power cords are the 2 meters (6.6 feet) long type, and two are required for redundancy.
DWDM to Channel Card Interconnect Cables	CNT supplies the cables connecting the channel cards to the DWDM module. These cables are pre-installed.
VT100 Access Cable	CNT supplies the cable for VT100 access.
Chassis Interconnect Cable	If an expansion chassis is ordered, CNT supplies the chassis interconnect cable. This cable is used to connect the DWDM of the expansion chassis to the DWDM of the main chassis.
Management Cable	If an expansion chassis is ordered, CNT supplies the 8-pin RJ45 cable used to connect the SNMP Agent/CMS Master card to the SNMP/CMS expansion card. The management cable is 3 meters (9.84 feet) long. The maximum distance between the main and expansion chassis is 3 meters.
Chassis Control Cable	If an expansion chassis is ordered, CNT supplies the cable (DB-9) used to connect the fan modules of the main chassis and expansion chassis.

Customer-supplied cables and equipment

In Table 7-3 we list the cables and equipment that the customer must supply for the UWM system.

Table 7-3 Customer-supplied cables and equipment

Type	Description
Power Receptacles	These should conform to the power receptacle requirements.
Interface Cables	Cables that connect the customer's equipment to the interface ports on the channel cards of the UWM (duplex preferred). These cables require SC connectors on the UWM side.
Line Cables	Cables that connect the local UWM to the remote UWM. If you have a single fiber connection, you must use SC/APC cables. If you have a fiber pair connection, you must use SC/UPC cables.
Patch Panels	Device allowing temporary connections to be made between incoming and outgoing lines.

7.2.7 Management

The Central Management System (CMS) can be accessed via two methods:

- VT100 control port
- SNMP interface

To allow for management by the CMS card, all cards and modules provide management functionality. Additionally, certain units can use the line to the remote counterpart for remote management of the system.

The management sections of the chassis can be connected using an 8-pin RJ45 cable. The sixteen channel system consists of one main chassis containing the SNMP Agent/CMS Master card and an expansion chassis containing an SNMP/CMS expansion card.

The SNMP agent uses Management Information Bases (MIB). MIBs define the variables which are used to control an SNMP device or to retrieve data from the device. The format of the MIBs as well as global sections are defined in the SNMP standard. MIBs are written in a special language (ASN 1) and are in plain ASCII text. This means that you can read MIBs using any available editor.

The SNMP/CMS card access via VT100, as well as SNMP, allows the user to retrieve operational data and configure the UWM system remotely. Both accesses must be protected against intruders as well as inadvertent configuration changes by the network manager.

Any time you disconnect and reconnect the VT100 terminal the password request turns up on the terminal.

SNMP's "community" mechanism allows you to regulate action to vital functions. A community is like a password. It grants access to a certain level of SNMP tasks. You must assign a community to each level. For the first few steps with SNMP, you may assign the same community to each level; however, you will want to set different communities later on. Designating different communities prevents potentially dangerous actions from being triggered accidentally. Also, this allows others to have access to simple, low level data retrieval operations.

Note: The "supervisor access" community should only be used temporarily, as this community provides complete access to any object.

7.2.8 Redundancy

The UWM is has redundant components as part of its construction. They are:

- Redundant power supplies fans and AC input
  - Coupled with the use of two power cords
- Hot-swappable channel cards, power supplies and fans
- No downtime in case of failure, upgrade or adding capability
- Alternate paths with fiber optic switch for line protection
  - Auto failover in case of line problems

7.2.9 Serviceability

All of the components in the UWM are hot-swappable and can be replaced while power is applied to the system.

This is coupled with remote service capabilities, on-site service, sparing at the customer site or at the CNT warehouse (with the parts shipped the next day), dedicated remote support for specific activities and installation services.

7.2.10 Monitoring and diagnostics

You can monitor or diagnose the UWM by using following features and by the remote dial-in access feature, as well as local access:

- RS232 maintenance interface
- VT100 terminal emulation
- Remote dial-in access
- Self test, BERT test and loop-back testing
- Optical spectrum and optical power measurement ports
  - The DWDM installed in the main chassis features a monitor port for measurement purposes. You can use this port to verify that the fiber optic components are operating properly without disturbing data transmission on the line. A small portion (5%) of the optical line signal is forwarded to this port. Connecting an optical spectrum analyzer (OSA) to the monitor port allows you to see a display of the spectral distribution of the light signals on the line. You can use this to verify the correct frequency of the optical channels.
  - The DWDM of the expansion chassis does not have the monitoring, management and expansion ports featured on the DWDMs described above, as they are already provided by the DWDM of the main chassis. This module can be used for either a single fiber or fiber pair connection.

7.2.11 Specifications

These are the specifications that relate to weight, height, power and operating temperature.

Environmental/electrical/power specifications:

- Height x width x depth
  - 270 mm (10.64 inches)
  - 485 mm (19.11 inches)
  - 530 mm (20.88 inches)
- Weight: 23 kg (50.7 lbs.)
- Operating temperatures: 5 to 40 °C (41 to 104 °F)
- Humidity: 10 to 90% non-condensing
- Power input: 115/230 VAC dual range auto-sensing
- Frequency: 47-66 Hz

Optical specifications

In Table 7-4 we provide the optical specifications for the UWM system.

Table 7-4 UWM optical specifications

Specification type	
Wavelength	1550 nm window (management port: 1300 nm window)
System optical output power	< 40 mW
Optical power budget	Dependent on installed components
Maximum reluctance of cabling	20 dB (duplex), 55 dB (simplex)
Auto power shutdown delay	< 50ms
Switch-over delay (back-up)	< 50 ms (whole system)

8

INRANGE 9801 SNS and Spectrum 2000

In this chapter we present an overview of the INRANGE distance products.

The products that will be covered are:

- INRANGE 9801 SNS
- INRANGE Spectrum 2000 (DWDM)

8.1 INRANGE 9801 Storage Networking System

The INRANGE 9801 Storage Networking System (SNS) extends remote high speed, volume performance storage devices (such as disk mirroring) to support disaster recovery environments and business continuance applications.

The 9801 SNS extends ESCON CNC channels over Telco networks at distances beyond specified ESCON distance limitations. This 9801 SNS product provides high-speed communications pipes using standard telephone networks, such as ATM OC-3, T3/E3, T1, and others.

In using standard telephone networks, the 9801 SNS architecture, therefore, incorporates data compression. Likewise, to meet high speed, volume performance storage device (such as disk mirroring) performance requirements, the 9801 SNS is designed with a low latency architecture.

8.1.1 Componentry

There are three major components to the 9801 storage networking system.

- Server: 9801H (Aspen) / 9801L (Hudson)
- ESCON/HSSI Channel Adapter Module
- Network Interface Modules (ATM, T3, HSSI, etc.)

The 9801 base system consists of the following assemblies:

- Server(s); 1 - min./2 - max.
- ESCON/HSSI Channel Adapter Module(s)
- Network Interface Module(s)
- Ethernet Module
- Modem (domestic shipments only)
- Network Hub
- Power Distribution Unit (PDU)
- Keyboard/Mouse/Status Monitor

The 9801 SNS Network Interface Modules are:

- ATM Module (mounted internally; 9801 Server)
- T3 Interface Module (mounted internally; 9801 Server)
- HSSI Interface Module (mounted internally; 9801 Server)
- 10/100 Ethernet Module
- 1000 Ethernet Module

The 9801 SNS optional interface assemblies are as follows:

- T1 Inverse Multiplexer for ATM assembly, for T1 interfaces (rack mounted assembly)

- OC3/T3 Converter assembly, for OC3/T3 interfaces (mounted on 9801 option shelf)
- OC3/E3 Converter assembly, for OC3/E3 interfaces (mounted on 9801 option shelf)

8.1.2 Protocols supported

The SNS supports the following protocols:

- OC-3: ATM; SM or MM fiber; duplex SC connector; SONET or SDH framing formats
- T3/E3: DS-3 framing; Belden 8281 or WECO 728A compatible cable with BNC connector
- HSSI: HDLC framing with standard 50-pin, high density (HD) connector
- T1/E1: DS-1 framing with DB-15 connector (using an internal ATM card and an external ATM to T1 inverse multiplexer).

8.1.3 Scalability

From a footprint and scalability view there are two cabinet options:

- There are two large or mid-size chassis per cabinet

  Chassis options
- There are three types which are large, mid-size, and small.

  We show each chassis's configuration in Figure 8-1.

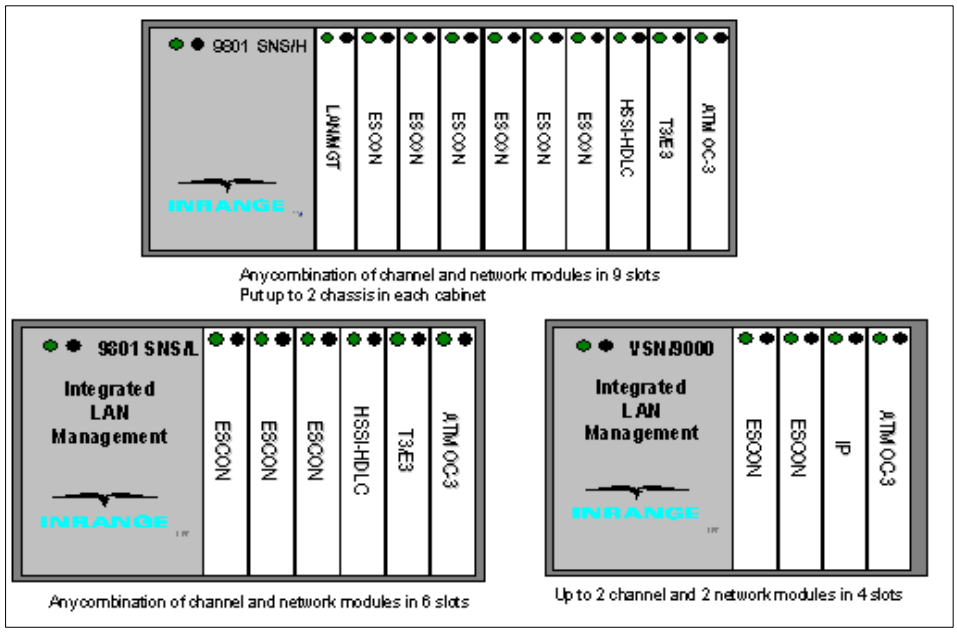


Figure 8-1 9801 SNS chassis

8.1.4 Data compression

The 9801 SNS provides individual compression for each channel. It compresses up to 64:1. But the validated compression ratio is about 4:1 to 12:1 for your production environment. It improves synchronous performance reducing back plane traffic as shown assuming 4:1 compression. This is shown in Figure 8-2.

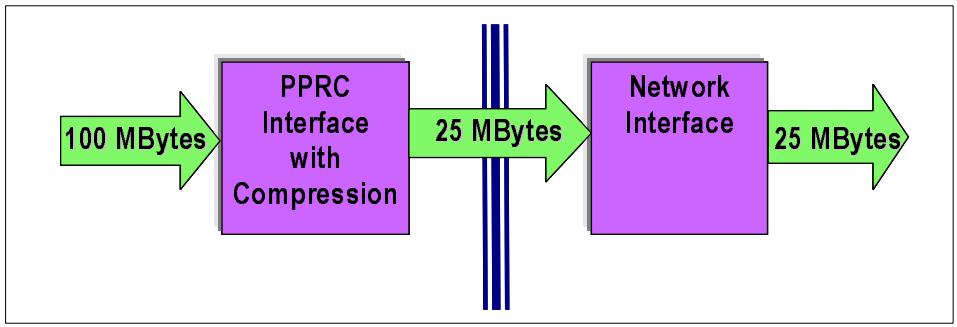


Figure 8-2 Channel card compression

8.1.5 Management

The SNS bandwidth and network cost control tool is performed by line and channel utilization monitoring (LUM). This feature visually represents both actual link activity and effective throughput (based on data compression) per link interface and channel adapter, at any time interval you choose.

You always know exactly how much bandwidth you need, and where you need it. And using SNMP you can obtain alarms for link and/or path status. This is shown in Figure 8-3.

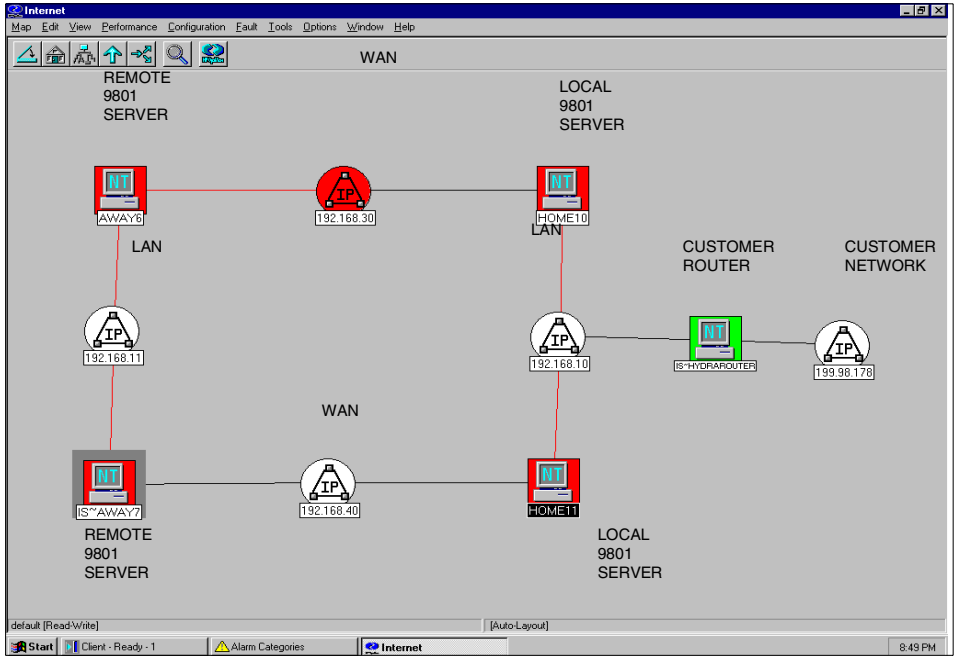


Figure 8-3 SNMP control

8.1.6 Redundancy

INRANGE 9801 has redundant power supplies, coolers, and dual AC sources. And, it has been configured for network link failover and network link recovery automatically, so that operator intervention is kept to an absolute minimum.

8.1.7 Specifications

The physical, electrical and environmental specifications are described here.

The physical specifications are:

- Height: 45 in
- Width: 19 in
- Depth: 36 in

The electrical specifications are:

- Voltage: 100-120 / 200-240
- Frequency: 50-60 Hz, Single Phase
- KVA: 1.0
- Channel Type: ESCON, HSSI/HDLC

The environmental specifications are:

- Temperature: 60-90F (15-32C)
- Relative Humidity: 20%-80%

8.2 INRANGE Spectrum 2000

The INRANGE Spectrum 2000 optical networking solutions deliver full bandwidth for metropolitan storage networking, LAN extension, open systems and more.

INRANGE offers the IN-VSN Spectrum Series fiber optic multiplexers. Designed expressly for enterprise applications, Spectrum solutions are based on optical wavelength division multiplexing, which combines multiple high bandwidth applications over a single fiber pair.

INRANGE has several distance products such as TDM (INRANGE Spectrum 1000) and WDM (INRANGE spectrum 1), and DWDM (INRANGE Spectrum 2 and Spectrum 2000).

However, we will describe only the INRANGE Spectrum 2000 as it is targeted for the high-capacity, high-density requirement of major data centers where large numbers of channels connecting are required and is the focus of this redbook.

In Figure 8-4 we show a picture of Spectrum 2000.

Figure 8-4 INRANGE Spectrum 2000

8.2.1 Management

The following components are used for management of the Spectrum 2000:

- Spectrum management suite
- Optical Supervisory Channel (OSC)
- SNMP agent
- HP Openview integration

Spectrum management suite

The intelligent and user-friendly point-and-click Spectrum management suite helps create new high-value, high-margin services featuring seamless end-to-end monitoring, control, provisioning and remote configuration. It includes an element manager, network manager, and network planner.

8.2.2 Serviceability

The Spectrum 2000 is easy to install and upgrade with no user downtime. And you can also analyze the problem from remote.

8.2.3 Scalability

Spectrum 2000 DWDM system supports up to 32 channels on a single fiber pair. And if you combine it with spectrum 1000 which is a TDM, it supports up to 256 applications and these can be transported over one single fiber pair.

8.2.4 Availability

Optical protection is a major requirement in metro optical networks especially for those services that are natively unprotected such as Fibre Channel.

The Spectrum 2000 provides line protection via a redundant fiber pair for point-to-point topologies. In case of fiber breakage a remote switch module detects loss of signal and switches all channels over to the protection line. The line protection covers line failures.

8.2.5 OADM

In Figure 8-5 we show the topology which allows us to add/drop up to 32 unprotected channels east and up to 32 unprotected channels west, in steps of four or eight channels.

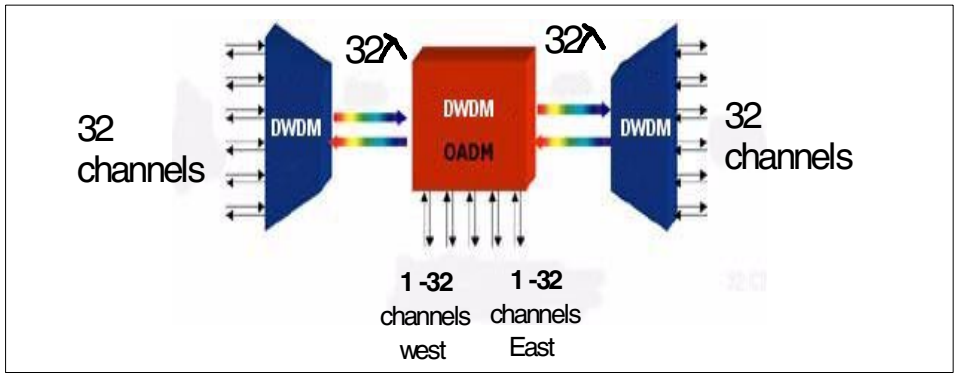


Figure 8-5 Linear add/drop links

The maximum pass-through capacity is 32 channels.

8.2.6 Protocols

The spectrum 2000 transparent fiber-in/fiber-out interfaces support all protocols between the range of 10Mbps and 10Gbps.

These include:

- OC-3/12/48
- STM-1/4/16
- Gigabit Ethernet
- Fast Ethernet
- FDDI
- ATM

- ESCON
- FICON
- One Gigabit Fibre Channel
- Two Gigabit Fibre Channel
- Coupling link
- Any proprietary protocol operating within the specified data range

8.2.7 Power requirements

The power requirements are:

- Voltage 90-132/180-264VAC, autoranging -60VDC to -38VDC
- Power consumption maximum 150 Watts

8.2.8 Cabling interface

The cabling interfaces are:

- Optical ports are mini SC.
- It needs Ethernet cable for management.

8.2.9 Wavelengths

The wavelengths are:

- Local port: 850/1310 nm
- Remote port: C/L-band
- Channel spacing: 200 GHz

8.2.10 Standards compliance

The Spectrum 2000 carries the ‘CE’ mark as an indication of conformity. Other agency compliances are FCC, TUV, GS, UL, CUL.

8.2.11 Clocking and bit racing

The clock options are as follows:

125/155/200/266/622/1062/1250/2488 MHz/10 GHz

8.2.12 Supported topologies

The supported topologies are:

- Point-to-point
- Linear add/drop

- Ring

We discuss these and their variations in 2.3, “DWDM topologies” on page 25.

While point-to-point topologies are largely prevalent in storage networking applications, there is a trend to move to ring topologies to support multiple add/drop locations efficiently within a metro area.

8.2.13 Distance

In terms of distance, the remote span is up to 60 miles/100 km (1550nm).

8.2.14 Componentry

The Spectrum 2000 packs plenty of flexibility into a small space. Its modular architecture comprises a rack-mountable 19 inches chassis and hot swappable modules that can selected to meet the network and application requirements.

8.2.15 Modules

Modules in Spectrum 2000 are:

- WDM channel modules
- TDM channel modules
- DWDM Multiplexer/Demultiplexer (MDXM)
- CWDM Multiplexer/Demultiplexer (CMDXM)
- Band Splitter Modules (BSM)
- Hub Module (HUB)
- Remote Switch Module (RSM)
- Optical Supervisory Channel Module (OSCM)
- Network Element Management Interface (NEMI)
- Device Element Management Interface (DEMI)

8.2.16 Specifications

The physical, environmental and optical specifications are:

- Weight 33 pounds/15 kg per base unit fully loaded
- Dimensions: width x depth x height 19 in. x 12 in. x 8,8 in. / 482 mm x 305 mm x 223 mm
- Optical ports mini SC
- Environmental
  - Temperature +5 to +50°C
  - Relative humidity 10% to 90% non-condensing

9

Nortel OPTera Metro 5300 Multiservice Platform

This chapter gives an overview of the Nortel OPTera Metro 5300 Multiservice Platform.

The Nortel OPTera Metro 5300 is an easy-to-deploy cabinet DWDM solution that extends Nortel OPTera Metro 5200 functionality and flexibility to metropolitan enterprise and customer premise environment applications. The OPTera Metro 5300 Multiservice Platform is a cabinet solution that extends DWDM service into the enterprise and customer premise environment spaces.

It is a customized packaging option that contains one or two OPTera Metro 5200 Multiservice Platform shelves, which are premounted, pretested and preconfigured for easy plug and play installation.

In Figure 9-1 we show the OPTera Metro 5300 and 5200.

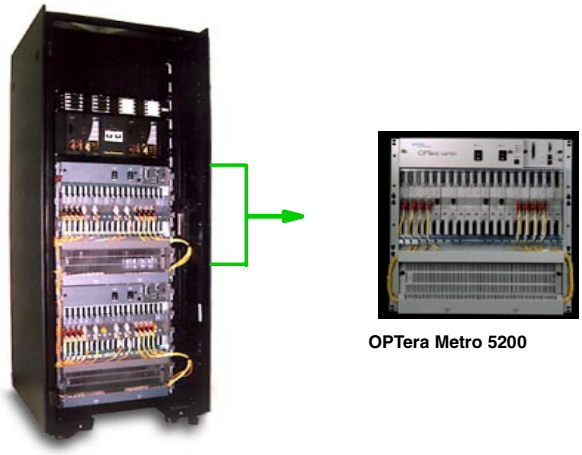


Figure 9-1 OPTera Metro 5300 and OPTera Metro 5200

The OPTera Metro 5300 provides a quick and easy cabinet deployment solution to help harness the DWDM capabilities of the OPTera Metro 5200, which delivers network scalability, wavelength-level manageability, and protocol and bit-rate independence.

The OPTera Metro 5300 meets the demands for high-speed I/O, network connections and data transfer for customers with multiple servers in different locations, as well as enterprise customers who need continuous access to data and applications, including e-business, business-intelligence and disaster recovery solutions.

The Nortel OPTera Metro 5300 is a forecast-tolerant, DWDM data-optimized, survivable platform for metropolitan access and interoffice application.

9.1 Typical uses

The OPTera Metro 5300 is suited for multiple applications involving multiple enterprise connectivity, such as:

- 16 Mbps to 10 Gbps
- Optical Ethernet over WDM (100M, 1G, 10G)
- Storage area networking
- Network attached storage
- Server-to-server networking through IBM's Geographically Dispersed Parallel Sysplex (GDPS): the ultimate e-business availability solution

- Voice, data and video

In this chapter we describe SAN connectivity.

9.1.1 Types of sites

There are three types of sites in an OPTera Metro 5200 network:

- Terminal sites
- OADM sites
- OFA sites

Terminal sites consist of OPTera Metro 5200 DWDM shelves that are provisioned as terminal shelves on the system manager. At this site, there must be a terminal shelf for every wavelength band used in the network: all wavelengths in the system terminate at this location. OPTera Metro 5200 networks can contain amplified terminal sites or unamplified terminal sites. Terminal sites are sometimes called hub sites when used in hubbed-ring configurations.

OADM sites consist of OPTera Metro 5200 DWDM shelves that are provisioned as OADM shelves on the system manager. At this site, single or multiple OADM shelves are placed to gain access to specific wavelengths in the system, so that some wavelengths are terminated, and some are optically passed through at that location. OPTera Metro 5200 networks can contain amplified OADM sites or unamplified OADM sites. OADM sites are sometimes called remote sites.

OFA shelves do not need to be collocated with terminal shelves or OADM shelves. If you provide independent Ethernet communications, you can install OFA shelves in intermediate sites; these sites are called OFA sites. You can also deploy OFA shelves in terminal sites or OADM sites as needed to overcome system losses. These shelves are provisioned as OFA shelves on the system manager.

Any client signal — such as SONET/SDH, Async FOTS, Gigabit Ethernet, or ATM — is connected to OPTera Metro 5200 using short-reach 1310 nm or 850 nm interfaces. The circuit packs are bit-rate and protocol independent up to 1.25 Gbit/s. A circuit pack is available that supports SONET/SDH at 2.5 Gbit/s.

In Figure 9-2 we show an example of Terminal site and OADM sites.

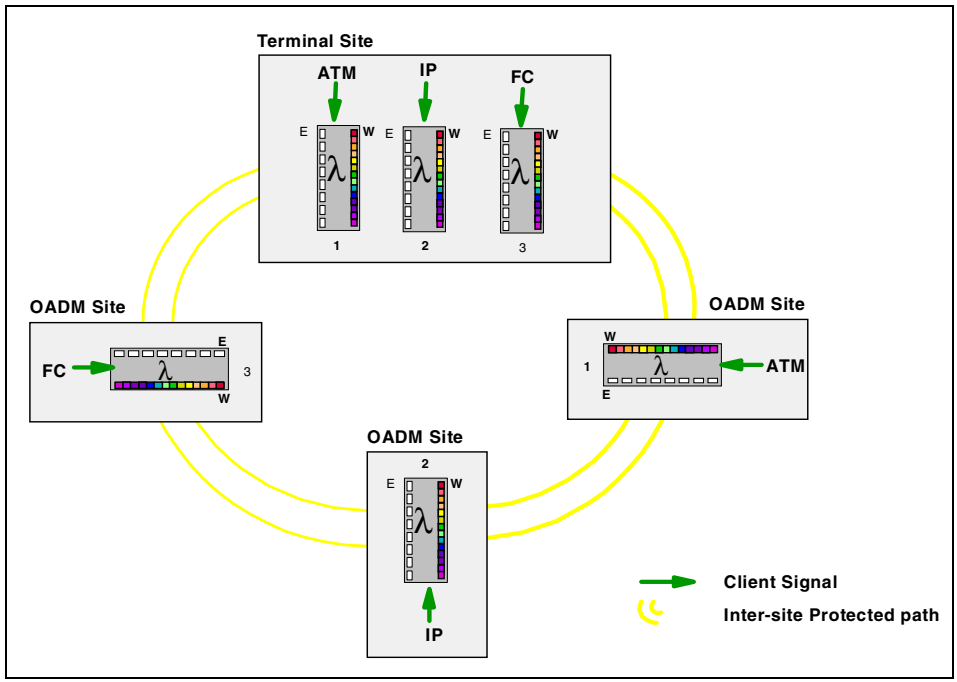


Figure 9-2 Example of OPTera Metro 5200 network sites

Note the different OADM sites and protection at the optical layer can be provisioned for any client signal irrespective of protocol.

9.2 Componentry

In the topics that follow we describe some of the components that make up this product.

9.2.1 OPTera Metro 5300 cabinet

The OPTera Metro 5300 cabinet contains the following:

- One or two OPTera Metro 5200 shelves (depending on site configuration) that support different hardware bands
- One fiber-optic patch panel
- One Ethernet hub (depending on site configuration)
- Optional trunk switches (also known as dual fiber switches)
- One AC power supply with two rectifiers

- Four casters and four leveling pads at the base of the cabinet for easy installation
- Front and rear doors that lock and are removable
- Side panels that are removable
- Fiber-optic cable access on the cabinet roof and base (for raised floor environments)

In Figure 9-3 we show the component details of OPTera Metro 5300 cabinet and OPTera Metro 5200 shelf.

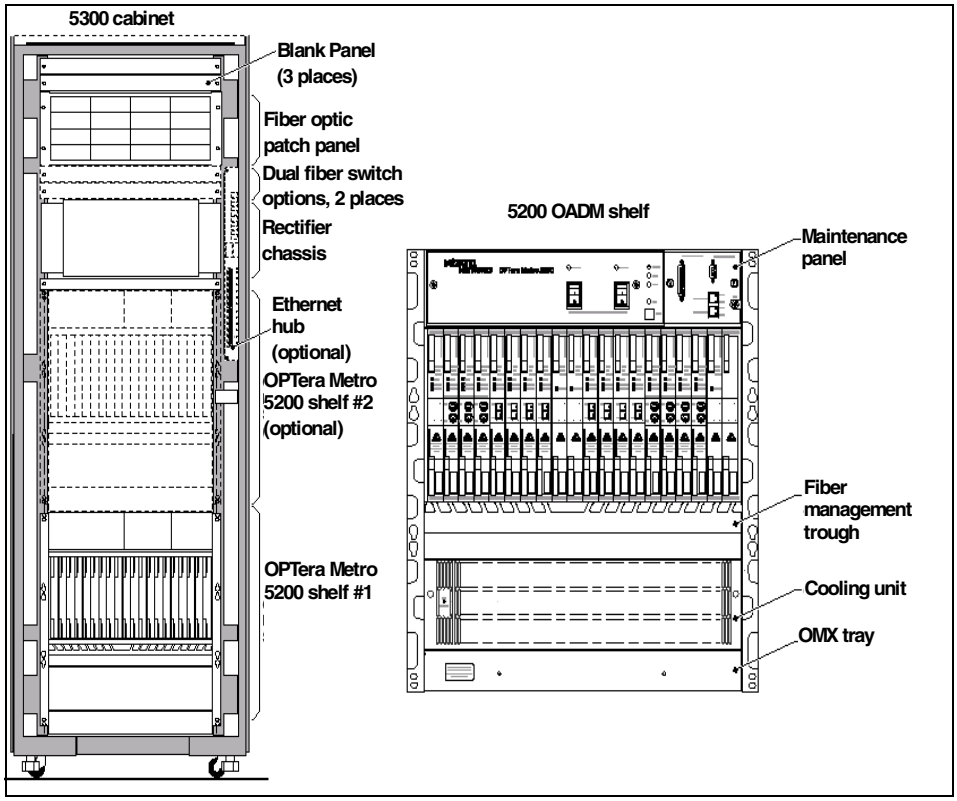


Figure 9-3 OPTera Metro 5300 and OPTera Metro 5200 component layout

9.2.2 OPTera Metro 5200 shelf

There are three types of shelves for the OPTera Metro 5200 system:

- Optical add/drop multiplexer (OADM) shelf
- Terminal shelf
- Optical-fiber amplifier (OFA) shelf

The OADM shelf contains traffic-carrying cards and provides an optical bypass. The terminal shelf also contains traffic-carrying cards, but provides electronic bypass only. The OFA shelf contains optical amplifiers that amplify C-band and L-band traffic.

OADM shelves provide passive optical pass-through that delivers true optical networking and reduces the need for repeated electrical-to-optical conversions.

Terminal shelves are shelves at which wavelength bands terminate in the network. Terminal shelves offer the same functionality as OADM shelves, except that they do not provide optical pass-through.

The OPTera Metro 5200 system is designed to meet the economic and technical requirements of metropolitan applications. The OPTera Metro 5200 can reach the distances required by most of these networks without using optical amplifiers. However, you can install amplifier shelves as the number of nodes in a network increases. OFA shelves reduce the signal degradation that occurs in networks as they expand.

Each OPTera Metro 5200 shelf has the following common equipment:

- One shelf processor (SP) circuit pack
- Two optical channel manager (OCM) circuit packs
- Optical channel interface (OCI) circuit packs with ESCON and ISC connectors (depending on the OADM or OFA shelf)
- Optical channel laser and detector (OLCD) circuit packs (depending on the OADM or OFA shelf)
- Optical multiplexers (OMX) for OADM shelves or equalizer coupler trays (ECT) for OFA shelves
- One cooling unit for every shelf

A fully configured system can transport up to 32 protected or 64 unprotected channels over each pair of optical fibers.

9.2.3 OPTera Metro 5200 shelf card cage

For the OPTera Metro 5200 shelf assembly (Standard 12 U high), Figure 9-4 shows the slot numbers and circuit packs in the card cage of the OADM or terminal shelf.

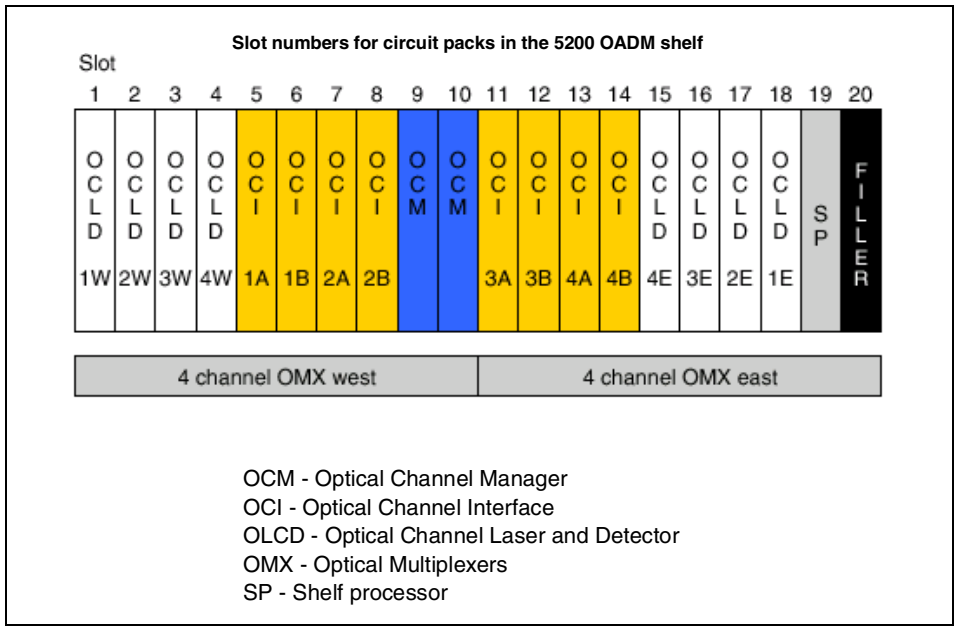


Figure 9-4 Slot numbers of OPTera Metro 5200OADF shelf

The shelf cover protects 20 slots that hold the following:

- Optical channel interface (OCI) circuit packs
- Optical channel laser and detector (OCLD) circuit packs
- Optical channel manager (OCM) circuit packs
- Optical Multiplexer modules (OMX)
- Shelf processor (SP) circuit pack
- Optical Supervisory Channel (OSC) (optional)

OCI circuit packs connect to customer equipment. The client signal travels across the backplane through the OCM and OCLD circuit packs to the multiplexer unit where it is sent to the network elements.

In Figure 9-5 we show the circuit pack interaction.

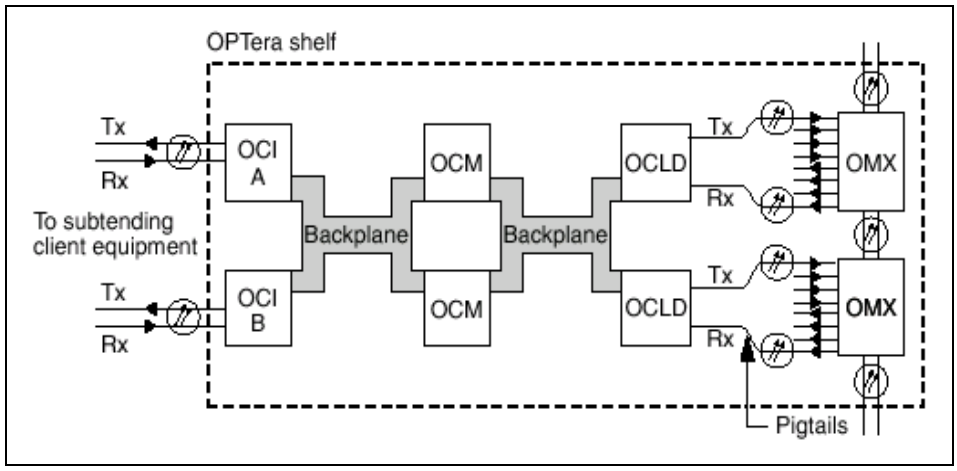


Figure 9-5 OPTera 5200 OADM circuit pack interaction

In the OPTera Metro 5200 Shelf Assembly (Standard 12 U high) shelf, the OMX -Optical Multiplexer modules are located in the OMX tray at the bottom of the shelf. They connect to the OCLD circuit packs through fiber-optic pigtails that are part of the OMX modules.

9.2.4 Optical channel interface (OCI) circuit pack

The OCI circuit packs provide an interface between subtending client equipment and the OPTera Metro 5200 system. Up to eight OCI circuit packs can be in an OPTera Metro 5200 shelf.

In Figure 9-6 we show the faceplate.

The following types of OCI circuit packs are available:

- OCI 622 Mbit/s 1310 nm - Duplex SC connector
- OCI 1.25 Gbit/s 1310 nm - Duplex SC connector
- OCI OC-48/STM-16 1310 nm - two FC connectors
- OCI 1.25 Gbit/s 850 nm - Duplex SC connector
- OCI ISC 1310 nm - Duplex SC connector
- SONET/SDH OCI - two FC connectors
- SRM OCI 1310 nm - four MT-RJ connectors
- SONET/SDH SRM OCI 1310 nm - four MT-RJ connectors

OCI ISC 1310 nm circuit pack

The OCI ISC 1310 nm circuit pack has these features:

- Supports Fibre Channel 1062 Mbit/s

- Supports open fiber control
- Supports single-mode fiber

9.2.5 Optical channel laser and detector (OCLD) circuit pack

OCLD circuit packs are identified by wavelength band (BAND 1 to BAND 8) and by channel within the wavelength band (CH1 to CH4). Up to eight OCLD circuit packs can be installed in the OPTera Metro 5200 shelf. The OCLD circuit pack does electrical-to-optical and optical-to-electrical conversions on a per-channel basis. The OCLD circuit pack detects optical and electrical performance degradations and failures.

The OCLD circuit packs add a per-wavelength optical service channel that provides supervisory and performance management for the OPTera Metro 5200 network. Overhead information is received and transmitted on the same optical path as the main payload channel but at a much lower bit rate, and it is also an out-of-band communication path.

The OCLD circuit pack has two FC panel connectors on the faceplate which is shown in Figure 9-6.

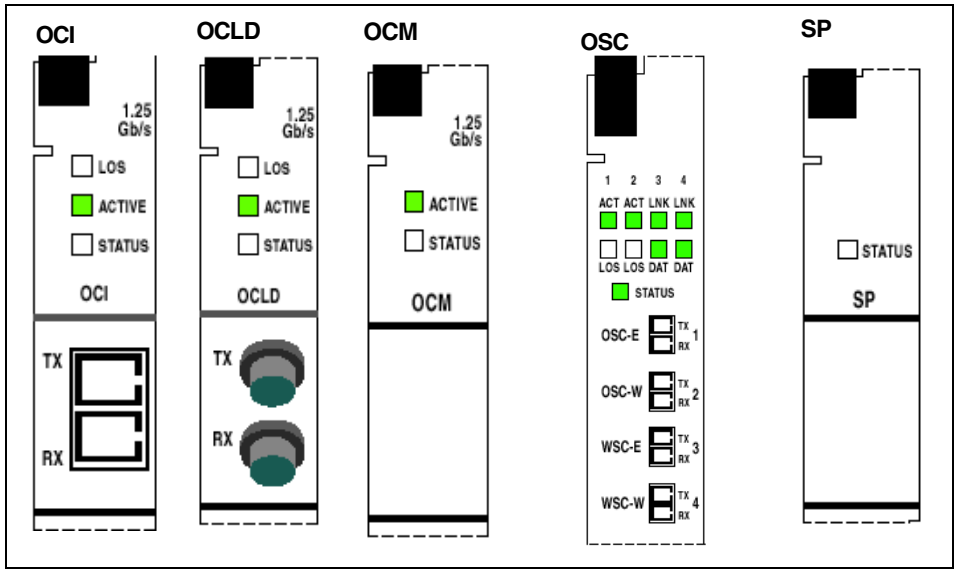


Figure 9-6 OPTera Metro 5200 circuit pack face plates

These connectors are for the single-mode fiber pigtails from the OMX module that carry the channel add and drop signals. The OCI, OCM, and OCLD circuit packs interconnect through the backplane.

9.2.6 OCM circuit pack

The OCM circuit pack does path protection switching in the OPTera Metro 5200 network. It switches paths as the result of optical fiber cuts, shelf failure, or performance degradation. Because the OCM circuit pack does path protection switching at the channel level, other channels in the wavelength band are not disrupted when a switch occurs. There are two OCM circuit packs in the OPTera Metro 5200 shelf.

The OCM circuit pack bridges the data channel between the OCLD and OCI circuit packs. There are no external connectors on the OCM circuit pack and connectivity to the OCI, OCM and OCLD circuit packs is accomplished through the backplane.

9.2.7 Optical Supervisory Channel (OSC)

The optional Optical Supervisory Channel (OSC) circuit pack and OSC module bundle together provide the ability to transmit and receive communication signals over a 1510 nm wavelength. This is along with bundled traffic wavelengths sent from one OPTera Metro 5200 site to the next. The OSC circuit pack is installed in slot number 20 of any OPTera Metro 5200 shelf, and provides the following functions:

- Intersite communication with adjacent sites through the OSC bidirectional ports.
- Intrashelf communication with the SP circuit pack through the backplane connections.

Communication with the customer data communication network and subtending equipment through the WSC bidirectional ports.

The OSC-E and OSC-W ports are connected using SM patch cords to the corresponding OSC module E and W, which is connected to the corresponding OMX module. The wayside supervisory channel (WSC) connector WSC-E and WSE-W are connected to the client equipment using MM patch cords. The OSC circuit pack communicates with the SP circuit pack using the system backplane.

9.2.8 Shelf processor (SP) circuit pack

The shelf processor circuit pack manages communication functions for OPTera Metro 5200. There is one SP circuit pack in an OPTera Metro 5200 shelf; see Figure 9-6.

The SP provides:

- Local management

- Alarm consolidation and telemetry connections
- Software and configuration management
- Shelf visibility
- Performance monitoring
- Inventory control for the shelf
- System communication

9.2.9 Optical multiplexer (OMX) tray

Each OPTera Metro 5200 OADM or terminal shelf has an optical multiplexer (OMX) tray that holds two OMX modules. Each OMX module has one wavelength band filter, and one channel filter. The OMX wavelength band must be the same for both OMX modules installed in the same shelf. The OMX wavelength band also determines the wavelengths of the optical channel laser and detector (OCLD) circuit packs that you install in the shelf.

Each OMX module contains passive optical filters that add and drop up to four channels in the wavelength band assigned to the OPTera Metro 5200 shelf. Other channels pass through the OMX unchanged.

The OMX module in OPTera Metro 5200 has two functional areas:

- Optical add section
- Optical drop section

In Figure 9-7 we show a block diagram of the OMX module.

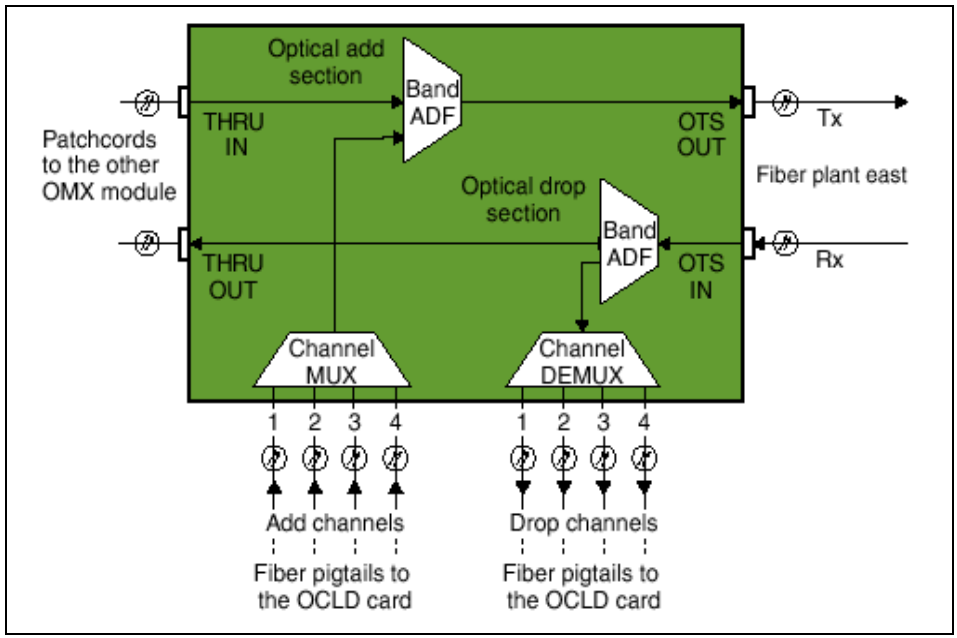


Figure 9-7 OPTera Metro 5200 - OMX module block diagram

The optical add section contains a band filter (ADF) and a channel multiplexer (MUX). The optical drop section contains a band filter and a channel demultiplexer (DEMUX). The ADF drops specific wavelengths while allowing other wavelengths to pass through the filter.

Two OMX modules interconnected in the OMX tray in a single-shelf OADM configuration.

In Figure 9-8 we show the interconnections between two OMX modules (east and west).

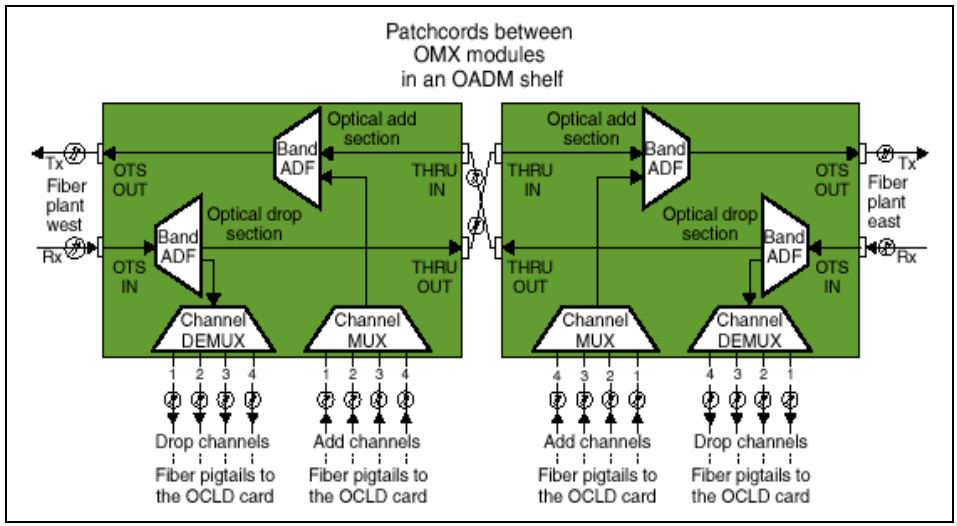


Figure 9-8 OPTera Metro 5200 OMX modules interconnection

The optical transmission signal (OTS) identifies connectors for pigtails or patch cords.

The OMX tray has two yellow indicator lamps. A yellow indicator lamp (normally off) is on the front of each OMX module. If an OMX module loses signal, the SP circuit pack sends a message and the indicator lamp of the module lights.

Each four-channel OMX module has eight optical fiber pigtails to connect the add/drop connections between the OMX module and its four related OCLD circuit packs labeled on the fiber pigtails. In addition each OMX module has a bank of connectors (SC Connectors) for fiber-optic patch cords. The patch cords are used to connect to OMX modules on other shelves at the same site and also to connect to the outside optical fiber plant.

There is an RJ45 connector on the side of each OMX module that connects to the maintenance panel through the OMX cable. This connector provides the electrical and monitoring interfaces to the OMX tray.

9.2.10 Ethernet hub

Every site with more than two shelves requires an Ethernet hub for inter shelf messaging.

In Figure 9-9 we show the Ethernet hub.

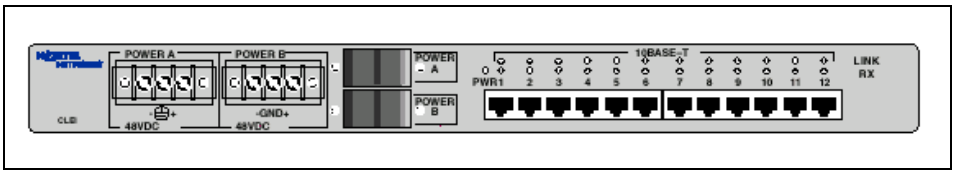


Figure 9-9 OPTera Metro 5300 - Ethernet hub

Intershelf messaging is achieved by connecting the 10Base-T 2X Ethernet port on the maintenance panel of each shelf to the Ethernet hub using shielded cross-over cable.

The Ethernet hub runs on –48 V dc power. If you do not have direct current power available at the site, you must use a rectifier to convert alternating current power to direct current power.

9.2.11 Rectifier chassis

Rectifiers convert alternating current power to direct current power. The rectifier chassis has two bays. Each bay holds one rectifier. Each rectifier has a circuit breaker on the front of the chassis labelled “Rect A” and “Rect B”, respectively. We show this in Figure 9-10.

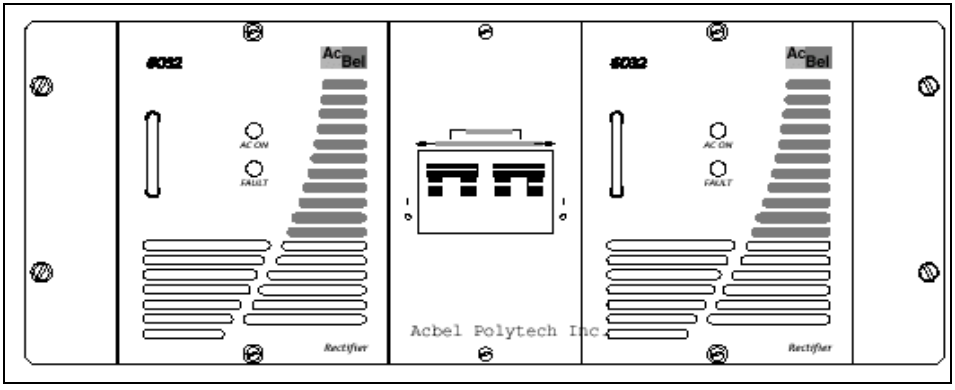


Figure 9-10 OPTera Metro 5300 rectifier chassis

Power connections to the OPTera Metro 5200 shelf (or shelves) are in the back of the rectifier chassis.

The power system operates with two single phase power cords. The cords divide the power equally under normal conditions. If power is interrupted on one of the power cords, the second cord supplies all the necessary power for the cabinet. The rectifier can operate at 110/120 V AC and 220/240 V AC with a single-phase frequency of 50 to 60 Hz.

9.2.12 Maintenance panel

The maintenance panel is located at the top of the OPTera Metro 5200 shelf. The maintenance panel has fault indicators, electrical circuit breakers for redundant power feeds, alarm indicator lamps, alarm cutoff control, and connectors for Ethernet and RS-232 interfaces.

In Figure 9-11 we show the locations of the indicators, switches and connectors.

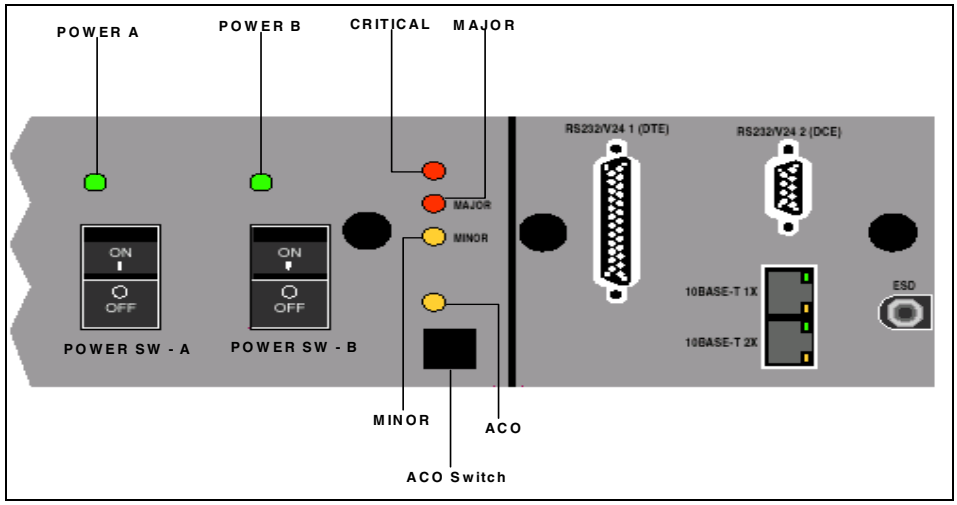


Figure 9-11 OPTera Metro 5200 shelf maintenace panel

There are two power switches on the maintenance panel, Power A and Power B. Each switch has a separate power feed. Indicator lamps above the switches indicate the status of the power feed.

Visual alarm indicators and alarm cutoff (ACO) are to the right of the power switches on the maintenance panel.

9.2.13 Fiber optic patch panel

The fiber-optic patch panel (FPP) supports MT-RJ, ESCON, and SC Duplex connectors. Fiber-optic patch cords connect the front of the OCI circuit packs in an OPTera Metro 5200 shelf to the back of the fiber-optic patch panel. Fiber-optic cables from the customer enter the base of the cabinet and connect to the front of the fiber-optic patch panel.

In Figure 9-12 we show the FPP labeling details.

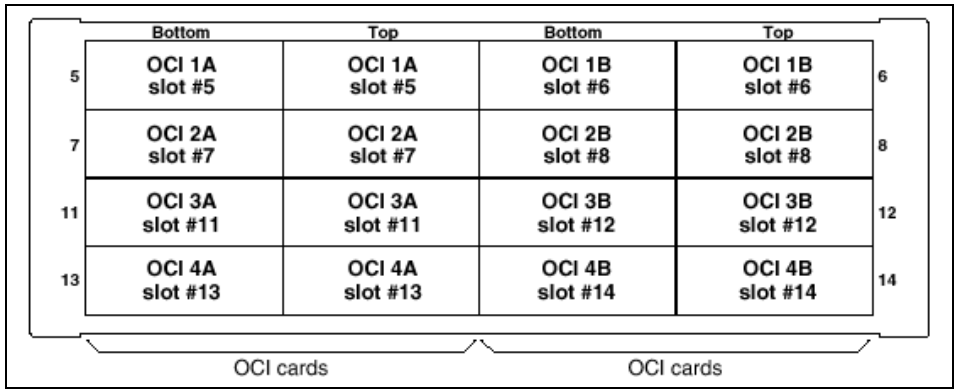


Figure 9-12 OPTera Metro 5200 fiber optic patch panel

The “Bottom” positions on the fiber-optic patch panel correspond to the slots for the OCI circuit packs in the lower OPTera Metro 5200 shelf in the cabinet. If you have a second shelf above the first one, the “Top” positions on the fiber-optic patch panel correspond to the slots for the circuit packs in the upper shelf.

9.2.14 Trunk switch

The optional trunk switch (also known as a dual fiber switch) is an option that offers protection against fiber trunk failures for base channels. We show this in Figure 9-13.

Figure 9-13 Trunk switch front and rear view

Trunk switches are installed in pairs, one at the end of each fiber trunk in a point-to-point unamplified configuration. An extra pair of optical fibers must be available as a backup for each working pair of optical fibers. This means that a two-fiber installation requires four fibers and a four-fiber installation requires eight fibers. The trunk switches attach to the first and last OMX at each site through fiber-optic cables.

If your configuration has only one pair of optical fibers, then the two trunk switches connect to either the first or last OMX at each site. Each trunk switch is 1U high and is located below the fiber-optic patch panel.

The fiber pairs (Tx and Rx) for both the primary path and the standby path are also connected to the front of the trunk switch. The OTS IN and OTS OUT of the OMX are connected to the front of the trunk switch. The back of the switch has a RJ-45 Ethernet port for configuring and operating the trunk switch. The redundant power connectors at the rear are coupled to the power supply.

9.2.15 Fiber management trough

The fiber management trough is located below the card cage. The trough consists of a shelf-width horizontal tray with a comb above it. The fiber management tray holds patch cords and fiber-optic cables that are routed to and from the circuit packs installed in the shelf. The comb above the tray helps distribute and route the patch cords from the tray to the connectors of the circuit packs.

Shelf cover

The OPTera Metro 5200 shelf has a removable door that covers the card cage and fiber management comb and provides EMI shielding.

9.2.16 OADM and mux/demux

Optical Add/Drop Multiplexer (OADM) functionality is provided in the OADM shelf by the two Optical multiplexer (OMX) modules in the OMX tray. Each OMX module contains passive optical filters that add and drop up to four channels in the wavelength band assigned to the OPTera Metro 5200 shelf. Other channels pass through the OMX unchanged.

The OMX module in OPTera Metro 5200 has two functional areas:

- Optical add section
- Optical drop section

The optical add section contains a band filter (ADF) and a channel multiplexer (MUX). The optical drop section contains a band filter and a channel demultiplexer (DEMUX). The ADF drops specific wavelengths while allowing other wavelengths to pass through the filter.

In Figure 9-14 we show a block diagram of the OADM section.

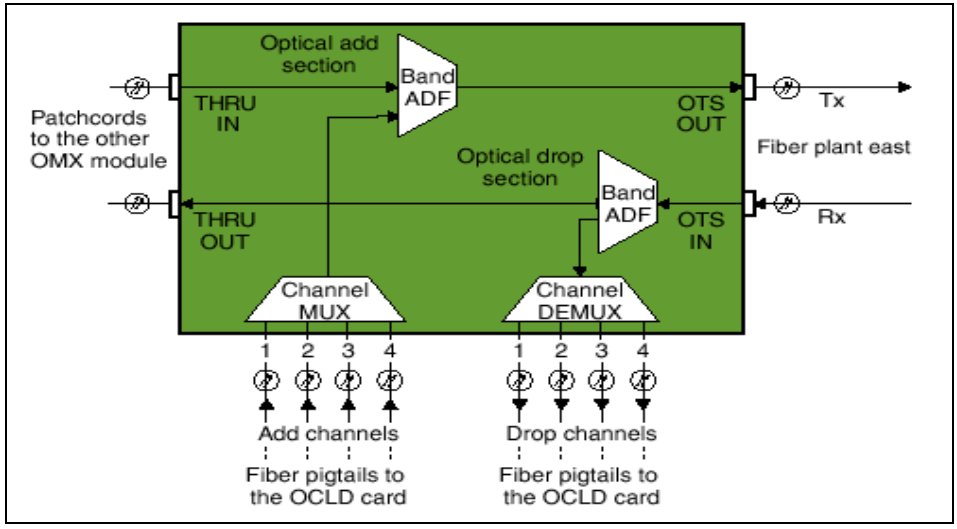


Figure 9-14 Block diagram of OADM section

9.2.17 Optical Amplifier (OA) /Optical-Fiber Amplifier (OFA)

The optical-fiber amplifier (OFA) shelf contains optical amplifiers that amplify C-band and L-band traffic.

OFA shelves do not need to be collocated with terminal shelves or OADM shelves. If you provide independent Ethernet communications, you can install OFA shelves in intermediate sites; these sites are called OFA sites.

OFA shelves can be deployed in terminal sites or OADM sites as needed to overcome system losses. These shelves are provisioned as OFA shelves on the system manager.

In Figure 9-15 we show the OFA shelf layout.

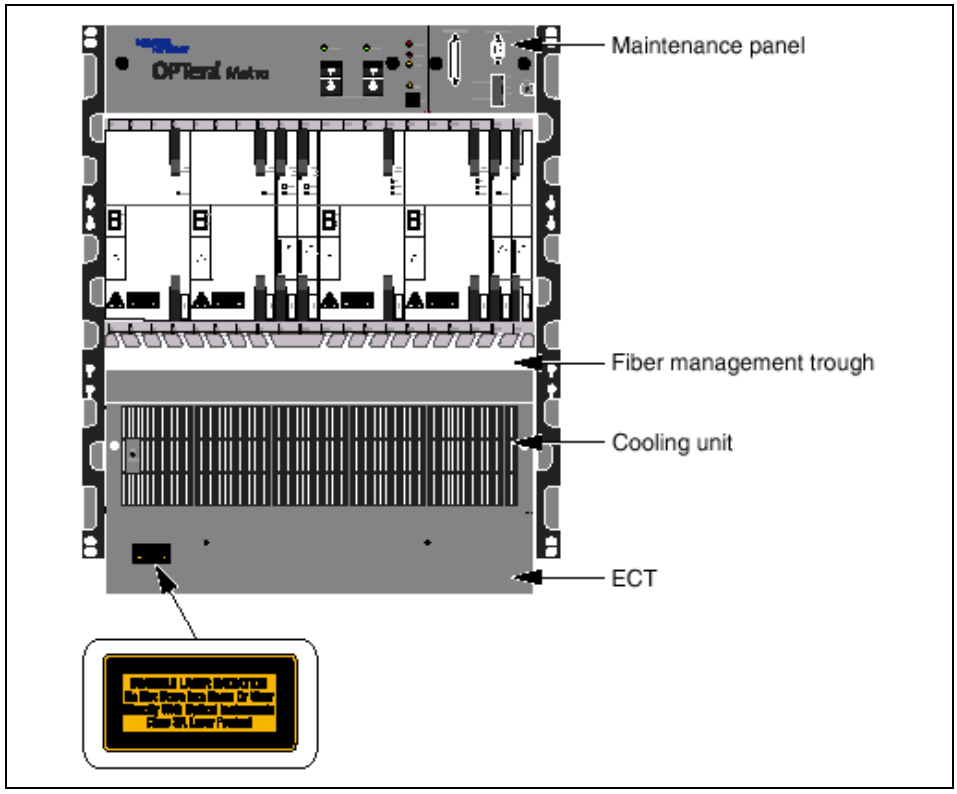


Figure 9-15 OPTera Metro 5200 OFA shelf layout

The 20 slots OFA shelf holds the following:

- OFA circuit packs
- OCM circuit packs
- SP circuit pack

In Figure 9-16 we show the slot details of circuit packs in the OFA shelf.

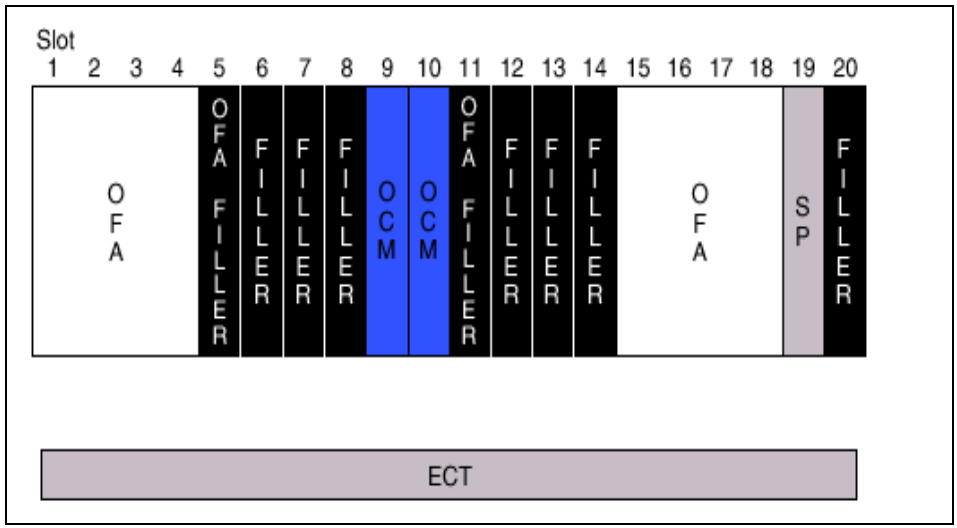


Figure 9-16 OPTera Metro 5200 OFA circuit pack locations

9.2.18 OFA circuit pack

The OFA circuit pack uses an erbium doped fiber amplifier (EDFA) to amplify C- and L-band signals. The OFA circuit pack reduces the signal degradation that occurs when you increase the number of nodes in the network. Two types of OFA circuit packs are available: the OFA C-band and the OFA L-band. Each circuit pack amplifies four bands with four wavelengths, for a total of 16 wavelengths per circuit pack. A maximum of two OFA C-band and two OFA L-band circuit packs can be in an amplifier shelf. Each OFA circuit pack uses four slots in the shelf.

9.2.19 Equalizer coupler tray

The equalizer coupler tray (ECT) installs in the OFA shelf. The ECT separates, equalizes, and combines conventional and long band traffic (C-band and L-band). If the OPTera Metro 5200 system is transmitting C- and L-band traffic on the same optical fiber, the ECT can separate the bands. The ECT then transmits the bands to the OFA circuit packs for amplification, and then combines the bands again. When the ECT transmits the combined bands to OADM nodes, each band is within the required power level. The equalizer also removes power differences in the bands before transmitting traffic to the amplifier.

The maintenance panel, shelf processor (SP), fiber management trough, and cooling unit are common on all shelves of OPTera Metro 5200 and have been described previously.

9.3 Supported topologies

OPTera Metro 5200 supports ring and linear topologies.

9.3.1 Ring topologies

The OPTera Metro 5200 supports the following ring configurations:

- Hubbed-ring
- Dual hubbed-ring
- Meshed-ring

Hubbed-ring configuration

The hubbed-ring configuration is used for traffic flows that are characteristic of access networks. In a hubbed-ring configuration, shelves in the OADM sites have a one-to-one correspondence with the shelves in the terminal site. You can connect a maximum of eight OADM shelves to the terminal site through one pair of optical fibers.

Dual hubbed-ring configuration

A dual hubbed-ring has two terminal sites. Shelves in the OADM sites have a one-to-one correspondence with the shelves in each of the terminal sites. This configuration allows a band to pass-through one terminal site electronically and end at the other terminal site.

Meshed-ring configuration

The meshed-ring configuration is used for traffic flows that are characteristic of interoffice networks. The meshed-ring configuration supports band meshing and channel meshing.

- Band meshing is the ability to add or drop all the wavelengths of a band at one or more nodes in the network.
- Channel meshing is the ability to add or drop any channel from one node at one or more other nodes in the network. Channel meshing is similar to band meshing, except that the entire band does not have to terminate.

In a meshed-ring configuration, you can have, but do not need, a terminal site.

9.3.2 Linear topologies

The OPTera Metro 5200 supports the following linear configurations:

- Point-to-point
- Linear OADM

Point-to-point configuration

In a point-to-point configuration, east terminal shelves have a one-to-one correspondence with west terminal shelves. You can provision a maximum of eight OPTera Metro 5200 shelves at each terminal site.

Linear OADM configuration

A linear OADM configuration has two terminating sites at each end and sites with OADM shelves in the center. This configuration can add or drop a signal at any of the sites within the same band.

The channel assignments provisioned in a linear OADM configuration are unprotected, because there is one path between the network elements.

9.4 OADM regenerator application

The regenerator shelf is a standard OPTera Metro 5200 OADM shelf configured as an electrical pass-through shelf. The shelf performs an optical-electrical-optical (O-E-O) conversion and extends the reach of the optical signal between two sites.

You can install a regenerator shelf where long optical fiber spans exceed the dispersion limit. The shelf regenerates the pulse, shape, and amplitude of the signal and resets the dispersion limit for an optical span.

9.4.1 33rd lambda wavelength

The Optical Supervisory Channel also referred to as the 33rd λ wavelength, enables the following functions:

- Fault isolation and link integrity
- Node visibility in meshed networks
- Multiple communication paths
- Remote access to OFA sites
- Transparent wayside channel
- Optical power equalization during initial deployment
- Monitoring capabilities

The OSC circuit pack transmits and receives network control information over an out-of-band wavelength (1510 nm) along with bundled traffic wavelengths from one OPTera metro site to the next, including OFA sites. The OSC circuit pack does not interfere with the traffic on the system.

The OSC circuit pack and OSC module bundle together to provide the ability to transmit and receive communication signals over a 1510 nm wavelength. This is along with bundled traffic wavelengths sent from one OPTera Metro 5200 site to the next. The OSC circuit pack is installed in slot number 20 of any OPTera Metro 5200 shelf, and provides the following functions:

- Intersite communication with adjacent sites through the OSC bidirectional ports.
- Intrashelf communication with the SP circuit pack through the backplane connections.
- Communication with the customer data communication network and subtending equipment through the WSC bidirectional ports.

9.5 Distance

The supported distance or fiber length in an OPTera Metro 5200 network is determined by a number of factors including:

- The configuration type (hubbed-ring or point-to-point)
- Dual Fiber Switch capability
- The number of shelves in the configuration
- Attenuation (dB loss) on fibers, splices and connectors throughout the OPTera Metro 5200 network

The maximum supported fiber pair length for all channel types in a point-to-point configuration is 50 km between the two sites. In a point-to-point configuration with Dual Fiber Switch (DFS), the maximum supported fiber length is 40 km. In hubbed-ring configurations, which can have up to eight locations, the maximum distance from the terminal site (hub site) to the farthest OADF site (remote site) is 35 km.

It is possible to cascade up to four networks, by connecting shelf channels from one OPTera Metro 5200 network to shelf channels on another OPTera Metro 5200 network. By cascading four point-to-point configurations, the maximum end-to-end fiber pair length is 200 km. Even in a cascaded configuration the maximum fiber length for a channel is still governed by the specifications of the attached device.

9.5.1 Loss/link/light budget

The optical link budget specifies the maximum loss supported for a connection between the point where it originates and the point where it terminates. Since there is no pre-defined configuration with these topologies and the order in which the shelves are connected together, the link budgets are calculated by adding the loss for each individual fiber section between the two end points of a connection. The loss must be calculated for each band, because the various bands are subject to different attenuation depending on the path and the number of network elements that it passes through. Link budget losses can be classified into two categories:

- Unamplified network link budget
- Amplified network link budget

The examples in this section refer to only the unamplified link budget. The link budgets for unamplified networks are based on per-band power calculations.

For example, in a hubbed ring topology, the optical link budget is calculated for the worst case optical path “east to west” and “west to east”.

In Table 9-1, we show the max loss (dB) for hubbed ring configurations and point-to-point configurations using all fiber types up to 40 km.

Table 9-1 Link budget for hubbed ring and point-to-point configurations

Hubbed ring configurations - all fiber types up to 40 km; up to 1.25 Gbits/s	
Number of OADF shelves	**Max Loss (dB)**
1	**21.7**
2	**19.9**
3	**18.3**
4	**16.8**
5	**15.2**
6	**13.7**
7	**12.1**
8	**10.6**
Point-to-point configurations - all fiber types up to 40 km; up to 1.25 Gbits/s	
Number of OADF shelves	**Max Loss (dB)**
1	21.7

Hubbed ring configurations - all fiber types up to 40 km; up to 1.25 Gbits/s	
Number of OADF shelves	**Max Loss (dB)**
1	**21.7**
2	**19.9**
3	**18.3**
4	**16.8**
5	**15.2**
6	**13.7**
7	**12.1**
8	**10.6**
Point-to-point configurations - all fiber types up to 40 km; up to 1.25 Gbits/s	
Number of OADF shelves	**Max Loss (dB)**
2	20.6
3	19.6
4	18.7
5	17.8
6	16.9
7	15.9
8	15.0

In addition to the link budget, connector loss or repair margin is also to be added. Link budgets are often specified between operating temperatures of 0 to 40 degrees centigrade. Link budget calculations are very important, often complex, and require a specialist's work.

9.6 Form factor, footprint and mounting

The OPTera Metro 5300 is a cabinet of 1795mm*648mm*686mm (H*W*D). We show this in Figure 9-17.

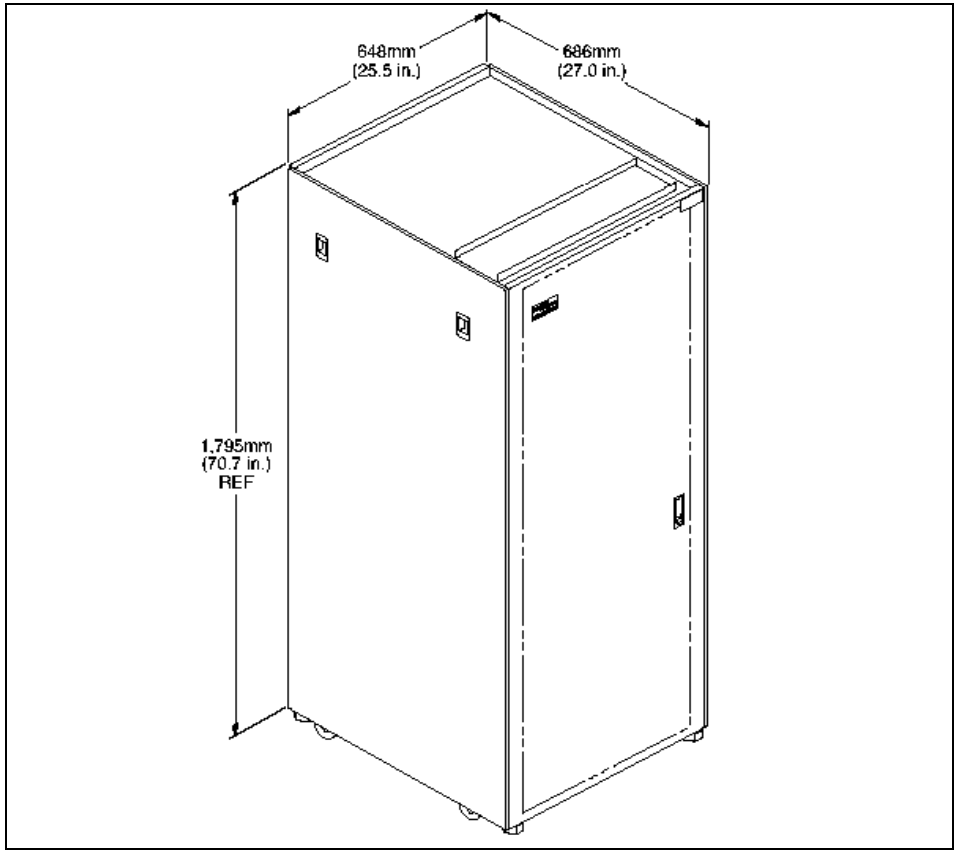


Figure 9-17 OPTera Metro 5300 cabinet dimensions

The weight of the cabinet depends on the number of OPTera Metro 5200s inside the cabinet. With one OPTera Metro 5200, it is 278 kg and with two OPTera Metro 5200s it is 321 kg.

The OPTera Metro 5200 Shelf Assembly is standard 12 U high, and each OPTera Metro 5300 can have two OPTera Metro 5200s.

There are two fans in the cooling unit of an OPTera Metro 5200 shelf that draw air through front of the shelf to cool the OPTera Metro 5200 circuit packs. The air exits through the top back of the shelf.

The OPTera Metro 5200 shelf general specifications are:

Dimensions

- Height 535 mm (21 in.) (12 U)
- Width 438 mm (17.25 in.)

- Depth 300 mm (11.85 in.)

Weight

- Fully loaded 43 kg (94 lb.)
- Empty chassis (shipped with maintenance panel, cooling unit, and omx tray installed): 27 kg (60 lb.)

Power requirements

- Nominal –48 V DC
- Minimum –38 V DC
- Maximum –72 V DC

The typical power dissipation of a fully loaded shelf is 524 Watts and the maximum power dissipation of a fully loaded shelf 675 Watts.

9.7 Management

You can manage the OPTera Metro 5200 network with the following software:

- System Manager and a Web browser
- Preside
- Preside Manager for OPTera Metro and an Ethernet connection
- TL1 and a Telnet connection
- SNMP for alarm surveillance

9.7.1 System manager overview

The OPTera Metro 5200 system manager is a Web-based graphical user interface (GUI) that allows you to access one or more shelves in an OPTera Metro 5200 network. You will use the system manager to provision, monitor, and maintain the network.

You can perform the following functions through the system manager:

- Equipment and facility configuration
- Alarm management
- Connections management
- Protection management
- Software download and upgrade
- Performance monitoring
- Event history review
- Login, security, and user administration

The system manager is accessible from any shelf in the network through the 10Base-T 1X port on the maintenance panel or remotely via a modem dial-up connection, using a RS232/V24 1 (DTE) interface.

A system manager applet is stored on the shelf processor (SP) circuit pack of each shelf. Logging on to the shelf enables the system manager. We recommend that you have a Windows NT system for system manager.

9.7.2 Preside manager for OPTera Metro 5200

The Preside manager for OPTera Metro 5200 is an optional network management system that you can use with your OPTera Metro 5200 network. The Preside manager provides a graphical user interface (GUI) and network-wide management capabilities as follows:

- Enrolls and de-enrolls network elements
- Detects OPTera Metro 5200 network elements in an OPTera Metro 5200 network
- Detects changes in network element attributes, such as name or IP address
- Provides communication between the OPTera Metro 5200 system manager and the Preside Graphical Network Browser (GNB)
- Security and session management
- Fault management
- Connection management
- Performance management

9.8 Availability and reliability

The OPTera Metro 5200 is designed for high availability, having redundant components and internal cross-connections. When using high availability channels, OCIs send signals to both shelf sides through cross-connected OCMs and OCLDs, reaching both OMXs to be multiplexed and sent down both fiber pairs.

The Dual Fiber Switch feature offers an intermediate, lower-cost availability option. It provides redundancy in the event of a fiber break but it does not provide the same level of hardware component redundancy as high availability channels.

The use of passive OMXs ensures traffic continuity through a shelf even in the event of a shelf or frame power failure: the other shelves can still send and receive their channel signals through the OPTera Metro 5200.

Each channel is equipped with in-band monitoring, which continually reports channel status. Whether the implementation is a point-to-point or hubbed-ring configuration, it can be managed from a single location with a PC.

It provides concurrent maintenance, because all cards and service elements are hot-pluggable. No regular maintenance is required except for periodically cleaning the air filter on the fan tray.

The OPTera Metro 5200 architecture provides data survivability for all types of traffic. The OPTera Metro 5200 network supports a maximum of:

- 40 optical add/drop multiplexer (OADM) or terminal shelves
- 24 optical-fiber amplifier (OFA) shelves
- 16 sites for ring topologies with OSC
- Nine sites for linear topologies using the same band at each site, without OSC
- Nine sites for unprotected ring topologies using the same band at each site, without OSC

9.8.1 Redundancy

OPTera Metro 5200 can be configured to enable path switching and equipment switching, this ensures that facility or equipment failures do not affect payload traffic.

Path switching

Path switching provides end-to-end protection for the signal carried between two OCI circuit packs in an OPTera Metro 5200 network. In a protected path, an OCI circuit pack at the local site shelf sends signals to and from an east and west OCLD circuit pack. The OCI at the remote site shelf selects which of the two paths to use. Path switching protects the channel assignments on the OCLD circuit packs; therefore, protection is provided on a per-path rather than on a per-shelf basis. Each direction of the channel is protected (unidirectional switching).

Note: Path protection does not exist on linear OADM configurations.

Equipment protection switching

Equipment switching uses redundant OCM circuit packs to protect all the channel assignments on a shelf. In an OPTera Metro 5200 system, there are two OCM circuit packs in each shelf. Each OCM circuit pack carries part of the traffic in a shelf. If an OCM fails or if you take an OCM out-of-service, the OCM sends a message to the other OCM circuit packs in the shelf to indicate that it is not available. The other circuit packs in the shelf automatically switch to the redundant OCM circuit pack.

Switching time

Protection switching is completed within 50 milliseconds (ms) following a 10 ms detect interval for both path protection and equipment protection.

9.8.2 Scalability

The OPTera Metro 5200 network supports a maximum of:

- 40 optical add/drop multiplexer (OADM) or terminal shelves
- 24 optical-fiber amplifier (OFA) shelves
- 16 sites for ring topologies with OSC
- Nine sites for linear topologies using the same band at each site, without OSC
- Nine sites for unprotected ring topologies using the same band at each site, without OSC

9.8.3 Serviceability

OPTera Metro 5200 allows hot insertion of the following circuit cards in an in-service system:

- OCI
- OCLD
- OCM
- SP
- OFA
- SRM

In fully protected channel configurations, replacing the circuit cards does not affect service, but with unprotected channel configurations, circuit card replacement, except for the OCM and SP cards, affect the service.

The OMX module (one or both) can be replaced, however this affects the service and will call for planned downtime.

9.8.4 Security

The OPTera Metro 5200 is protocol independent and does not view the data being transported. It does no error checking or correction on the data.

Access to the System Manager software is controlled by user-supplied passwords for different access levels. Access may be as:

- Administrator (can view and change the OPTera Metro 5200 configuration)
- Operator (can view the OPTera Metro 5200 configuration and status)
- Observer (can view the status only)

A maximum of four users can be logged on to a OPTera Metro 5200 shelf at the same time.

9.8.5 Interoperability

DWDM solutions are typically not interoperable, though they would follow the ITU-T guidelines. For example, the CISCO 15540 operates between the ITU_T recommended wavelengths of 1533.33 nm and 1562.23 nm, and has a channel spacing of 100 GHz. The Nortel OPTera Metro 5200 operates between the ITU_T recommended wavelengths of 1528.77 nm and 1605.73 nm, and has a channel spacing of 200 GHz.

However, different client equipment, like FC Switches, FC Directors, and Storage can be connected to the same DWDM equipment.

9.8.6 Connectivity

OPTera Metro 5200 connects to client equipment on one end and to the transport network on the other end. The client signal is converted into standards based wavelengths and then sent over a single fibre link to the other site. At the remote site the standard wavelength signal is converted into client signal. This is an O-E-O-E-O (Optical to Electrical to Optical to Electrical to Optical) operation in the DWDM equipment between the client equipment at two sites that are connected using this DWDM equipment.

9.8.7 OADM regenerator application

The regenerator shelf is a standard OPTera Metro 5200 OADM shelf configured as an electrical pass-through shelf. The shelf performs an optical-electrical-optical (O-E-O) conversion and extends the reach of the optical signal between two sites.

You can install a regenerator shelf where long optical fiber spans exceed the dispersion limit. The shelf regenerates the pulse, shape, and amplitude of the signal and resets the dispersion limit for an optical span.

9.8.8 Bandwidth, channels and channel spacing

OPTera Metro 5200 channels meet ITU-T recommendations. Each wavelength band is made up of four channels with 200 GHz spacing.

In Figure 9-18 we show that the system supports wavelengths between 1528 and 1606 nm (centered at 1550 nm) that have 200 GHz channel spacing. The 32 wavelengths are divided into eight bands of four channels, all of which are transmitted over a single optical fiber and can be managed separately.

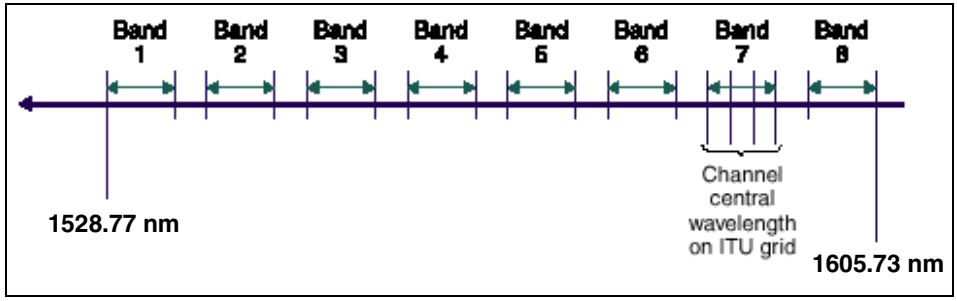


Figure 9-18 OPTera Metro 5200 wavelength bands and channels

A shelf has one wavelength band. A completely configured OPTera Metro 5200 system (16 shelves) can transport up to 32 protected or 64 unprotected channels over each pair of optical fibers. Each channel normally operates from 50 Mbit/s to 2.5 Gbit/s, but can support 16 Mbit/s traffic. This configuration allows for a maximum transport capacity of 80 Gbit/s protected or 160 Gbit/s unprotected.

9.8.9 Protocols

OPTera Metro 5200, which delivers network scalability, per wavelength manageability, and protocol and bit-rate independence, supports the following signals:

- SONET OC-1/OC-3/OC-12/OC-48/OC-192
- SDH STM-1/4/16/64
- Async Fiber Optic Systems (FOTS)/PDH
- Gigabit Ethernet, 10 GbE
- ESCON, FICON
- Fibre Channel
- HPPI

- D1 Video
- 2.5G, 10G POS

However, in this chapter the overview has been limited to Fibre Channel (FC) and SAN.

9.8.10 Power requirements

OPTera Metro 5300 requires a dual-source grounded AC outlet within 1.5 m (5 ft.) of the cabinet. The power requirement is 110/120 V AC or 220/240 V AC, depending on the country. And in addition the site should have provision for two additional AC outlets (three-prong grounded with 20 Amp circuit breakers) for the system manager computer.

9.8.11 Interconnect cables

Interconnect cables are used for:

- Intershelf messaging and local connection to a system manager computer
- Interconnecting OMX modules between OPTera Metro 5200 shelves

The types of interconnect cables are:

- **Ethernet cables**
  - Straight-through cables—used to connect an OPTera Metro 5200 shelf to the system manager computer
  - Cross-over cables—used for inter shelf messaging
- **Fiber-optic patch cords**

  Fiber-optic patch cords are used to interconnect OMX modules between site shelves so that signals can flow through the backbone of the network. In each pair of patch cords, one patch cord has gray boots and the other has black boots. The different colors are to help you connect pairs of ports on two shelves at the same time without crossing the signal paths.

9.8.12 Microcode and firmware

The OPTera Metro 5200 has software tools options. The software right-to-use licenses (RTU) has to be ordered for the function. The current software release is Version 3.2. Some examples of the RTU are:

- NE Base Release 3.2
- System Manager Protection Switching (Line)
- SNMP Northbound Interface
- Wavelength meshing RTU

- TL1 Northbound IF RTU
- Up to 2.5 Gbit/s connections
- Software Delivery Kit (CD-ROM)

9.9 General specifications

The OPTera Metro 5300 is a cabinet of 1795mm*648mm*686mm (H*W*D). This is shown in Figure 9-17. The OPTera Metro 5200 shelf general specifications are as follows:

Dimensions
- Height 535 mm (21 in.) (12 U)
- Width 438 mm (17.25 in.)
- Depth 300 mm (11.85 in.)

Weight
- Weight: fully loaded 43 kg (94 lb.)
- Weight: empty chassis (shipped with maintenance panel, cooling unit, and omx tray installed): 27 kg (60 lb.)

Power requirements
- Nominal –48 V DC
- Minimum –38 V DC
- Maximum –72 V DC

Power dissipation of a fully loaded shelf
- Typical power 524 Watts
- Maximum power 675 Watts

Temperature limits
- Operating: 0 to 55 °C (32 to 131 °F)
- Maximum rate of change: 8.3 °C (46.94 °F) per hour
- Shipping and Storage: –40 to 66 °C (–40 to 150 °F)

Other environmental conditions
- Relative humidity (non-condensing) limits: operating and storage 5 to 95%
- Thermal loading: room ambient conditions 3444 W/m^2
- Floor loading: 732 kg/m^2 (150 lbs/ft^2)
- Operating altitude: 80 to 1800 meters (260 to 5900 ft.) above sea level
- Earthquake Zone: 4
- Dust conditions: A maximum of 100,000 dust particles 0.5 microns/$ft.^3$ (MIL-STD-801D)

9.9.1 Standards compliance

OPTera Metro 5300 cabinet meets with the European Conformity and carries the “CE” mark as an indication of conformity.

In North America, the cabinet meets with the Underwriters’ Laboratories (UL), the Canadian Standards Association (CSA), and Telcordia specifications. The cabinet complies with Telcordia GR-63, GR-487, and GR-1089.

The system limits radio frequency (RF) emissions to meet the Federal Communications Commission (FCC) Class A requirements.

The OPTera Metro 5200 system is European Telecommunications Standards Institute (ETSI) and Network Equipment Building System (NEBS) compliant.

OPTera Metro 5200 channels meet ITU-T recommendations. Each wavelength band is made up of four channels with 200 GHz spacing.

OPTera Metro 5200 CLASS 3A Laser Product.

10

Sorrento Networks GigaMux and EPC

Sorrento Network's GigaMux DWDM Fiber Optic Transmission System multiplexer provides carrier-class dense wave division multiplexing to extend Storage Area Networks or Ethernet/IP-based LANs.

The GigaMux delivers the gigabit speeds, high availability, scalability, and multi-location networking and management that business critical applications demand. It accomplishes this by turning one optical fiber into as many as 64 virtual fibers, or 32 full duplex channels. These channels can then simultaneously transport multiple independent applications.

The GigaMux is designed for flexibility. It can be deployed in numerous configurations. In addition, providing a mix-and-match capability on a per channel basis, the GigaMux can offer support for a wide range of applications (up to 10 Gbps per channel) and network topologies. Fast and Gigabit Ethernet, ESCON, Fibre Channel, FICON, FDDI, SONET/SDH signals from OC-3/STM-1 to OC-192/STM-64, and Coupling Link can be mixed in one box, all with full performance and economy.

In the topics that follow we introduce the GigaMux and some of its important features.

10.1 Product overview

Using a pair of optical fibers, GigaMux can handle 64 duplex channels with aggregate bandwidth of up to 640 Gbps. With its sub-rate multiplexing feature, the duplex channel capacity increases to 512, and with Sorrento's new 16-port SONET/SDH EPC, the maximum full duplex capacity increases from 512 channels to 1024 channels.

The GigaMux system is a Fiber Optic Transmission System that transforms an existing fiber plant from a dedicated single-application medium to a versatile conduit capable of simultaneously transporting up to 64 independent applications (channels).

We show a picture of the GigaMux rack populated with five equipment shelves in Figure 10-1, and also a single shelf which is all that is needed to get started.

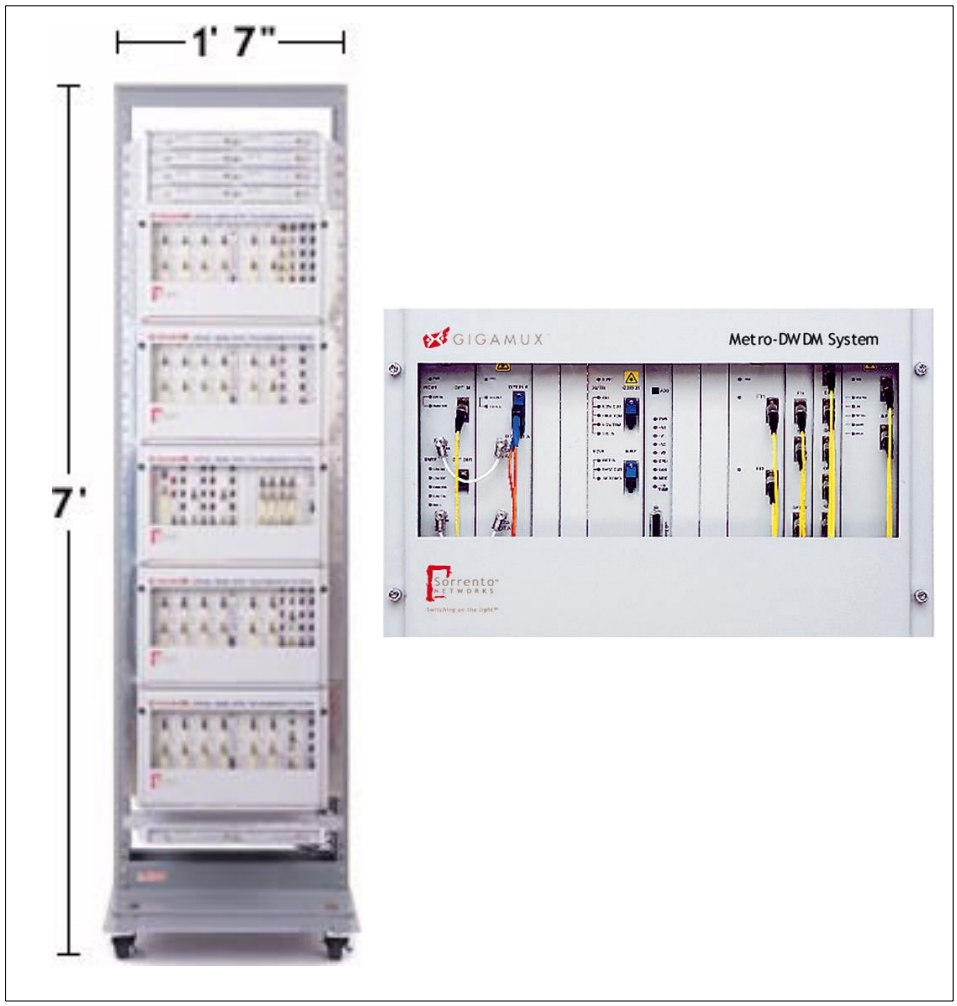


Figure 10-1 GigaMux rack — showing its dimensions — and shelf

Using Dense Wave Division Multiplexing (DWDM), the GigaMux System offers breakthrough technology that enables you to transition to an all optical network. The GigaMux system supports up to 10 Gbps of traffic per channel. Each of the GigaMux's 64 channels provides a fully independent optical pathway so even dissimilar data types can easily share the same fiber.

The GigaMux System prepares an existing fiber plant to meet growing transmission needs effortlessly and its modular design provides you with the flexibility to meet emerging needs. It meets industry standards such as Network Equipment Building Systems (NEBS) standards and the ITU-T channel spacing recommendations.

In Figure 10-2 we show the GigaMux application difference.

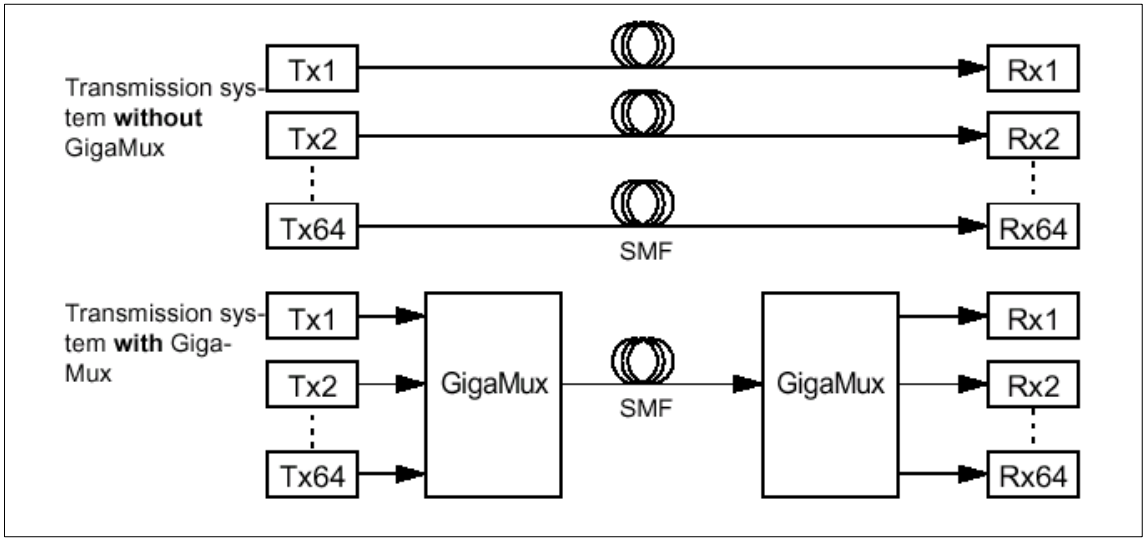


Figure 10-2 GigaMux application

10.2 Architecture

GigaMux is a modular system, allowing you to only buy what you need today. As requirements grow, the GigaMux System is simply expanded to meet those needs.

Each GigaMux System is composed of basic building block modules. A fully functional GigaMux system can start with one active channel and later be expanded without ever interrupting existing channel traffic. Depending on your application, adding GMEx, GMEI/GMDE, GMLEI, or GMFE expansion modules allows your system to grow without interrupting services.

The GMWD module, also referred to as Sidestreet Channel (refer to 10.6, “Sidestreet channel” on page 292 for more information), allows you to integrate GigaMux DWDM signals with existing fibers carrying 1310 nm legacy signals. This integration allows you to upgrade your system without the cost of purchasing all new equipment.

10.2.1 Topologies

GigaMux supports all types of topologies of the most commonly designed linear and ring configurations.

These are the supported linear configurations:

- Point-to-Point
- Linear Add/Drop

These are the supported ring configurations:

- Ring Add/Drop
- Star
- Mesh

10.2.2 Point-to-point

A point-to-point configuration interconnects two terminal points. In a point-to-point system the east terminal GigaMux connects to the west terminal GigaMux.

For 32 channel systems Sorrento Networks supports a maximum of eight spans (nine nodes), where each span has a loss no greater than 23 dB. If the fiber loss is approximately 0.25 dB/km and there is approximately 2 dB loss in the node connection for each span, the effective transmission distance for a point-to-point system exceeds 600 km.

We show a point-to-point configuration in Figure 10-3.

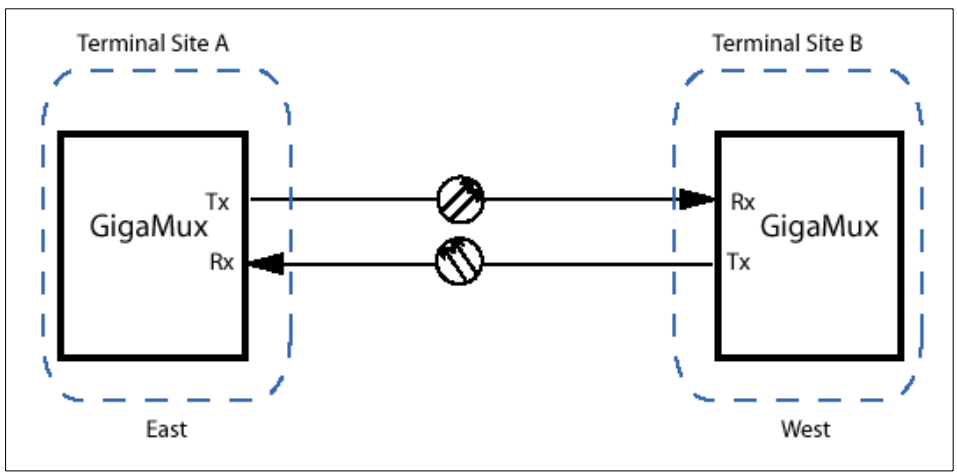


Figure 10-3 GigaMux point-to-point

10.2.3 Linear add/drop

GigaMux linear add/drop systems can be designed with one or more add/drop sites and two terminal sites. Up to 64 channels can be added or dropped. 32 unprotected channels are supported in single fiber bidirectional system.

Sorrento Networks supports full add/drop in a maximum of nine nodes (8 spans, 32 channels), where each span does not have a loss greater than 23 dB. The effective transmission distance in an add/drop network exceeds 600 km.

We show linear add/drop in Figure 10-4.

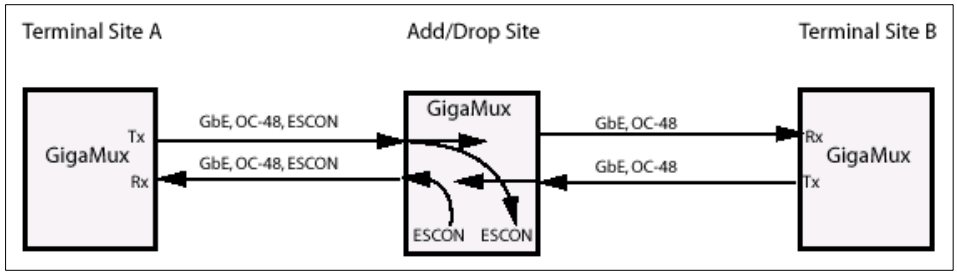


Figure 10-4 Linear add/drop

10.2.4 Ring system

A GigaMux ring system consists of three or more sites, where at least one is an add/drop site. A ring topology system offers a high level of protection for your system. When a failure occurs traffic is re-routed via an adjacent site.

Sorrento Networks supports 32 unprotected channels in single fiber, bi-directional ring system. For a 32 channels system, Sorrento Networks supports a maximum of nine nodes in a ring system. Each node is capable of add/drop of all channels. The maximum size of a ring system is 400 km.

As shown in Figure 10-5, these sites form a closed ring.

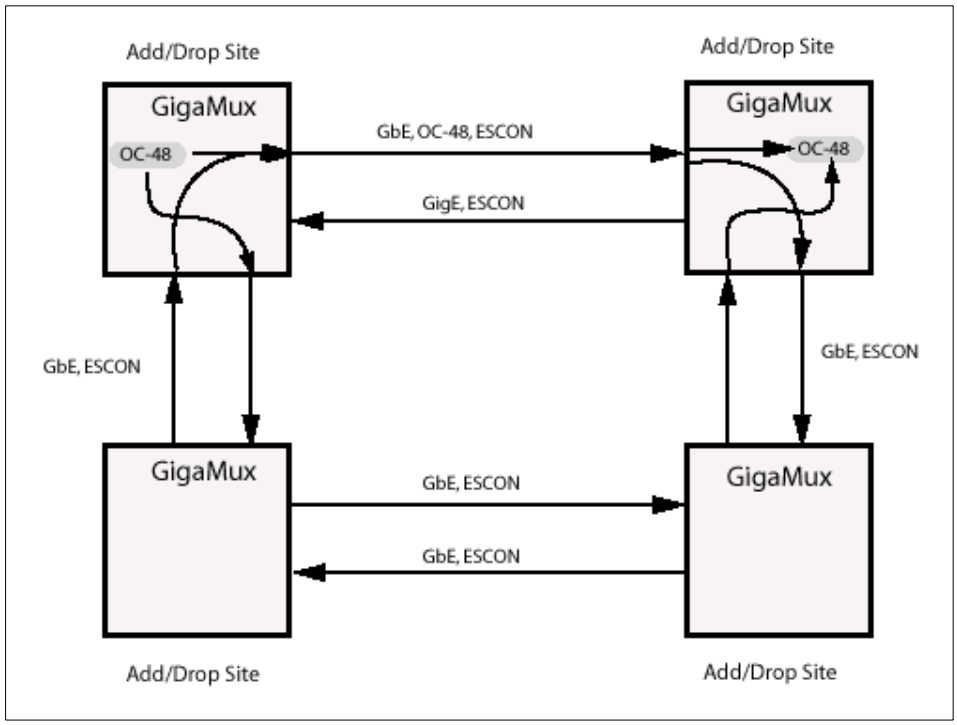


Figure 10-5 Uni-directional ring add/drop

10.2.5 Star system

A star topology is compromised of a hub and connecting sites. The hub site provides a convenient centralized cross-connect location, which facilitates maintenance and management. A star configuration is flexible in that it supports different applications in traffic patterns, such as a ring and bus. Also, this type of topology is useful when future expansion is required.

Sorrento Networks supports 32 unprotected channels in single fiber, bi-directional star system, and 64 unprotected channels in a standard dual fiber uni-directional star system.

We show a star topology system in Figure 10-6.

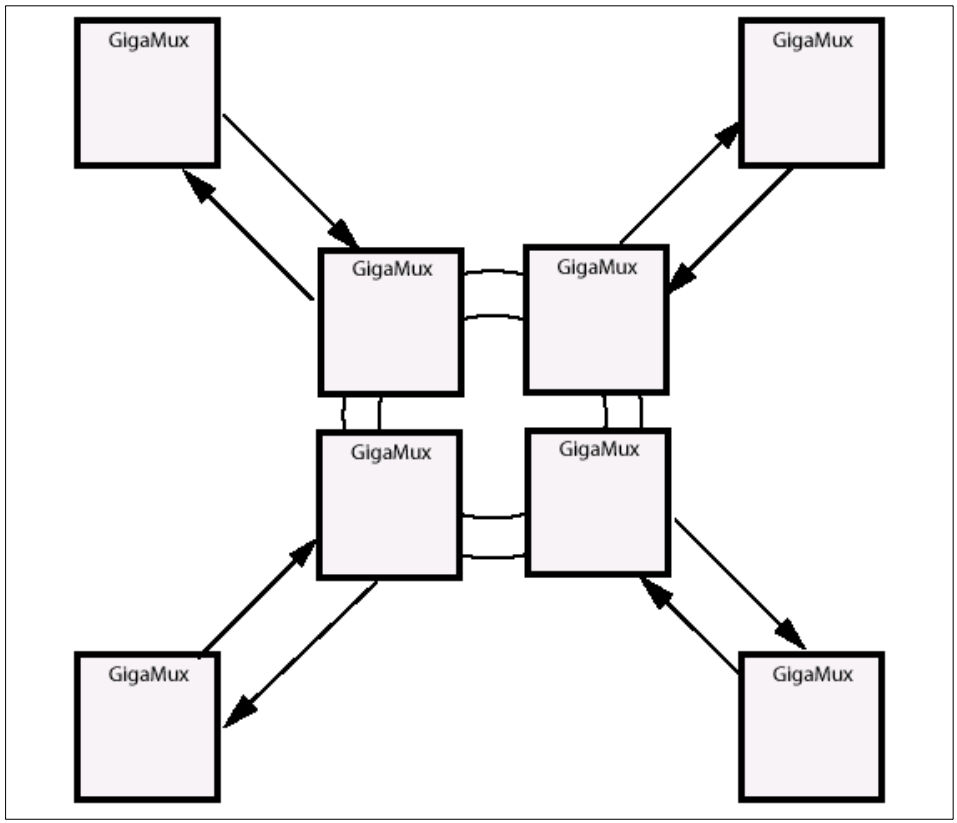


Figure 10-6 Star topology system

10.2.6 Mesh system

A mesh system topology consists of several sites that provide diverse routing. A mesh topology provides a high level of protection. When a failure occurs traffic is re-routed via the clear sites.

Sorrento Networks supports 32 unprotected channels in single fiber, bi-directional star system and 64 unprotected channels in a standard dual fiber uni-directional mesh system.

We show a mesh system in Figure 10-7.

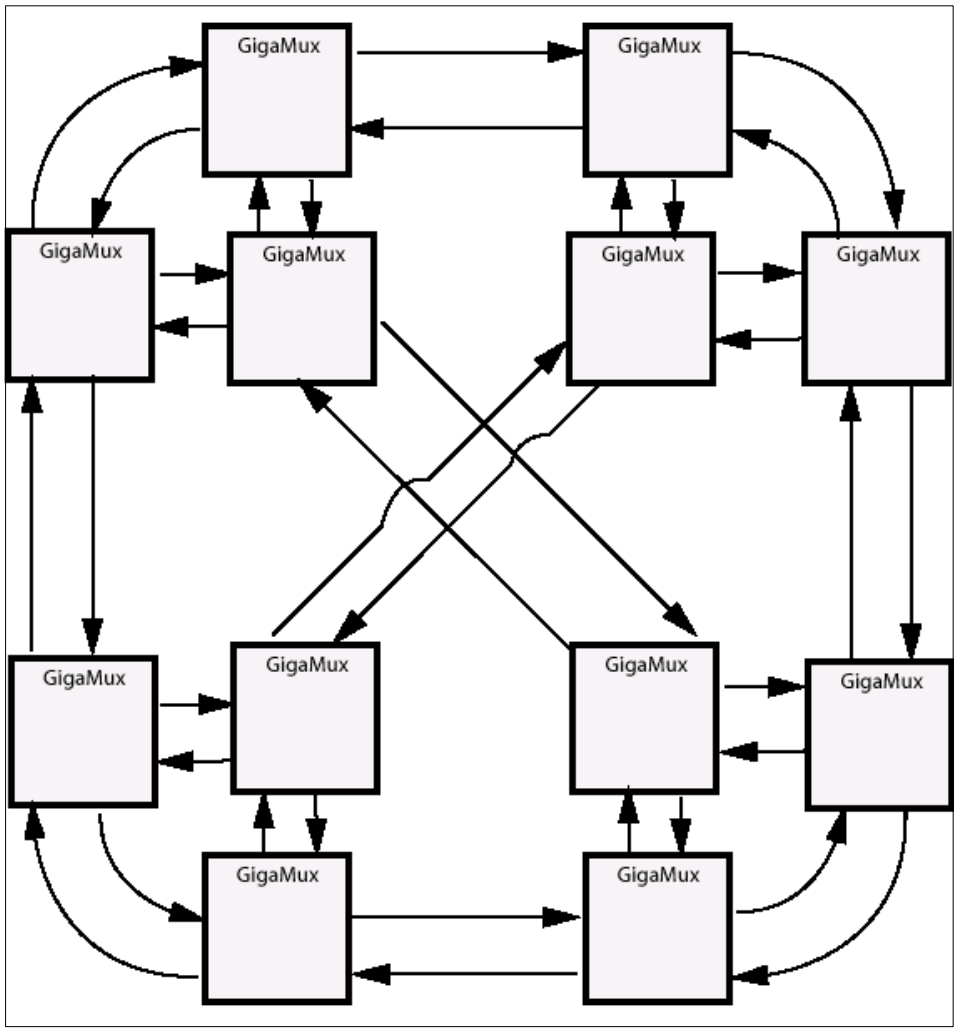


Figure 10-7 Uni-directional mesh system

10.2.7 Protocol independent

GigaMux provides unique mix-and-match flexibility on a per channel basis. A single fiber can share Switched Gigabit Ethernet and Fibre channel (using the GMI-1GESX or GMI-1GnX modules), FDDI and ESCON (using GM-ESC8 modules), OC-1 through OC-48/STM-16 (using GMOX-2.5G, GMCR-2.5G, GM-SMX-2.5G, and GM-SMX-10G modules), OC-192/STM-64 (using GMOX-10G and GMCR-10G modules) and proprietary digital signals.

10.2.8 Performance monitoring

The performance monitoring (PM) feature provides a method to observe and record the performance characteristics of a digital data stream on SONET and SDH networks. The performance monitoring feature enables you to detect performance degradations before a failure occurs on equipment and facilities in your system. It is currently available on the GMCR-2.5G, GMOX-10G, GMCR-10G, GM-SMX-2.5G, and GM-SMX-10G modules.

The following PM parameters are monitored:

- Optical Power Received
- Optical Power Transmitted
- Transmitter Laser Bias Current
- SONET Section B1 and J0 bytes monitoring
- Loss of Signal (LOS) (Alarm)
- Loss of Frame (LOF) (Alarm)

To display PM status and set PM thresholds use the GigaNest Manager or TeraManager.

10.2.9 Traffic flow

The GigaMux system supports uni-directional (simplex) and bi-directional (duplex) traffic flow. By definition there are two wavelengths per channel, and therefore GigaMux supports up to 64 channels uni-directionally and up to 32 channels bi-directionally.

In Figure 10-8 we show the simplex and duplex nature of the GigaMux data flow.

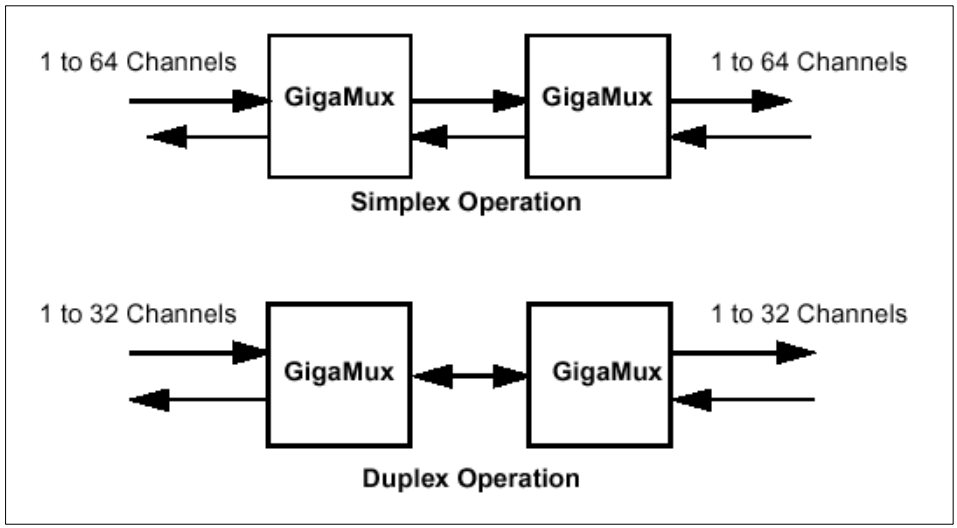


Figure 10-8 GigaMux simplex and duplex applications

10.3 System hardware

The GigaMux has a modular design, with consideration for non-disruptive upgrades, allowing quick restructuring in the field as requirements change and without interrupting existing channel traffic.

The system's components are installed in a 17-slot equipment shelf, and for most configurations, it requires only one floor tile of space. There is only one slot that is reserved for maintenance, other than that any circuit pack can be configured in any slot. Connections are all located on the front panel for easy installation and configuration.

10.3.1 GigaMux system components

GigaMux is a fiber optic transmission system that supports traffic up to 10 Gbps (OC-192/STM-64) per channel. Typically, its components are installed in the 17-slot GN-CS GigaNest equipment shelf. A GigaMux System consists of four elements, as described in the following sections:

- An equipment shelf such as the GN-CS
- Dual power supply and fan assembly
- Management and control
- Optic channel modules (active and passive)

10.3.2 GigaMux equipment shelf

The GigaMux system supports three types of equipment shelves

- GN-CS
- GN-CSP
- GN-2

The GN-CS equipment shelf houses active and passive modules. The GN-CSP houses GMCM and filter modules only. Unlike the GN-CS, the modules equipped in the GN-CSP do not require power, and therefore this shelf is not equipped with a backplane. The GN- 2 is a two slot passive equipment shelf,

In Figure 10-9 we list the various features of the equipment shelves.

GN-CS	GN-CSP	GN-2	Features
X	X		17 slots
X			1 slot reserved for management card
X			Redundant power supply
X			Integral fan assembly with dual power connectors and replaceable air filter
X	X	X	Cable-routing areas permit easy and clean cable connections; no cables extend outside the shelf
X	X		Removable front panel with poly carbonate window
X	X		Reversible mounting brackets for installing in a 19-inch equipment rack or cabinet
X	X	X	Reversible mounting brackets for installing in a 23-inch equipment rack or cabinet
		X	2 slot, for single slot passive modules for a 23 inch rack or cabinet

Figure 10-9 Equipment shelf features

10.3.3 Power supplies

Dual AC or DC power supplies provide fully-redundant protection for the devices in the GigaNest.

- Modular design permits easy removal and replacement
- Mounting brackets for both 19-inch and 23-inch installations

10.3.4 Battery connections

The battery connectors on the rear panel of the DC power modules each have a NEG (negative) and POS (positive) terminal. Two –48 VDC sources can be connected to ensure uninterrupted operation.

10.3.5 Management card

The Management Card occupies the center slot of the GN-CS and provides comprehensive management and control functions for all the channel cards installed in up to 16 slots. In installations where a channel card occupies two slots, the GigaNest Manager can be configured to control devices in two shelves. See the Installation Guide, 42-02022-01 for more information.

Also, note that a Management card is required on each shelf when a node is equipped with a GNC module. Refer to “GigaMux node control card” on page 276 for more information.

The Management Card’s front panel features an async terminal interface and LED indicators that monitor the status of the power supplies, shelf temperature, and integral components,

The rear panel provides 10Base-T Ethernet, auxiliary async, and an 8-pin modular connector for the fan. Local and dial-in terminal access support async ANSI and VT100 terminal emulation. The Ethernet port supports TCP/IP, Telnet, TFTP, and SNMP V1.

10.3.6 Management interface

The Management Interface Card provides the following interfaces and connectors:

FAN
RJ45, 8-pin modular connector links the Management card’s FAN port to up to two fan trays. This allows monitoring of up to six fans.Use a straight Ethernet UTP cable to connect between the FAN port and the Monitor Out port of the fan tray.

ALARM
Provides an alarm signal to an external device when there is either a power supply failure or a shelf high temperature condition. The alarm signals the external device using either an open or short between two alarm terminals. If the external alarm is connected to COM (Common) and NC (Normally Closed), the alarm contacts are closed when the power supply is functioning. A power supply

failure indicates an open between the two alarm terminals. If the alarm is connected to COM (Common) and NO (Normally Open), the alarm contacts are open when the power supply is functioning. A power supply failure indicates a short between the two alarm terminals.

LAN 1
RJ45, DTE, Ethernet 10BaseT connector that links the Management Card's internal Ethernet hub to the Local Area Network (LAN). Use a customer-supplied 10BaseT drop cable to connect the LAN 1 interface to a Media Access Unit (MAU) on the LAN.

LAN 2
RJ45, DTE, Ethernet 10BaseT connector that links the Management Card's internal Ethernet hub to the LAN card of a PC or workstation. Use a crossover UTP cable between two Ethernet interfaces.

LAN 3
RJ45, DTE, Ethernet 10BaseT connector that links the Remote Management Card's internal Ethernet hub to the Service Channels card's Ethernet interface. Use crossover UTP a cable between two Ethernet interfaces.

AUX
RJ45, modular jack connector that connects the micro controllers serial interface port to a PC's or workstation's COM ports. Use a straight UTP cable between connections. The AUX port provides a secondary console interface for direct access of management screens.

10.3.7 GigaMux node control card

The GigaMux node control card (GNC) provides comprehensive OAMP functions through TL-1 for all channel cards.

The GNC card is located in the first chassis of the first equipment rack. The GNC occupies three slots: slot 9, slot 10, and slot 11. The GNC provides EMS support through an embedded TL1 parser. This parser allows the GNC to execute and respond to commands sent by EMS.

The GNC can also send autonomous messages to the EMS for asynchronous traps. TL1 traffic can be routed through the console ports on the front and back panels of the GNC. TL1 users can also use Telnet to open a TL1 session through a LAN connection. Ethernet Port 3 is dedicated to this purpose.

Communications between the GNC and channel cards in the same chassis are accomplished using internal SLIP connections. The GNC and all management cards are chained together through the assigned Ethernet ports to form a local network referred to as the shelf LAN. The GNC connects to this LAN through Ethernet Port 1. Through this port, the GNC communicates with all of the other channel cards through the management cards residing in each chassis along with the channel cards. Ethernet Port 0 is used as a backup connection to the management cards. This port is active only after a break is detected in the Shelf LAN.

Multiple GigaMux Nodes can also be connected using the Ethernet ports on the GNC to create an OSC LAN for managing the networking system. The OSC LAN allows an EMS user to access all NE's on the entire system complex. Like the shelf LAN, the OSC LAN also forms a loop to provide protection against cut cables and other physical signal interruptions.

Alarm indication

The Critical, Major and Minor LEDs are used to indicate the highest severity alarm condition on the system. The Ready LED will be used to indicate whether there is live traffic on the system. The LED will be on or off. If the LED is on, the color will be Green. The NE will use the facility state to determine whether to turn the LED on or off. If all facilities are out of service for whatever reason, the Ready LED will be turned off. The GNC also provides 16 alarm inputs, and 16 relay outputs. All inputs and outputs are user-programmable.

10.3.8 Network control

The GigaMux offers flexible and reliable SNMP support for optical network management. Its network management system remotely monitors key system elements, such as bandwidth allocation, system performance and anomalous conditions. Optional remote dial-in management via V.32 modem opens a backup management path should the primary fail. In addition optional plug-in modules provide node-to-node communications control for enhanced network management.

10.4 Management and control

Sorrento Networks provides node (GigaNest Manager) and network (GigaView and TeraManager) management software for GigaMux.

10.4.1 GigaNest Manager

The GigaNest Manager user interface software provides full management of the GigaMux and access to the terminal management screens of all the devices installed in the shelf.

Terminal management screens are provided for the various channel cards used in the GigaMux system. The following management methods are available:

- ANSI terminal - direct access of the management screens through a terminal connected to the Management Card's CONSOLE interface
- Telnet - access of the management screens from any PC or ANSI terminal on the network using Telnet protocol
- SNMP - access the SNMP agent using Simple Network Management Protocol
- TFTP - remote login is supported for TFTP-based software update maintenance

For more information on operating the GigaMux system using the GigaNest Manager, see GigaMux Operations Guide, 42-01033-01.

The GigaMux provides node and network management software for GigaMux uisng the Network Management System.

10.4.2 GigaView

GigaView is a graphical user interface (GUI) element management system in which you can perform operations, administration, maintenance and provisioning (OAMP) functions for GigaMux. GigaView is designed to run on a single user platform. It is available for Windows 98, 2000 and NT. GigaView utilizes SNMP protocol for all messaging.

For more information on operating the GigaMux system using GigaView, see the GigaView User Guide, 42-01040-01. GigaView is an optional software package, for information on this package contact your Sorrento Networks representative.

10.4.3 TeraManager

TeraManager is a graphical user interface (GUI) element management system in which you can perform OAMP functions for GigaMux. It utilizes TL-1 for all messaging and provides a topical view of all network elements in a system. TeraManager enables users to visually monitor several nodes, links and alarms in a system quickly and easily.

TeraManager is an optional software package, for information on this package contact your Sorrento Networks representative.

10.4.4 Channel modules

Sorrento Networks provides three types of modules:

- Interface (active)
- Filter (passive)
- Span modules

Each type of module has a distinct functional purpose when designing a system. Active modules support individual channel requirements and convert source signals to ITU-T compliant DWDM signals.

Dependable passive modules perform multiplexing and de-multiplexing functions.

Span modules, such as amplifier, protection and DCM modules, impact the signal at the span sections of a network. All Channel modules are slot-independent and can function together in the same shelf or different shelves.

10.4.5 Active modules

Sorrento Network's active modules can accept and transmit the following types of signals: digital bit streams from 16 Mbps to a full 10 Gbps (OC-192/STM-64), including Fast Ethernet 100 Mbps and Gigabit Ethernet, ESCON, ETR, ISC and proprietary digital bit streams. Low speed modules offer economical support for 16 Mbps through 2.5 Gbps applications. High speed modules extend this range to support traffic up to 10 Gbps. All protocol and bit-rate independent active modules automatically match the incoming optical/electrical digital signal rate including both synchronous and asynchronous data. Bit-rate and protocol independent circuitry is used to provide flexibility with different application formats.

GMOX-xx and GMCR-xx transponder modules

The GMOX-xx, GMCR-xx transponder modules convert a source optical signal into a ITU-T wavelength for DWDM. Sorrento provides different types of transponder modules:

- GMOX-2.5G
- GMOX-ISC
- GMOX-25LR
- GMOX-10G
- GMCR-2.5G
- GMCR-10G

These modules support optical inputs in the 1310 and 1550 nm bands.

GMOX-2.5G

The GMOX-2.5G module supports 16 Mbps to 2.5 Gbps. This module performs a 2R operation (reshape and re-amplify). It supports both single mode and multi-mode fiber applications. Specifically, Fibre Channel, OC-3 and OC-12 SONET speeds, direct ESCON transport and ETR (external timing reference) protocols are supported. The GMOX-2.5G has added hardware control that enables the transmitter and receiver to be controlled directly from software. This feature enables the software to override the APSD function to allow testing and set-up during installation or for troubleshooting.

GMOX-ISC

The GMOX-ISC module supports 1 Gbps. This module performs a 2R operation (reshape and re-amplify) and OFC (Open Fiber Control). It supports IBM Geoplex environment applications, such as ISC (Inter-System Coupling Link). The GMOX-ISC has added hardware control that enables the transmitter and receiver to be controlled directly from software. This feature enables the software to override the APSD function which allows testing and set-up during installation or for troubleshooting.

GMOX-25LR

The GMOX-25LR accepts signals from 50 Mbps to 2.5 Gbps. This module performs a 2R operation (reshape and re-amplify). This module is designed for regional applications that are up to 600 km.

GMOX-10G

Sorrento Networks supports two 10 Gbps modules: GMOX-10G-PIN and GMOX-10G-APD. These modules can be used for applications that are up to 40 km and 80 km, respectively. These modules support SONET signals only. The GMOX-10G-PIN and GMOX-10G-APD modules perform the 3R (re-shape, re-amplify and re-clock) operation. The difference between the two is in the diode, PIN or APD diode. This module also supports performance monitoring (PM), such as the loop back test feature. This feature is software configured to provide either routine testing or on demand testing.

GMCR-2.5G

This is a 3CR (clock recovered) transponder with two speed clock recovery of OC-48 and OC-12 and includes SONET/SDH performance monitoring. It has 1310 nm SMF optics.

GMCR-10G

Similar to the GMCR-2.5G, this is an OC-192/STM-64 transponder. It has 1310 nm SMF optics.

GM-FNTn broadband transmitter module

The GM-FNT1 and GM-FNT2 modules are primarily used in cable TV applications. These modules convert a source broadband signal to an optical signal. The GM-FNT1 supports 50- 750 MHz up to 18 QAM channels per wavelength. This module brings multichannel narrow cast signals from a source location to a remote location. The GM-FNT2 module supports 5-65 MHz QPSK. This module is a reverse path transmitter that brings the multichannel narrow cast signals from the remote location back to the source. Optical inputs in the 1310 nm and 1550 nm bands are supported for these modules.

GM-FNRn broadband receiver module

The GM-FNR1 and GM-FNR2 modules are the counterparts to GM-FNT1 and GM-FNT2 modules, respectively. These modules translate an optical signal from the GigaMux System back to the customer's RF. The GM-FNR1 module is a forward path receiver, it receives an optical signal from the source. The GM-FNR2 module is a reverse path receiver, it receives data from the remote location.

GM-ESC8 EPC modules

The GM-ESC8 EPC module is a proprietary eight channel full duplex multiplexer that optimizes wavelength efficiency by transporting multiple channels of varying protocols. These protocols are configured using the GigaNest Manager software. We explain the concepts associated with EPC in 10.8, "Electronic Photonic Concentrator" on page 297.

GMI-1GnX and GMI-GESX-2 interface modules

The GMI-1GSX and the GMI-GESX-2 interface modules are designed to meet the Gigabit Ethernet and Fibre Channel specifications for SX (850 nm). The GMI-1GLX module is designed meet Gigabit Ethernet and Fibre Channel specifications for LX (1310 nm). Each module provides a duplex SC optical connector.

GMTR transceiver module

The GMTR Transceiver Module is a combination transmitter and receiver. The transmitter converts a source electrical signal into an ITU-T DWDM signal. The receiver converts an ITU-T DWDM signal into the customer's electrical signal. Sorrento Networks provide two transceiver modules: a 1.5 Gbps module, GMTR-1.5G and a 2.5 Gbps module, the GMTR-2.5G.

GMOA optical amplifier module

The GigaMux Optical Amplifiers (GM-OA) are optional modules that can be used for extended distance applications and critical applications when extra optical power is required. Sorrento Networks provides these types of optical amplifiers:

- GM-OA-T5
- GM-OA-T6
- GM-OA-T7

The GM-OA-T5 is a double pump amplifier, and GMOA-T6 and GM-OA-T7 are triple pump amplifiers that are typically used in cascading amplifier applications. The GM-OA-T7 is used as a booster amplifier.

10.4.6 Passive modules

Passive modules form the optical networking core. They are used to create 4, 8, 16, 32 and 64 channels systems. Passive modules perform multiplexing and de-multiplexing functions. They do not require any power and can be installed on GN-CS, GN-CSP and GN-2 equipment shelves.

GMMD modules

The GMMD modules can simultaneously serve as multiplexers and demultiplexers. As a multiplexer, the GMMD module accepts up to four narrow line width optical signals and provides a combined optical output to a GMEx module or the fiber trunk. As a demultiplexer, it accepts a narrow line width optical signal from a GMEx module or the fiber trunk and provides up to four optical outputs.

Sorrento Networks has designed GMMD modules for C and L Band applications.

GME expansion module

The GME Expansion Modules (GME1, GME2, GME1C, GME2C, GME-1L or GME-2L) perform a multiplexer/demultiplexer function. The addition of these modules enable support of up to 16 channels.

GME and GME2

The GME and GME2 modules are two port expansion modules.

GME1C and GME2C

The GME1C and GME2C modules are C band 4 port modules. These modules facilitate add/drop functionality.

GME1L and GME2L

The GME1L and GME2L modules are L band 4 port modules. These modules facilitate add/drop functionality.

GMDE, GMEI and GMLEI expansion module

Sorrento Networks provide three expansion modules capable of supporting 32 channels:

- GMDE
- GMEI
- GMLEI

The GMDE Dual Expansion Module is used to allow bidirectional capability over one fiber, independent of wavelength assignment.

We illustrate this in Figure 10-10.

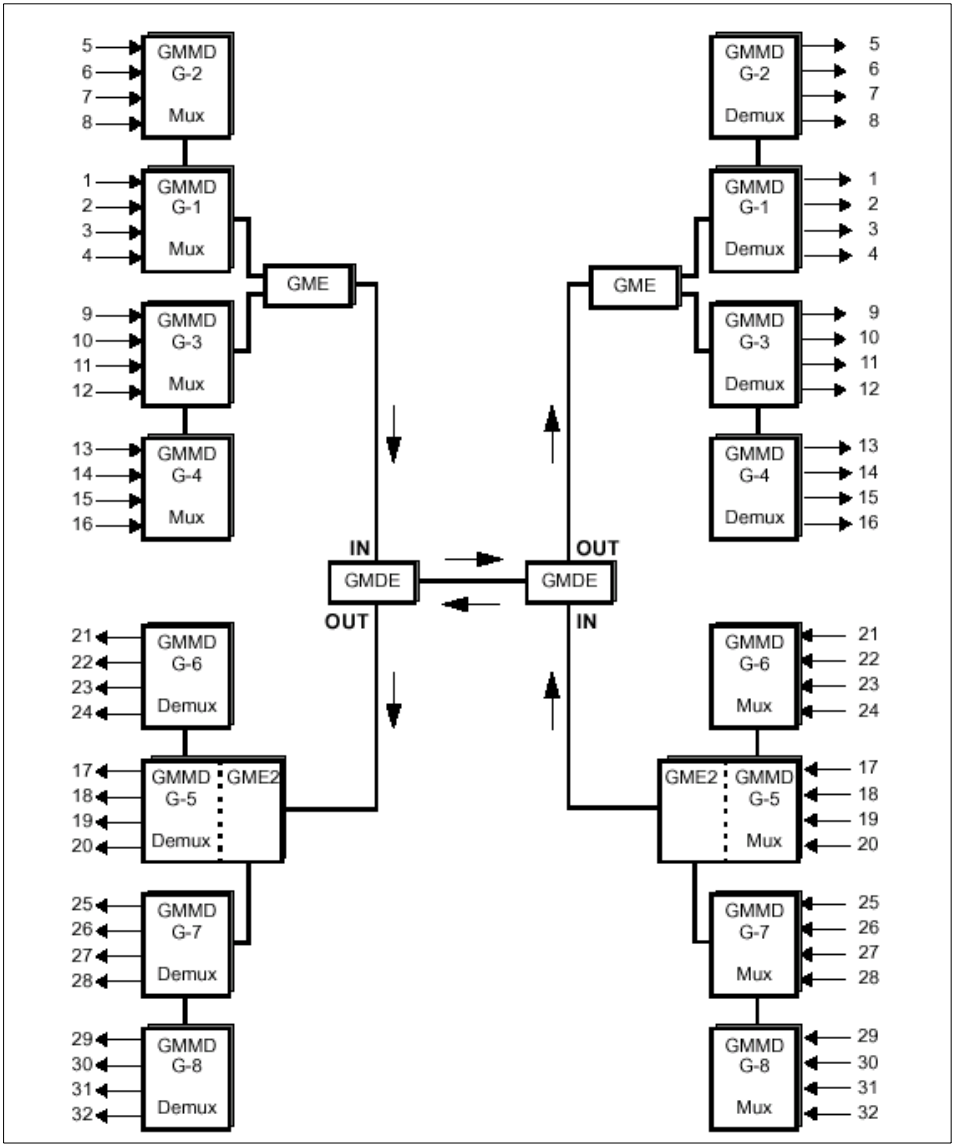


Figure 10-10 GMMD, GME, GME2 and GMDE functional diagram

The GMEI Dual Independent Expansion Module is used when creating a network capable of supporting 32 channels with directional independence.

We illustrate this in Figure 10-11.

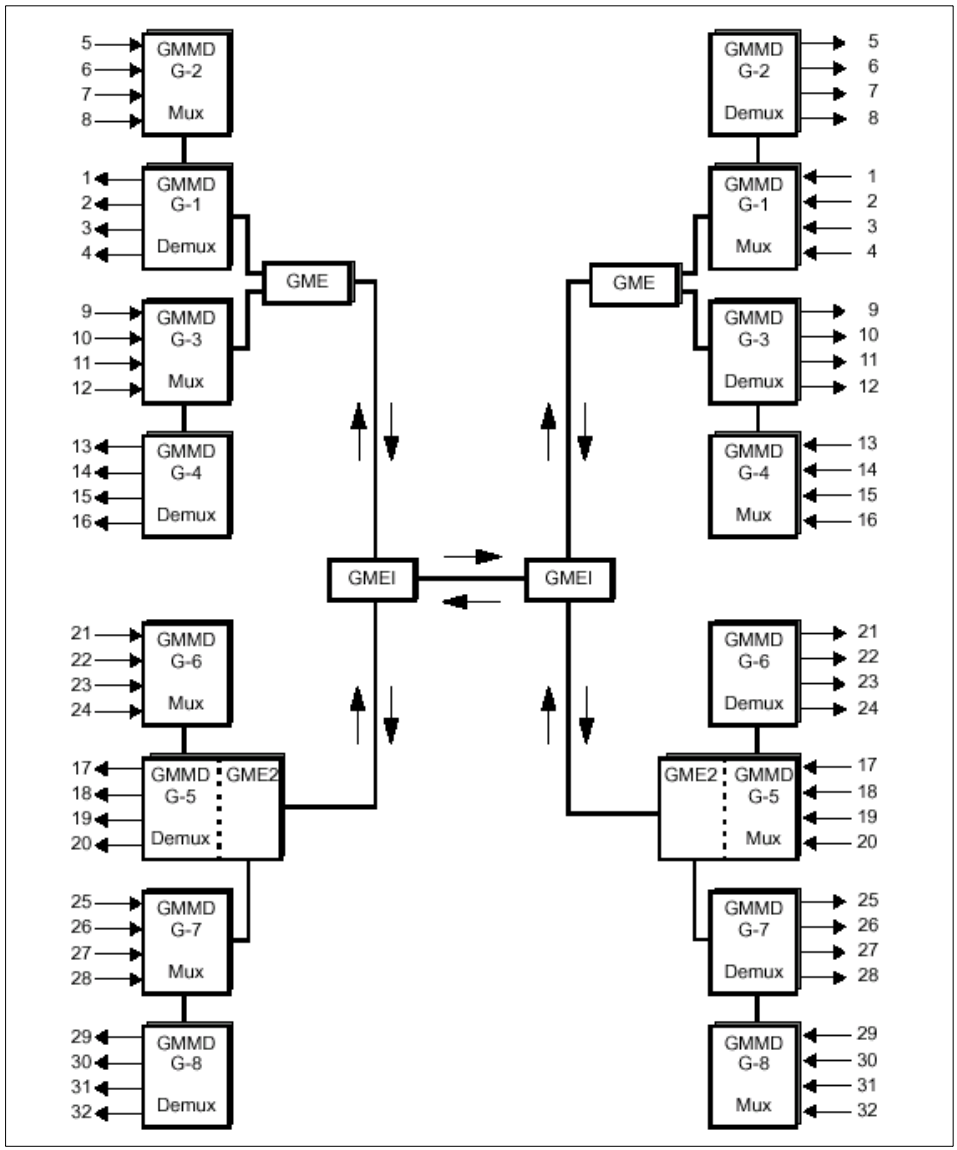


Figure 10-11 GMMD, GME, GME2, GMEI functional diagram

The GMLEI Dual Independent L Band Expansion Module is used when creating a network capable of supporting 32 channels in the L Band.

We illustrate this in Figure 10-12.

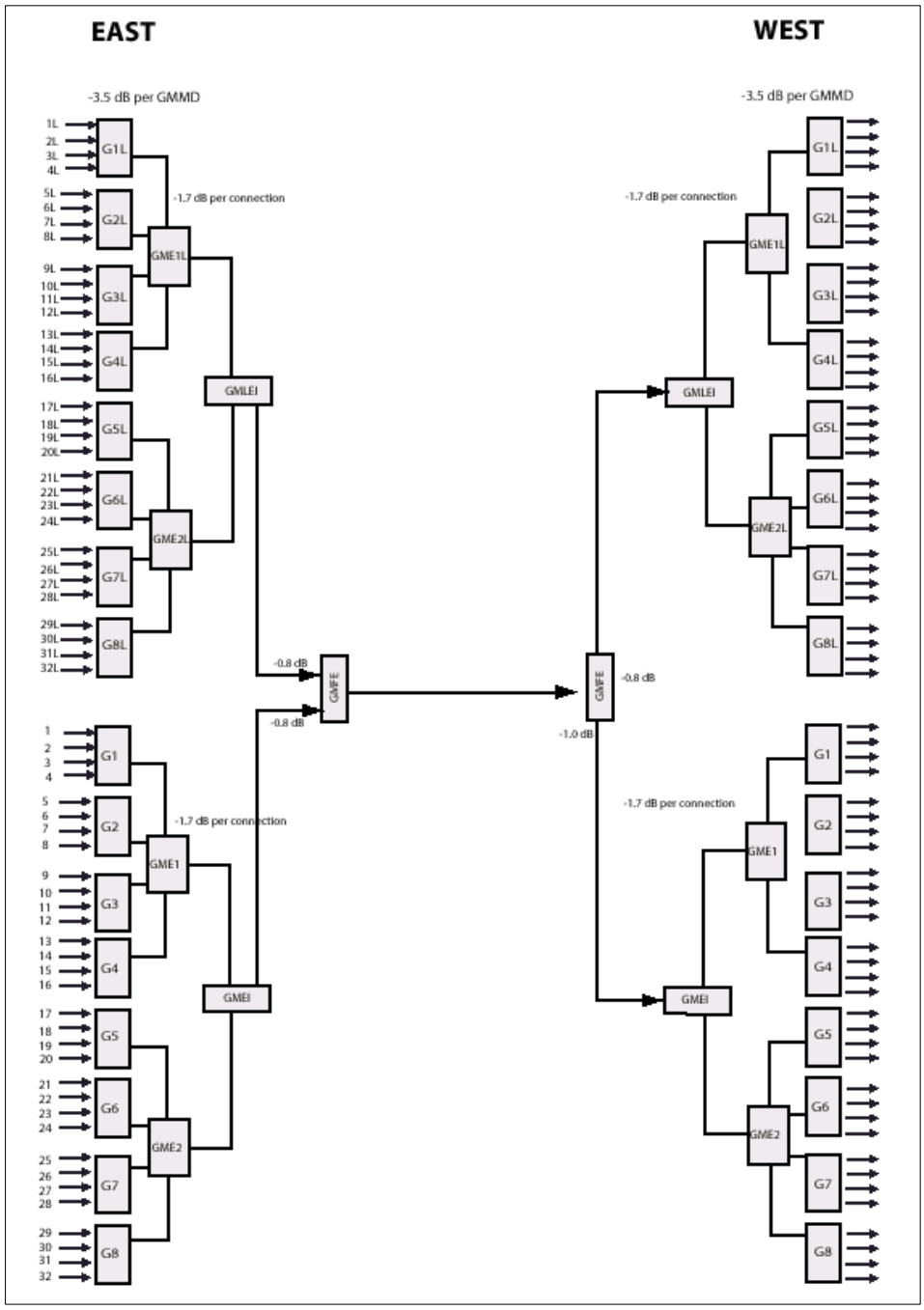


Figure 10-12 GMMD, GMLE1, GME1L and GME2L functional diagram

GMFE expansion module

The GMFE is an L Band Expansion Module, which performs a multiplexer/ demultiplexer function. The GMFE is a bidirectional module with three ports; one common and two outputs/inputs. This module is used when creating a network capable of supporting 64 channels (32 C band and 32 L band). The GMLEI and the GMEI/ GMDE modules connect to this module.

GMWD module

The GMWD module, also referred to as the Sidestreet Channel, is used to integrate GigaMux DWDM signals with existing fibers carrying 1310 nm legacy signals. The GMWD module supports all GigaMux channel configurations in addition to processing the legacy signals.

10.4.7 Span modules

Span modules are modules that affect the signal at the span section.

Optical service channel modules

The Optical Service Channel (OSC) connects GigaMux nodes to one another and provides a dedicated reporting channel for the network management system (NMS). Two trunk fibers are required to operate the OSC. The OSC cannot be used if only one fiber is used to interconnect the nodes. Remote login is supported for Telnet, SNMP-based management access, TL1 based management access and TFTP-based software update maintenance. The OSC operates at the standard 1510 nm wavelength and is outside any optical amplifier's pass band. The OSC consists of two modules:

- GM-OSC is the active communications module.
- GM-SCF is the passive filter module to integrate the 1510 nm OSC Tx and Rx onto the fiber trunks.

GM-OSC

The GM-OSC is an active DWDM laser transceiver that connects via two fiber jumpers to the GM-SCF filter. The GM-OSC can be placed in four locations in an equipment shelf: the two far ends and the slots surrounding the MGMT card (slots 10 and 11). This positioning allows the RJ45 connector in the back of the GM-OSC module to match up with the access holes in the shelf, and therefore make RJ45 connector accessible.

GM-SCF

The GM-SCF is a passive, dual-filter module that is installed between the duplex fiber trunk and the existing GigaMux trunk input and output. Typically two such filter and transceiver modules are installed at a node to support redundant dual fiber routes.

Protection Modules

The Fiber Trunk Switch protection modules (GM-FPMT, GM-FPMR and GM-FPS2W) are designed for systems with redundant fiber trunks (Fiber 1 as primary and Fiber 2 as redundant). All fiber protection modules are designed for single mode fiber, and should not be used on multi-mode fiber. The GM-FPMR status screen in the GigaNest Manager (see GM-FPMR Module Status in the Operations Manual) informs the user which fiber trunk is active and its condition (normal or failed).

By using Sorrento Networks TeraManager Network Management system, protection modes can be set.

GM-FPMT

The GM-FPMT module is installed at the transmit end. It contains a 1:2 splitter that splits the source signal into two identical signals. The GM-FPMT module can be used for simplex or bidirectional fiber trunk applications.

GM-FPMR

The GM-FPMR module is installed at the receive end and should be used with fibers carrying uni-directional traffic. It has two optical inputs and a common output. The optical threshold is set at -25 dBm 0.5 dB. An alarm event is generated indicating a failure when the optical level drops below this threshold. If the optical level returns above this threshold, an alarm event is generated indicating fiber restoral. Input FT1 on the module is for the working trunk and input FT2 on the module is for the protection trunk.

GM-FPS2W

The GM-FPS2W is similar to the GM-FPMR module, except it is used for bi-directional traffic protection. While allowing bi-directional traffic the design of the GM-FPS2W requires careful channel allocation for proper operation. The optical threshold is set at -23 dBm 0.5 dB. An alarm event is generated indicating a failure when the optical level drops below this threshold. If the optical level returns above this threshold, an alarm event is generated indicating fiber restoral. Input FT1 on the module is for the working trunk and input FT2 on the module is for the protection trunk. The GM-FPS2W module is designed so that the wavelengths in the Tx direction fall within one of two bands: either 1530 to 1544 nm (red band) or 1547 to 1565 nm (blue band) bands and the wavelengths in the receive direction fall into the complimentary band.

GMAD add-drop module

The GMAD add-drop module is a passive optical module that allows a selected channel to be dropped while passing all other channels. The GMAD provides an economical and efficient method of adding and dropping specific optical channels (wavelengths) along a fiber path while allowing the remaining channels to pass through.

In Figure 10-13 we show a GMAD functional diagram.

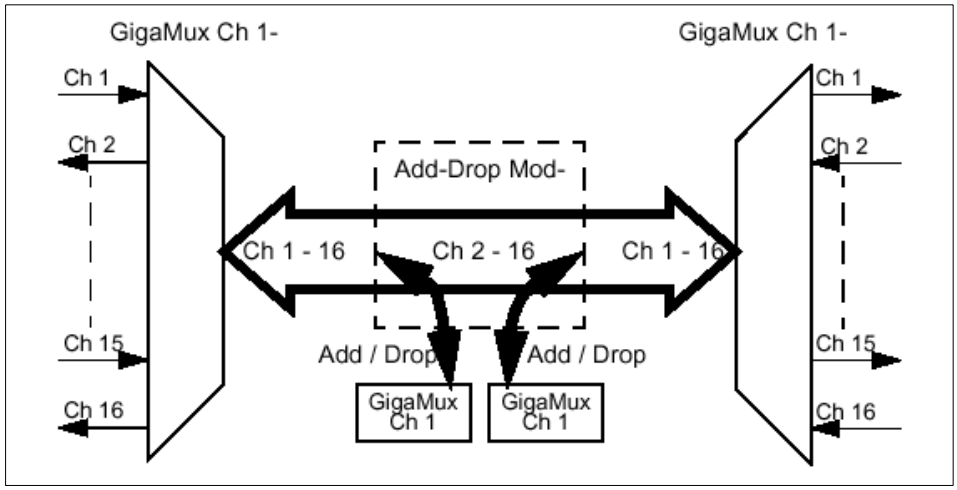


Figure 10-13 GMAD functional diagram

GMCM dispersion compensation module

The GigaMux dispersion compensation modules (DCM), GMCM-20, GMCM-40, GMCM-60, are optional modules that can be used to counter dispersion in the C Band. These modules decrease the effect of chromatic dispersion by adding the correct amount of opposite dispersion.The GMCM-20, GMCM-40 and GMCM-60 modules compensate 20 km, 40 km and 60 km of SMF-28 fiber, respectively.

10.5 Assigning channels

Once you have determined your system topology and the number of channels required in your system, you can begin assigning channels to groups. This topic describes the general guidelines for efficiently assigning channels in point-to-point and ring systems.

10.5.1 Point-to-point

In a point-to-point network, where no intermediate add/drop sites are involved, optical channels should be assigned sequentially, beginning with Channel 1.

Simple point-to-point uni-directional systems

A point-to-point uni-directional system is interconnected by two fibers with two identical systems, operating back-to-back. Channel assignments are identical for both fibers.

In Figure 10-14 we show assigning channels in a uni-directional ring.

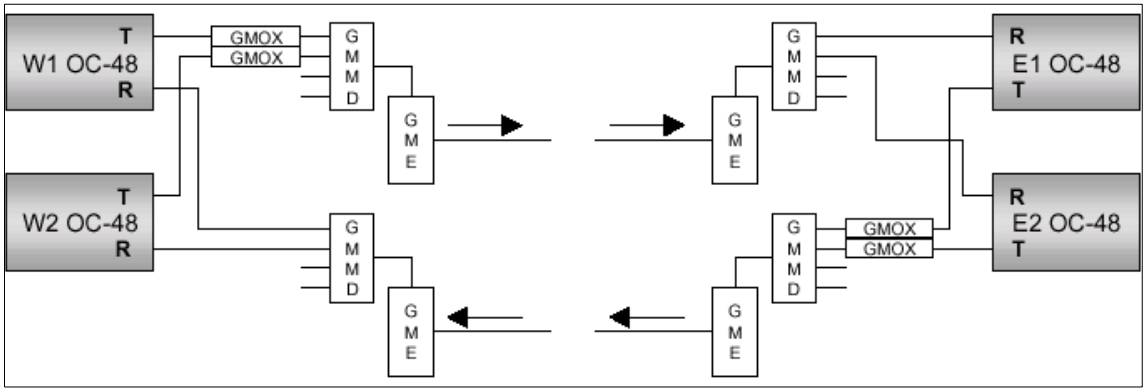


Figure 10-14 Assigning channels in a uni-directional ring

This illustrates how the first OC-48 (W and E) is assigned to Channel 1 for both transmit and receive. The second OC-48 (W and E) is assigned to Channel 2. The underlying principle is that number 1 OC-48 will always use Channel 1, whether it is east to west or west to east, number 2 OC-48 will always use Channel 2, and so on.

Simple point-to-point bi-directional system

If the point-to-point network uses only one interconnecting fiber, channel designation and expansion is still assigned sequentially, beginning with Channel 1.

Figure 10-15 displays channel assignment in a bi-directional system.

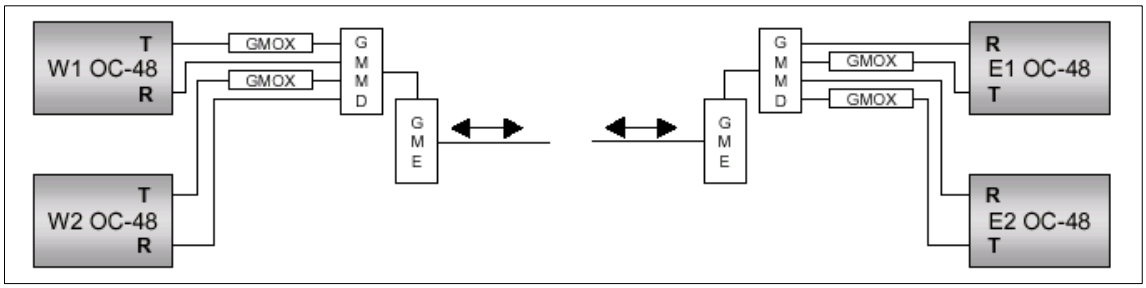


Figure 10-15 Bi-directional channel assignment

W1 OC-48 uses Channel 1 to transmit and Channel 2 to receive. W2 OC-48 number 2 transmits on Channel 3 and receives on Channel 4. E1 OC-48 receives on Channel 1 and transmits on Channel 2. Similarly, E2 OC-48 receives on Channel 3 and transmits on Channel 4.

Point-to-point with add/drop

When add/drop functionality is added, channel assignment should be arranged in logical groups of 4, where each group corresponds to one GMMD. Where possible, the filter group break points, 4, 8, 16, and 32 channels should be employed.

Use the following rules to assign channels:

- Segregate channels in common spans, that is, place spans that terminate at the same point together.
- Arrange them in sets of 4 (each set will account for 1 GMMD group).
- For 16 channel systems, use GMMD Groups 1 and 2 for longest distance spans; use GMMD Groups 3 and 4 for add/drop nodes.
- For 32 Channel Systems, use GMMD Groups 1, 2, 5 and 6 for longest distance spans; use GMMD Groups 3, 4, 7 and 8 for add/drop nodes.

10.5.2 Ring

A ring network consists of point-to-point spans and, as such, follows the same principles as point-to-point systems. Channels should be grouped according to the commonality of their spans and segregated into 4-channel groups.

We assume that the system is the ring illustrated in Figure 10-16.

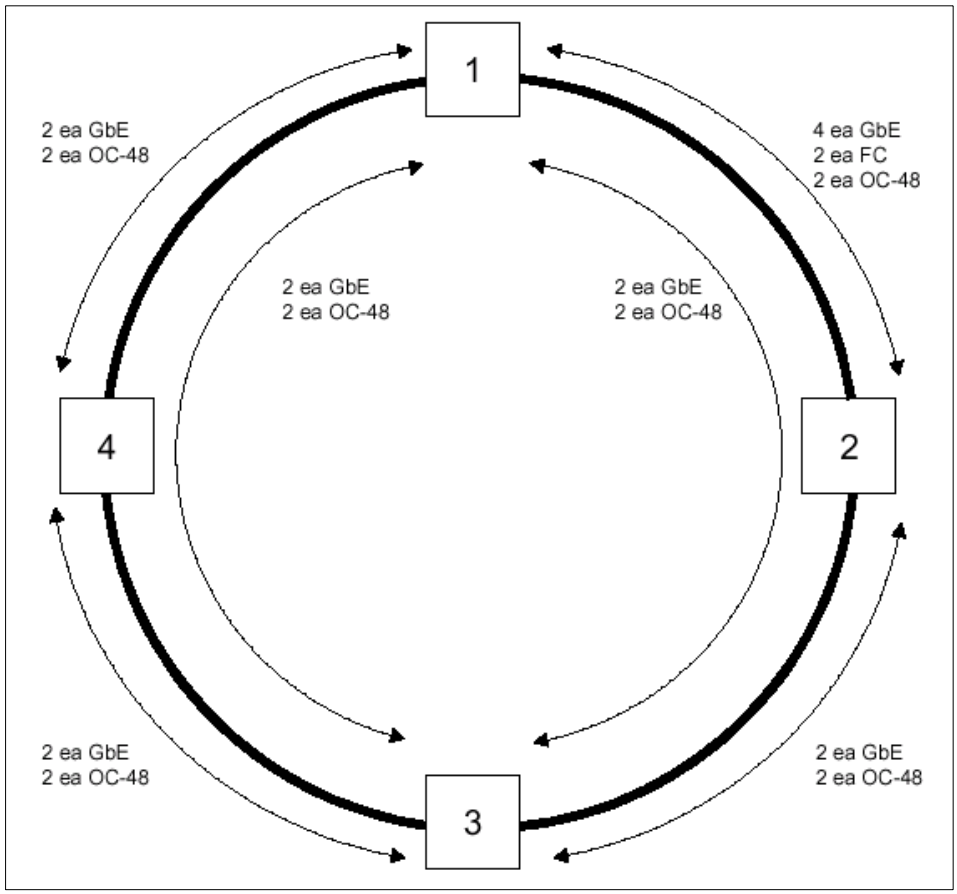


Figure 10-16 Assigning channels in a ring system

The network is now a four node ring where Node 5 and Node 1 are the same node. This illustrates how a ring can be broken open and displayed as a set of point-to-point spans.

10.6 Sidestreet channel

Sorrento Networks now offers Sidestreet Channel which is an optical solution that enables metro area carriers to inject DWDM-based, multi-gigabit bandwidth into any span of their legacy networks. Sidestreet Channel enables carriers to target big bandwidth delivery only to those sites that require it, while allowing other sites on the same metro ring to maintain their legacy architecture without costly system-wide upgrades.

By utilizing Sorrento's Sidestreet Channel, carriers can now have the ability to migrate their networks to multi-gigabit capacities one site at a time and avoid costly, time-consuming “forklift” upgrades of their entire networks.

Some of the main features are:

- More efficient networks by creating optical side streets
- Open new pathways for enhanced DWDM performance at selected nodes only
- Provides a seamless bridge to economically capture new services and revenues
- Negates the need to merge all your traffic into wavelengths

In the metropolitan area network, it is difficult to predict which spans will require extra bandwidth. A typical metro optical ring will have sites that experience greater demand for bandwidth, while others sharing the same ring will grow more slowly. With Sidestreet Channel, carriers now have the means to add big-bandwidth optics on a site-specific basis. By utilizing Sidestreet Channel it will allow you to upgrade, as needed, to DWDM capacity in less time and at lower cost.

Sidestreet Channel creates two rings on the same fiber, a low-cost single-channel ring for legacy equipment and a multi-channel DWDM ring for additional and enhanced features, which can both be managed as a single, logical network. The ability of a carrier to manage both rings on one network can allow all nodes to be integrated to support both legacy and DWDM-enhanced transport simultaneously.

Sidestreet Channel supports virtually every protocol including SONET, ATM and IP.

Sidestreet Channel also allows carriers to shape raw capacity into a form that can be used directly by the attached network equipment, aggregate multiple lower-speed connections for more flexible wavelength utilization, and provide connectivity and restoration in multi-ring, hub and ring/mesh combination networks.

10.7 Summary

Here is a summary of the features and specifications of the GigaMux.

The benefits are:

- Aggregate throughput up to 640 Gbps

- From 16 Mbps to OC-192/STM-64 over the same fiber, simultaneously running SONET/SDH, IP, ESCON, ATM, Fibre Channel, Ethernet, Video and more over the same fiber
- Combine up to 64 independent, high-speed signals
- Point-to-Point, Linear Add-Drop, Ring, Mesh or hybrid architectures
- Simplifies network while reducing costs and increasing bandwidth

The system specifications are:

- Multiplex Configurations (per fiber)
  - 64 channel duplex or mixed unbalanced duplex (32 ITU C band / 32 L band)
  - 100 GHz Spacing
- Capacity
  - Up to 9.953 Gbps per wavelength
- Data Rates
  - 16 Mbps to 2.5 Gbps NRZ
  - 9.953 Gbps NRZ, 3R
  - ITU Grid Wavelength
- Optical Connector
  - FC/PC
  - SC
  - ST
- Fiber Protection
- Switching Time
  - < 25 ms
- Protocols Supported
  - SONET/SDH
  - ATM
  - IP
  - ESCON
  - Gigabit Ethernet
  - Fibre Channel
  - Fast Ethernet
  - Ethernet
  - Video
  - EPC
- Topology Supported

- Point-to-point
- Linear Add-Drop
- Ring
- Mesh
- Hybrid architectures

The GigaMux modules are:

- EPC sub-rate multiplexer (more details are included in 10.8, “Electronic Photonic Concentrator” on page 297)
  - Asynchronous ESCON, FDDI, OC-3/STM-1
  - 8 ports each with ST connectors
- Synchronous OC-3/STM-1 and OC-12/STM-4
  - 8 ports each with ST, SC, or FC connectors
  - 16 ports each with ST, SC, or FC connectors
- Synchronous OC-48/STM-16
  - 4 ports each with ST, SC, or FC connectors
- Synchronous Gigabit Ethernet
  - 2 ports each with full-rate 1.25 Gbps GigE; ST, SC, or FC connectors
  - 4 ports each with fractional 1.25 Gbps GigE, 50 Mbps increments; ST, SC, or FC connectors
- Synchronous Fibre Channel
  - 2 ports each with 1.0625 Gbps FC; ST, SC, or FC connectors
- GMOA band C optical amplifier
  - Models
    - Booster
    - In-Line
    - Pre-Amplifiers with or without Monitor Ports
- Optical add/drop mux/demux
  - Models
    - One Channel (Use GMMD for More Than 1 Channel Add/Drop)
  - Spacing
    - 200 GHz
- Ethernet service channel
  - Optical Input/Output
    - Wavelength = 1510nm
    - Bit rate = 10 Mbps

- Optical Connector
  - FC/PC
  - SC
  - ST
- Ethernet Input/Output
  - RJ45 UTP cable to 100 ft. at 10BASE/T
  - GM-SCF Dual OSC trunk filter

The GigaMux management features are:

- GigaMux management
  - Connectivity
    - Local Async Terminal
    - Dial-Line
    - 10/100M Ethernet
  - Parameter Control
    - Menu driven screens for all parameters
- GigaView - Enterprise Management System (EMS) SNMP Based
- Craft Interface
  - SNMP MIB2 plus enterprise MIBS support all terminal access functions
- Host Platforms
  - Sun/Solaris: GigaView with HP OpenView NNM Support
  - Windows/NT: GigaView for Windows
- TeraMAN - Carrier Management System (TL1 Based)
- Craft Interface
  - TeraMAN Craft Interface Terminal (CIT)
  - Supports TL1 and provides a GUI identical to that of the EMS
- Element Management System
  - TL1 based EMS conforming to Telcordia and ITU standards
  - Supports northbound CORBA interface
  - Provides FCPS functionality through an easy-to-use GUI
- Host Platforms
  - CIT: Windows NT PC or Laptop
  - EMS: Sun/Solaris
- GigaNest Chassis Environmental Dimensions
  - 9" wide x 10.5" high x 12" deep

- Power Supply
  - 100 - 240 VAC 400 Watts
  - 40 - 57 VDC 400 Watts
- Certifications
  - NEBs Level 3
  - CE EN 50082-1 and EN 50022

10.8 Electronic Photonic Concentrator

With data traffic surpassing voice traffic and the explosive growth of the Internet, the insatiable appetite for bandwidth continues. New applications and services are stressing the already over-burdened metro networks. It is critical that service providers are capable of meeting bandwidth demands cost-effectively.

We show the Electronic Photonic Concentrator (EPC) in Figure 10-17.

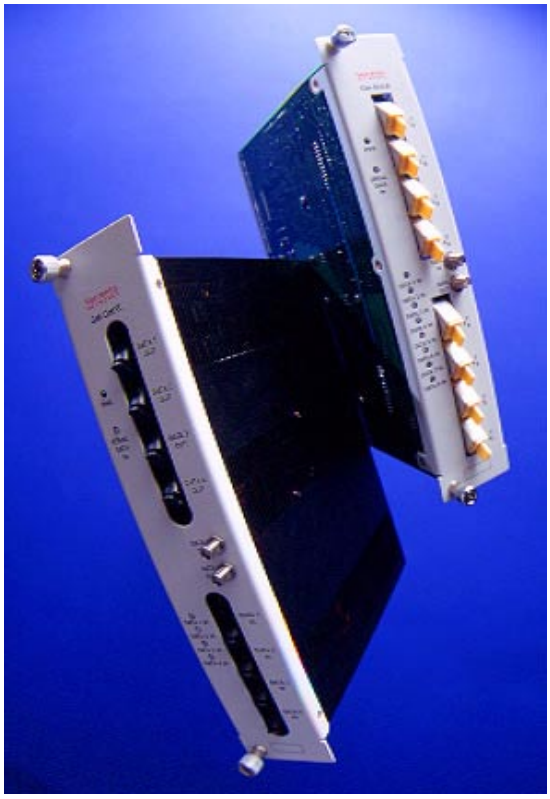


Figure 10-17 Electronic Photonic Concentrator

10.8.1 Increasing network efficiency

Sorrento Networks' EPC allows service providers to meet the growing variety of traffic demands on their networks by carrying traffic in its native formats. While extending the service providers' network reach to the access layer, EPC removes latency by eliminating the optical-electrical-optical (O-E-O) conversion and providing seamless network manageability. EPC is a sub-rate multiplexer that cost-effectively accesses equipment operating at lower speeds (10 Mbps to 200 Mbps), combines the signals onto one wavelength and transports the various signals across the entire optical network. Traditionally, when high-bit-rate SONET equipment (for example, OC-48, OC-192) is deployed in metropolitan optical networks, there is a granularity problem.

Maximizing bandwidth

Specifically, many metro client signal formats only take up a fraction of a 2.5 Gbps (or greater) pipe, and hence the remaining available bandwidth can be wasted unless additional, expensive SONET TDM multiplexing equipment is deployed. This equipment increases the cost and complexity of service provider networks, and moreover, payload mapping is restricted to a few specific sub-rate tributary signal formats.

With EPC, network service providers can maximize each wavelength's capacity as much as 16:1 by directly placing sixteen separate signals onto each DWDM wavelength with Sorrento's new 16-port SONET/SDH EPC.

For example, if a customer needs a 200 Mbps signal transported, the EPC will allow that signal to be combined with up to seven other signals onto one wavelength instead of wasting the entire 2.5 Gbps (or greater) bandwidth on a singular sub-rate data stream.

The EPC greatly improves efficiency and allows operators to realize reduced network operating costs.

With each fiber wavelength capable of up to 10 Gbps capacity, it is important to place as much traffic as possible onto wavelength, therefore maximizing the capacity that each fiber allows. Utilizing the standardized ITU grid, EPC works in conjunction with the GigaMux and TeraManager to provide an end-to-end metro optical network solution.

10.8.2 Products

The EPC product family covers the following protocols:

- ESCON EPC
- Fibre Channel EPC
- OC-48/STM-16 EPC

- OC-192/STM-64 EPC
- GigaBit Ethernet EPC

Fibre Channel EPC

Here is an overview of the Fibre Channel EPC features:

- Two-slot module
  - Multiplexes 2 Fibre Channels/FICON signals (1.0625 Gbps) into an OC-48
- Field-installed transceivers
  - 850 nm, MM
  - 1310 nm, MM
  - 1310 nm, SM (2 km, 10 km, 20 km)
- Fully compliant to T11 Spec
- SONET/SDH PM
- Loop-backs
- Built-in transponder
  - Short or Long reach versions

OC-48/STM-16 EPC

Here is an overview of the 8 port OC-48EPC features:

- Two-slot module, 8 Ports
  - Multiplexes 4 OC-12 or 8 OC-3 or a mix of OC-3/OC-12 into an OC48
- Software selectable port speeds
- Tributaries support IR-1 reach
  - 15 km to the terminal device
- Traffic aggregated on a single wavelength
- Synchronization
  - External
  - Internal
  - Line
- SONET/SDH performance monitoring
- Loop-backs
  - Terminal and line
- Built-in transponder
  - Short or Long reach version

OC-48/STM-16 EPC

Here is an overview of the 16 port OC-48EPC features:

- Three-slot module, 16 Ports
  - Multiplexes 16 OC-3 or 4 OC-12 or a mix (OC-3/OC-12) into an OC-48
- Software selectable port speeds
- Tributaries support IR-1 reach
  - 15 km to the terminal device
- Traffic aggregated on a single wavelength
- Synchronization
  - External
  - Internal
  - Line
- SONET/SDH performance monitoring
- Loop-backs
  - Terminal
  - Line
- Built-in transponder
  - Short or Long reach version

OC-192/STM-64 EPC

Here is an overview of the OC192 EPC features:

- Two-slot module, 4 Ports
  - Multiplexes 4 OC-48 into an OC-192
- Tributaries support field upgraded transceivers
  - IR-1 (up to 20 km)
  - SR-1 (up to 2 km)
  - 850 nm, multi-mode
- Carries SDH traffic transparently
- SONET/SDH
- Traffic aggregated on a single wavelength
- Synchronization
  - External
  - Internal
  - Line

- SONET/SDH performance monitoring
- Loop-backs
  - Terminal
  - Line
- Forward Error Correction (FEC)

GigaBit Ethernet EPC

Here is an overview of the GigaBit Ethernet EPC features:

- Two-port full rate module
  - Multiplexes 2 full-rate GigE signals into an OC-48
- Four-port fractional rate module
  - Multiplexes 4 fractional GigE's into an OC-48
  - Individual port speeds selectable from 50 Mbps to 1 Gbps
- Gigabit Ethernet ports
  - Field installed transceivers
    - 1310 nm, 2 km
    - 1310 nm, 10 km
    - 850 nm, 500 m
- MAC layer flow control
- RMON and SDH statistics
- Loop-backs
- Built-in transponder
  - Short or Long reach versions

10.8.3 Typical SAN applications

To show the savings that can be made just in terms of fibers, we have two examples that show a "before" and "after" scenario.

In the "before" example, we are using a total of 1536 fibers and this is shown in Figure 10-18.

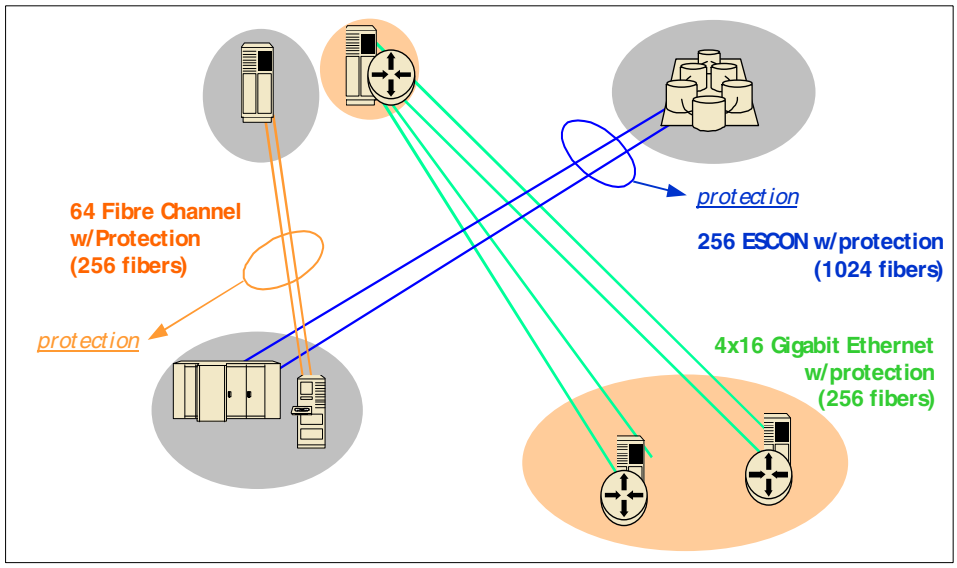


Figure 10-18 Before total fibers used = 1536

By utilizing GigaMux and EPC, we are able to save 1530 fibers as shown in Figure 10-19.

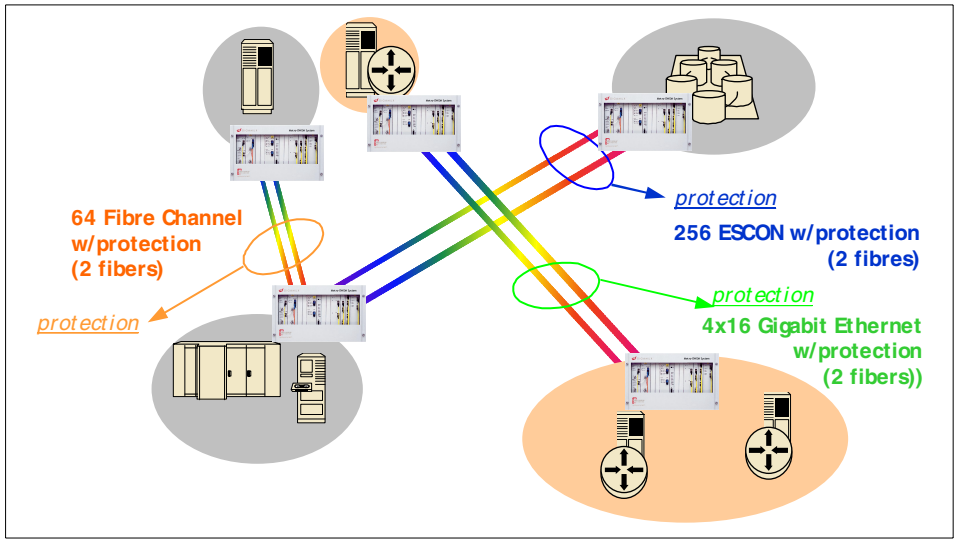


Figure 10-19 After — total fibers used = 6

We have now reduced the amount of fiber needed from 1536 to 6 by using a combination of GigaMux and EPC.

Part 2

Solutions

In this part we take the components and technology that we have introduced and put them all together to build a robust distance solution.

11

IBM TotalStorage SAN Switch distance solutions

In this chapter we will discuss some solutions that can be implemented using the IBM TotalStorage SAN Switch products included in the IBM portfolio. We will combine these with some of IBM's and other vendors' DWDM solutions to describe high distance SAN solutions.

For the features and function of the IBM TotalStorage SAN Switch, refer to:

- *IBM SAN Survival Guide*, SG24-6143
- *IBM SAN Survival Guide Featuring the IBM 2109*, SG24-6127
- *Implementing an Open IBM SAN,* SG24-6116

In this chapter we present various implementations and uses for a SAN in high availability graphically dispersed environments. The considerations for a high availability SAN design based on the IBM TotalStorage SAN Switch are covered in the following topics.

By reducing or eliminating single points of failure in the enterprise environment, SANs can help to improve overall availability of business applications. By utilizing highly available components and solutions as well as a fault-tolerant design, enterprises can achieve the availability needed to support 24x7 uptime requirements.

High distances in these solutions enables the SAN to cover a whole new host of IT availability issues, for example, we can deploy disaster tolerant solutions.

In networked systems, such as SANs, with their associated servers, fabric, and storage components, as well as software applications, downtime can occur even if parts of the system are highly available or fault tolerant. To improve business continuance under a variety of circumstances, SANs can incorporate redundant components, connections, software, and configurations to minimize or eliminate single points of failure.

With the emergence of the Internet and the proliferation of global e-business applications, more and more companies are implementing computing infrastructures specifically designed for continuous data and system availability. Today, even applications, such as company e-mail, have become mission critical for ongoing business operations. Faced with increased customer and internal user expectations, companies are currently striving to achieve at least 99.999 percent (the five “nines”) availability in their computing systems — a figure equivalent to less than 5.3 minutes of downtime a year. Additional downtime can severely impact business operations and cost valuable time, money, and resources.

To ensure the highest level of system uptime, companies are implementing reliable storage networks capable of boosting the availability of data for all the users and applications that need it. These companies typically represent the industries that demand the highest levels of system and data availability — the utilities and telecommunications sector, brokerages and financial service institutions, and a wide variety of service providers.

11.1 High-level availability objectives

Developing highly available networks involves identifying specific availability requirements and predicting what potential failures might cause outages. The first step is to clearly define availability objectives, which can vary widely from company to company and even within segments of the same company. In some environments, no disruption can be tolerated, while other environments might be only minimally affected by short outages. As a result, availability is a function of the frequency of outages (caused by unplanned failures or scheduled maintenance and upgrades) and the time to recover from such outages.

Many companies are addressing their availability requirements by implementing networked fabrics of Fibre Channel devices designed to provide high-performance storage environments. These flexible SANs are based on the following principles:

- A thorough understanding of availability requirements throughout the enterprise
- A flexible design that incorporates fault tolerance through redundancy and mirroring
- Simplified fault monitoring, diagnostics, and repair capabilities to ensure fast recovery
- A minimal amount of human intervention required during failover events
- A reliable backup and recovery plan to account for a wide variety of contingencies

Note: The IBM 2109 family of switches (S08, S16 and F16) have two and four ASICs respectively, each of these ASICs has four ports assigned to it. The buffer credit allocation is on an ASIC level, so four ports will share a pool of buffer credits. This is a key consideration when introducing distance and it should be the aim of a low level design to spread the allocation of long distance ports among the ASICs for that switch.

All long distance links within the fabric must be configured as extended fabric links in order to obtain optimum I/O performance. The correct extended fabric level must be chosen and long distance is classified in this switch at three levels. We describe these in Table 11-1.

Table 11-1 Extended fabric settings

Distance in km	Extended fabric
0 to 10	Level 0
11 to 50	Level 1
51 to 100	Level 2

Extended fabric is an optional license and must be installed before long distance connectivity is enabled. Refer to IBM technical support for information on acquiring and installing the license.

To ensure systems can avoid or withstand a variety of failures, SANs incorporate a wide range of capabilities, including:

- Highly available components with built-in redundancy and hot-plugging capabilities
- No single points of failure
- Intelligent routing and rerouting
- Dynamic failover protection

- Non-disruptive server and storage maintenance
- Hardware zoning for creating safe and secure environments
- Predictive fabric management
- Extended fabrics concepts, as deployed by the IBM 2109 switch family.

Here we describe the IBM 2109 with a multi-node DWDM configuration that spans four sites and provisions optical services.

There are four switches, with each switch's E_Ports connected over a DWDM channel that includes dual paths for transmitting and receiving. This is shown in Figure 11-1.

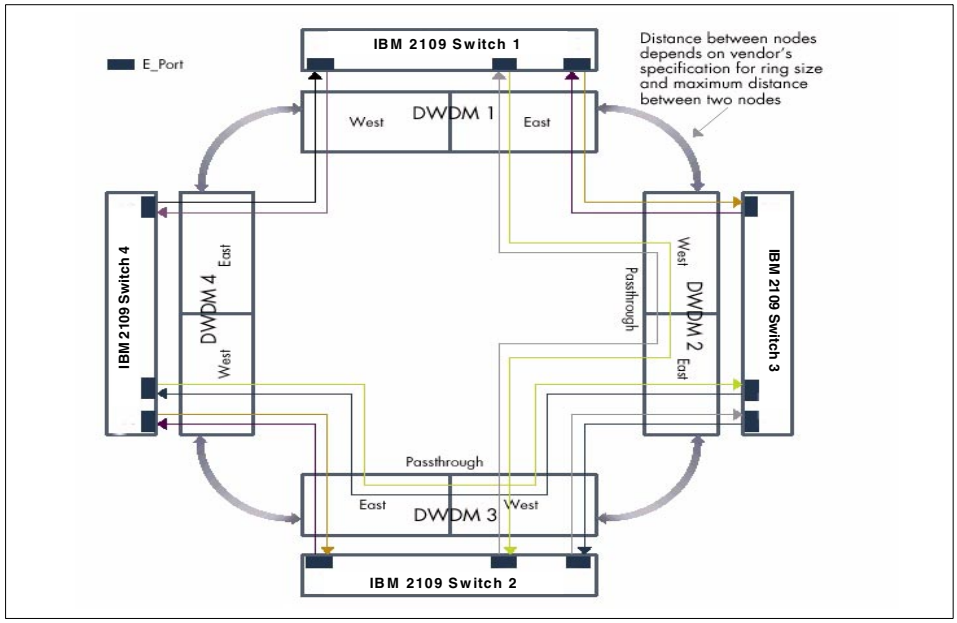


Figure 11-1 Extended fabric over DWDM ring

Each path has its own wavelength. The DWDM pass through feature enables non-contiguous sites to connect over an intermediate site as if they were directly connected. The only additional overhead of the pass through is the minimal latency (5 microseconds/km) of the second link. The pass through has no overhead since it is a passive device. This fabric would logically appear as a fully meshed topology.

Each of the links can operate in protected mode, which provides a redundant path in the event of a link failure. In most cases, link failures are automatically detected within 50 ms. In this case, the two wavelengths of the failed link reverse directions and reach the target port at the opposite side of the ring. If the link between DWDM 1 and 4 fails, the transmitted wavelength from 4 to 1 would reverse direction and reach 1 through 3 and 2. The transmitted wavelength from 1 to 4 would also reverse direction and reach 4 through 2 and 3.

Calculating the distance between nodes in a ring depends on the implementation of the protected path scheme. For instance, if the link between DWDM 2 and 3 fails, the path from 1 to 3 would be 1 to 2, back from 2 to 1 (due to the failed link), 1 to 4, and finally 4 to 3. This illustrates the need to utilize the entire ring circumference (and more, in a configuration with over four nodes) for failover.

Another way to calculate distance between nodes is to set up the protected path in advance (in the reverse direction) so the distance is limited to the number of hops between the two nodes. In either case, the maximum distance between nodes determines the maximum optical reach.

An example of this specification is 80 to 100 km for a maximum distance between nodes, and 160 to 400 km for maximum ring size. These distances should continue to increase as fiber optic technology advances.

11.2 Disk consolidation with a remote disk

This remote disk consolidation solution is an extension of the most basic SAN solutions of disk consolidation; see Figure 11-2.

Disk consolidation provides a means for storage managers to take advantage of the capabilities of Fibre Channel fabric-based Storage Area Networks to organize storage in more flexible and efficient ways, easing many aspects of storage administration.

With IBM 2109 extended fabric functionality, extended long wave GBIC's it is possible to achieve distances up to 100 km. However, testing your systems and applications in this architecture is highly recommended.

This is an attempt to extend the distance between the storage and the servers, in order to provide business benefits such as consolidation, physical separation of disk storage from the server across sites.

This solution extends the distance up to 10 km apart, using long wave GBICs in the SAN fabric, and single mode fiber cable interconnect. The SAN fabric is comprised of:

- IBM 2109 - S08 8 Port FC switch
- IBM 2109 - S16 16 Port FC switch
- IBM 2109 - F16 16 port 2Gb switch (today the connections to the ESS would still be at 1Gb)

The host servers are connected to the SAN fabric via SW GBIC's using MM Fibre cable.

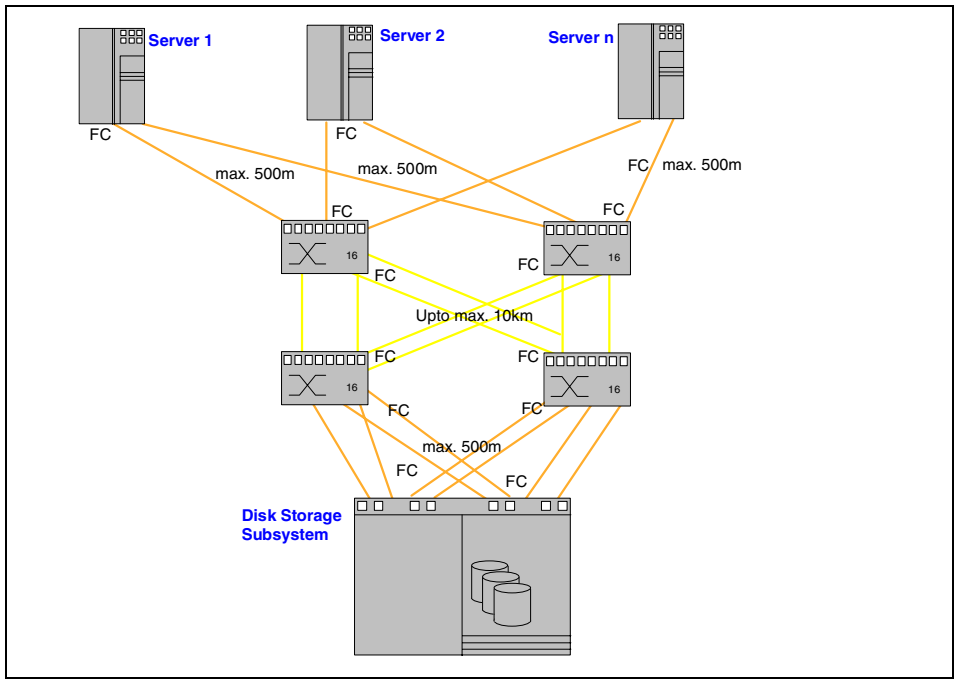


Figure 11-2 Remote disk consolidation

Checklist

This is a list of many actions and items that will need to be considered for this implementation:

- Installing Fibre Channel host bus adapters
  - Servers will need Fibre Channel adapters — adding and configuring
- Installing host software
  - Install SDD for dual path connectivity
- Installing ESS Software on appropriate workstations — not specified here
  - IBM StorWatch Specialist

  - IBM ESS Expert Software
- Configuring storage
  - Define hosts
  - Allocate LUNs
  - Assign LUNs to hosts
- Connectivity
  - Attaching ESS to switches
  - Attaching servers to switch
- Validate connectivity
  - Test server to storage on both paths
  - Tests of failover/failback operations
- List of firmware and driver revisions of servers, HBA, storage, switches and update where necessary
- The IBM TotalStorage SAN Switch Specialist software version must be compatible with the ESS firmware level

Performance

In this simple implementation, performance will depend on the number of HBAs available on each server and the number of storage connections and most important the distance between storage and servers. It is important to consider application behavior over the distance, and this varies from application to application and we have assumed that the performance is adequate for applications running in the servers.

The IBM 2109 switch family supports any to any connectivity so it will not affect performance by itself. It is important to profile your applications to understand their I/O characteristics. This will enable you to make informed choices about distance and bandwidths.

Without having the exact requirements we may consider that for a high profile server, a server to storage ratio of 6:1 is acceptable. This is only a starting point, we will then implement some measurement system or use statistical data to decide whether we need to add more connections or whether we have more bandwidth than required.

Scalability

The IBM 2109 switch family supports the concurrent addition of more switches, so we can scale this solution by adding more servers or storage devices without disrupting operation, although we may cause a fabric initialization process to happen briefly while the fabric integrates any new switches added. With four ISL between the switches on the host side and the switches on the storage side, we can have 26 servers with two HBAs each, and still keep four spare ports in a fully populated switch.

Security

Here are some security considerations:

- Disk Storage ESS LUN masking by WWN will allow each server access only to configured LUNs
- IBM TotalStorage SAN Switch Specialist user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions. There are currently two levels to this: administrator and user.
- Physical Switch security is recommended, for example, locked cabinet, restricted access site
- “What if” failure scenarios
- These are some theoretical assumptions:
  - Host HBA failure — SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN.
  - ESS host adapter failure — The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters, a single adapter failure will reduce available bandwidth by 25%. For an average workload and five servers as shown it should not impact performance. When the host adapter is replaced we need to update zoning with the new WWN.
  - Switch port failure — The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or ESS Host adapter failure. The cable can be moved to a spare port. AIX and SDD will have to be re-configured to pickup the new path information if it was a storage port. Physical access to switch and EFC Manager user with maintenance rights are required. AIX root access may be required.

– Fiber failure — Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to switch and attached device are required.

11.3 Two sites at 10 km apart

One can assume that a greater distance between IT sites results in greater security against wide spread disaster. However, increased distance has its price in terms of interconnection cost, business relocation effort, and network cost. Before you start to expand your storage network, you have to consider if the following solution can meet your business purpose.

The solution as shown in Figure 11-3, can be used for accessing a database at remote site as well as mirroring data at an alternative site over greater distances.

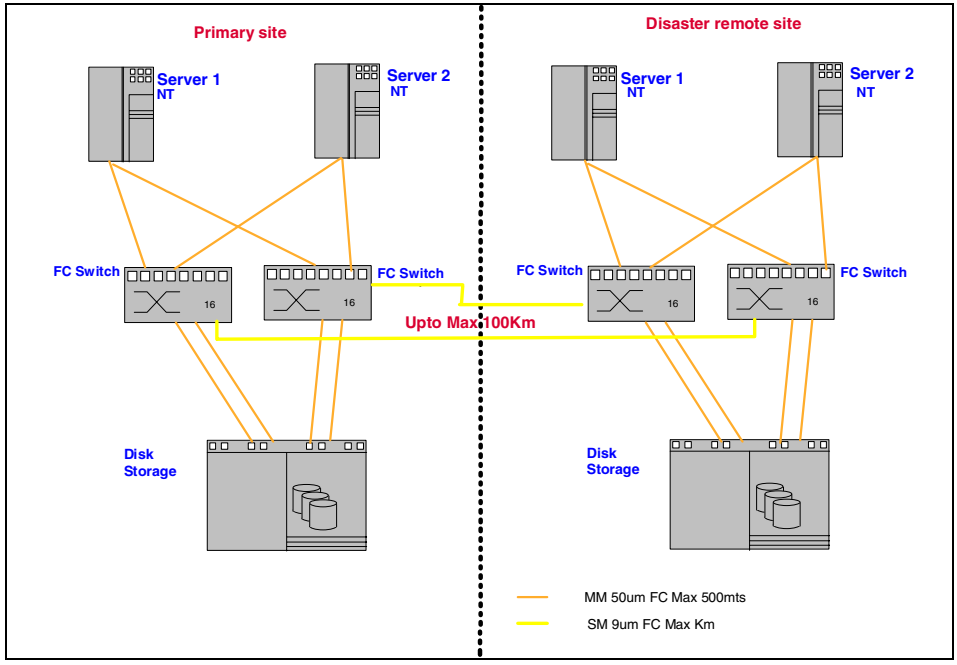


Figure 11-3 Two sites, 10 km apart

We have assumed for our solution that:

- Both sites have SAN infrastructure.
- Server to IBM 2109, IBM 2109 to storage use multiple mode fibre.

- UNIX servers in production which are being used for mission critical production need data mirroring for disaster recovery.
- NT servers at site A need to access the database in storage B at site B, in addition to using data in storage A.
- UNIX servers at site B are in standby mode, and being used for backup servers at ordinary times.
- Multi-mode FC cables are used for connection between servers and IBM 2109, and between IBM 2109 and storage devices. Single mode FC cables are used between IBM 2109s at each site.
- Software mirroring is used in UNIX servers.
- No DWDM and Channel Extender proposed.
- No PPRC proposed.

Checklist

The following items should be considered to implement the solution:

- Fibre channel supported on servers
- Operating systems supported S/W mirroring
- Shortwave HBAs on our servers
- Longwave laser GBICs (enabling a transmission distance of up to 10 kilometers) and extended long-wave GBICs (enabling a transmission distance of up to 100 kilometers) for a maximum of 128 GBICs
- Any application performance implications of longer distances

Performance

The major performance consideration with a long distance solution is calculating the correct number of lines between sites. This number can only be determined by performing a detailed performance profile of the servers and storage that will be remote. Over-estimating the number of lines will increase costs dramatically, under sizing the number of lines will dramatically effect the performance of the SAN.

It is vital that detailed performance data is available prior to sizing the number of lines required.

Typically, latency will increase over long distances, a good rule of thumb is 4.8 microseconds per kilometer. Brocade have performed comprehensive testing over distances in excess of 100 km and have found no performance implications with the switch.

The default E_D_TOV and R_A_TOV values do not need to be modified for this distance.

Scalability

Refer to our scaling table, Table 11-1, “Extended fabric settings” on page 307.

Security

Our solution assumes we own private lines between sites so encryption is not required.

For leased lines or managed services where lines are shared, encryption is normally an option available from the service provider.

“What if” failure scenarios

Here are some theoretical assumptions:

- If the primary route is further than the secondary — This may have performance implications and will need verification.
- If a normal HBA is used — You will only be able to locate servers 500 meters from the switch.

11.4 Multiple site - ring topology DWDM solution

In Figure 11-4 we show a high availability SAN design for multiple servers deployed with dual switch and redundant fabric at each site.

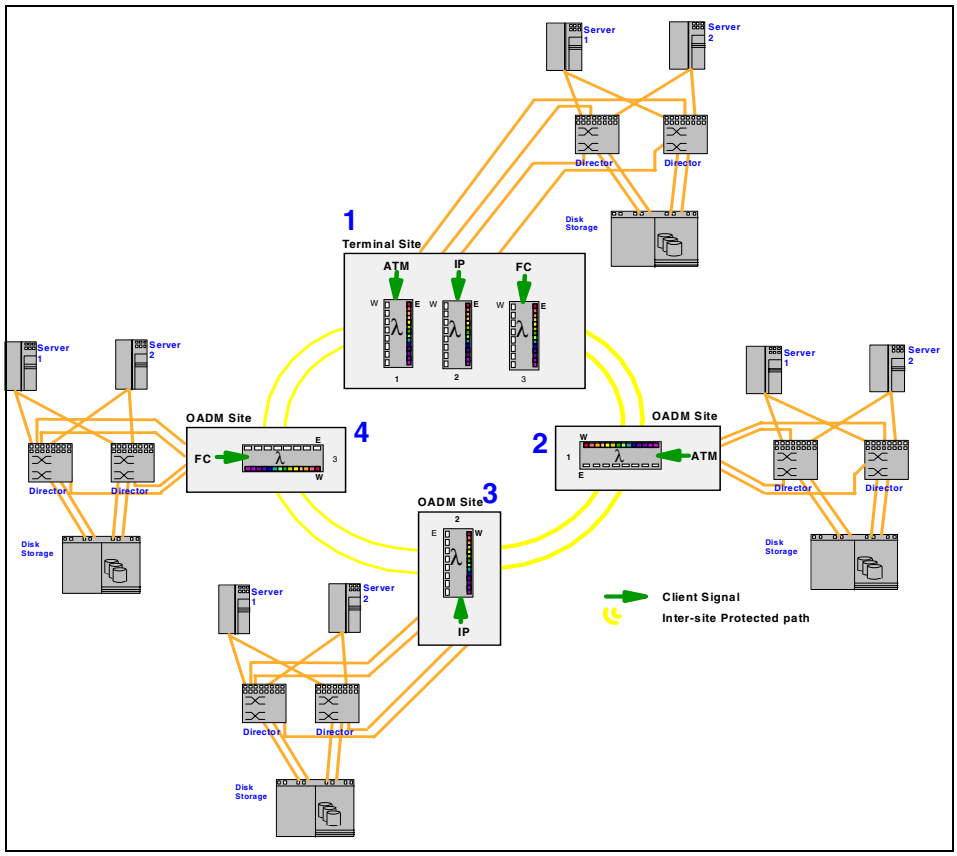


Figure 11-4 Multiple site ring topology SAN

This design will support a small number of servers at each site, however connectivity will also be available to other locations via the DWDM across a MAN. These switches uses a DWDM solution for the inter-site ISL connections.

This remote disk consolidation solution is an extension of the most basic SAN solution of disk consolidation.

Disk consolidation provides a means for storage managers to take advantage of the capabilities of Fibre Channel fabric-based Storage Area Networks to organize storage in more flexible and efficient ways, easing many aspects of storage administration.

This is an attempt to extend the distance between the storage and the servers, in order to provide business benefits like consolidation, physical separation of disk storage from the server across sites.

With the IBM 2109 extended fabric functionality, extended long wave GBIC's, and the DWDM infrastructure it is possible to achieve distances up to 100 km. However, testing your systems, and applications in this architecture is highly recommended.

The SAN fabric is comprised of:

- IBM 2109 - S08 8 Port FC switch
- IBM 2109 - S16 16 Port FC switch
- IBM 2109 - F16 16 port 2Gb switch (today the connections to the ESS would still be at 1Gb)

The host servers are connected to the SAN fabric via SW GBIC's using MM Fibre cable.

Checklist

This is a list of many actions and items that will need to be considered for this implementation:

- Installing Fibre Channel host bus adapters
  - Servers will need Fibre Channel adapters — adding and configuring
- Installing host software
  - Install SDD for dual path connectivity
- Installing ESS Software on appropriate workstations — not specified here
  - IBM StorWatch Specialist
  - IBM ESS Expert Software
- Configuring storage
  - Define hosts
  - Allocate LUNs
  - Assign LUNs to hosts
- Connectivity
  - Attaching ESS to switches
  - Attaching servers to switch
- Validate connectivity
  - Test server to storage on both paths
  - Tests of failover/failback operations
- List of firmware and driver revisions of servers, HBA, storage, switches and update where necessary
- The IBM StorWatch Specialist software version must be compatible with the ESS firmware level

Performance

In this simple implementation, performance will be depend on the number of HBAs available on each server and the number of storage connections and most important the distance between storage and servers. It is important to consider application behavior over the distance, and this varies from application to application and we have assumed that the performance is adequate for applications running in the servers.

The IBM 2109 switch family supports any to any connectivity so it will not affect performance by itself. It is important to profile your applications to understand their I/O characteristics. This will enable you to make informed choices about distance and bandwidths.

Without having the exact requirements we may consider that for a high profile server, a server to storage ratio of 6:1 is acceptable. This is only a starting point, we will then implement some measurement system or use statistical data to decide whether we need to add more connections or whether we have more bandwidth than required.

Scalability

The IBM 2109 switch family supports the concurrent addition of more switches, so we can scale this solution by adding more servers or storage devices without disrupting operation, although we may cause a fabric initialization process to happen briefly while the fabric integrates any new switches added. With four ISL between the switches on the host side and the switches on the storage side, we can have 26 servers with two HBAs each, and still keep four spare ports in a fully populated switch.

Security

Here are some security considerations:

- Disk Storage ESS LUN masking by WWN will allow each server access only to configured LUNs.
- IBM StorWatch Switch Specialist user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions. There are current two levels to this: administrator and user.
- Physical Switch security is recommended for example locked cabinet, restricted access site.
- “What if” failure scenarios.
- Here are some theoretical assumptions:

- Host HBA failure — SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced, the zoning information and ESS host definition will have to be updated with the new WWN.
- ESS host adapter failure — The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce available bandwidth by 25%. For an average workload and five servers as shown it should not impact performance. When the host adapter is replaced we need to update zoning with the new WWN.
- Switchport failure — The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or ESS Host adapter failure. The cable can be moved to a spare port. AIX and SDD will have to be reconfigured to pickup the new path information if it was an storage port. Physical access to switch and EFC Manager user with Maintenance rights are required. AIX root access may be required.
- Fiber failure — Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to switch and attached device are required.

DWDM considerations

We have shown a ring topology here; this gives a logical mesh, potentially giving any to any connectivity. We have implemented this as two rings, one ring connects to one switch in each site, the other ring connects to the remaining switch. This gives us two discrete SAN fabrics.

We would implement the DWDM solution as a protected ring to ensure availability. In this example, we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation and here we have assumed a light workload.

We show here an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one switch in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits. However, this is not unique to DWDM, more so to the general SAN solution over distance. The DWDM is a core architecture deployment and because of its independence from protocol, it can be used for lots of other traffic; that is to say, it extends beyond the SAN environment. The distance we have shown here is 25 km between nodes.

An example of ring specification is 80 to 100 km for a maximum distance between nodes and 160 to 400 km for maximum ring size. These distances should continue to increase as fiber optic technology advances.

11.5 Two sites: channel extender and WAN extension

Channel extenders typically use telecommunication lines for data transfer and therefore enable application and recovery sites to be located over longer distances apart. The use of channel extenders provides the separation for disaster recovery purposes and avoids some of the barriers imposed when customers do not have a “right of way” to lay their fiber cable.

In Figure 11-5 we show a typical SAN distance extension using Optical Channel extenders at both primary and secondary sites and the ATM network.

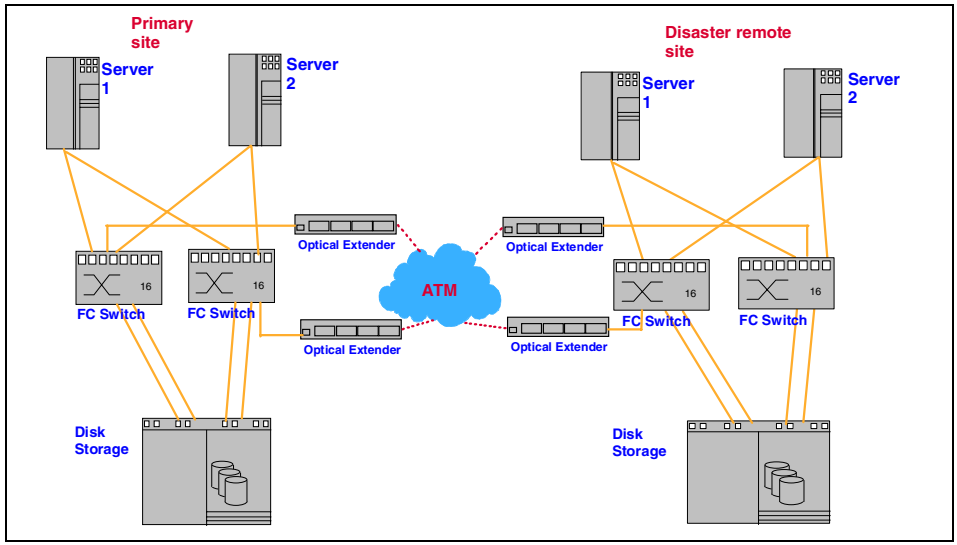


Figure 11-5 SAN extension over (ATM) WAN using channel extenders

The SAN fabric is comprised of:

- IBM 2109 S08 FC switch

- IBM 2109 S16 FC switch
- IBM 2109 F16 FC switch

The host servers and storage are connected to the SAN fabric via SW GBIC's using MM Fibre cable.

We assumed some considerations as follows:

- The data movement/copy from primary to secondary site is done at the OS level in the host or by storage subsystem if it supports it.
- The ATM network is installed, configured and available. The ATM network technical details are beyond the scope of this example.
- As with any extended solution "distance = some performance sacrifice". Any individual transaction will be effected by the latency. Latency typically needs to be validated based on the specific end-to-end devices deployed. An extended distance configuration that worked with one server/application might fail with time-outs in another.

Checklist

We checked the following items:

- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and microcode level for proposed configuration
- IBM switches SW GBIC count
- MM FC cable laying and termination
- Nickname assignments so we can quickly cross reference WWNs to devices
- I/O interface (for example, FC or ESCON) selection and installation on the Optical channel extender
- Network interface (for example, ATM OC3) selection and installation on the Optical channel extender
- Pre-installation testing of the network interfaces (ATM network)

Performance

The major performance consideration with a long distance solution is calculating the correct number of interconnecting links between sites. This number can only be determined by performing a detailed performance profile of the servers, storage that will be remote. It is vital that detailed performance data is available prior to sizing the number of interconnecting links required.

Channel extenders generally compress the data before sending it over the transport network, however the compression ratio needs to be determined based on the application characteristics and the distance.

You must especially consider an amount of the updated data for a period of time (peak time), and reflect them on calculating the number of interconnecting links and the number of data volumes for a SAN/WAN solution.

Scalability

Here are some scalability points:

- Most optical channel extenders support multiple I/O interfaces for client equipment. However in a installation, it is recommended to have a single protocol interfaces on a physical channel extender.
- Scalability rules for the SAN fabric of IBM switches remain unchanged as in previous examples.

Security

Here are some assumptions about security:

- Disk Storage LUN masking by WWN will allow each server access only to configured LUNs.
- IBM StorWatch Switch Specialist user IDs, passwords and rights are defined and defaults removed so only authorized personnel can perform management functions.
- Physical switch security — locked cabinet, restricted access site.
- The physical security of fiber connections and patch panels should be considered.

“What if” failure scenarios

Here are some theoretical assumptions:

- **Host HBA failure**

  SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. EFC Manager user with Product Administrator rights and ESS Specialist access are required.

- **Storage host adapter failure**

  The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters, a single adapter failure will reduce available bandwidth by 25%. For an average workload and five servers as shown it should not impact performance. When the host adapter is replaced we need to update zoning with the new WWN. We also need to reconfigure OS and SDD to pickup the new path information. EFC Manager user with Product Administrator rights and OS root access are required.

- **Switch port failure**

  The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port.

- **Fiber failure**

  Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to switch and attached device are required.

- Switch completely down, storage completely down, or site down (power, air conditioning, site damage) will cause an interruption in normal operation.

- Physical damage to storage causing data loss (fire, flood). We will need to restore data from backup tapes.

- **Channel extenders failure**

  In the current solution, channel extenders are not configured in a redundant mode, however they can be.

- **Transport network failure**

  The telco link failure is not planned for, in the current solution; however, a redundant link can be acquired from the vendor or a separate telco may be considered depending on the business recovery policy.

11.6 Remote tape vaulting

For an existing large corporation with multiple sites, tape library resources can be consolidated to a separate site. This simplifies data movement logistics and centralizes backup software configurations. By doing this the cost of doing business is reduced as the infrastructure is efficiently utilized and less IT personnel required for backup data management. In addition, in the event of a disaster, the data is already located on tape in remote location and there is no

longer a need to ship the data to another site. Another application is for outsourcing services companies like storage service providers (SSP), that are interested in providing backup and backup management services to smaller companies who may not have the infrastructure or need it based on their business recovery policy.

In Figure 11-6 we show the basic layout for a remote tape vaulting solution, using the IBM TotalStorage SAN Switch.

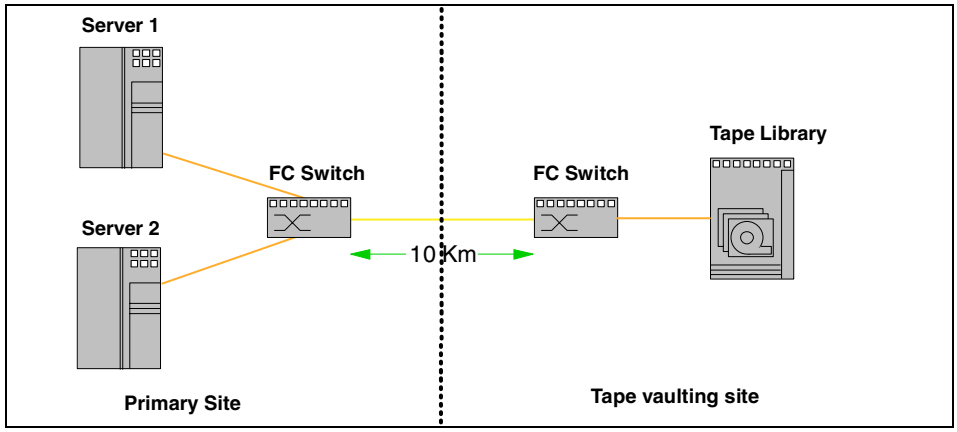


Figure 11-6 Remote tape vaulting

This solution extends the distance up to 10 km apart, using Long wave GBICs in the SAN fabric, and Single mode fiber cable interconnect. Distance up to 80 km are possible with Finisar repeaters.

Note: Two 2109s connected via longwave GBICs can be placed 10 km apart.

There will be no change to the longwave GBIC distances when the 2Gb SAN technology is available.

Architecturally, the Fibre Channel connection extensions could be provided by dark fibre, Fibre Channel repeaters, or dense wave division multiplexing (DWDM) devices. Dark fibre is fibre provided by one of the telcos with the telco providing the repeaters.

The host servers are connected to the SAN fabric via SW GBIC's using MM Fibre cable. In the case of the IBM TotalStorage SAN Switch, the distance is limited to 10 km.

Most tape libraries Fiber Channel ports run in Fibre Channel Arbitrated Loop (FC-AL). They can be directly attached to the IBM switch, because they automatically sense the port type and it will be assigned as FC-AL.

Zoning can be established by WWN so we can limit access of each server or each group of servers to specific tape drives.

In order to avoid human errors that can affect operation of other servers, zone changes should only be performed by designated personnel, and proper procedures must be in place to make sure that operations personnel are aware of the available devices to each server according to the zones currently active.

Checklist

In addition to the items considered in the disk consolidation example we must now consider:

- Host HBA supported for tape drive/library attachment
- Host tape device driver levels
- Host operating system levels compatible with tape library fiber requirements
- Zone configuration allowing tape access to required servers
- Host software tape sharing capabilities
- Switches LAN connection, firmware and licence code where appropriate

Performance

Depending upon the drive interface LTO or Magstar, the performance will vary.

Performance basically will depend on the number of HBAs available on each server and the number of storage connections and most important the distance between storage and servers. It is important to consider application behavior over the distance, and this varies from application to application and we have assumed that the performance is adequate for applications running in the servers.

Scalability

Potentially we can scale the solution in terms of adding additional tape drives by also increasing the number of additional IBM 2109 switches. However, physical space and cost should be considered.

Security

Here are the security assumptions:

- Zoning can be used to restrict access to devices to specific servers when required. This may also be used to change workload profiles to tape drives,

for example, over night may be when only the backup servers can see all the drives.

- Proper tape management procedures will avoid servers contending for the same tape device. You would need to size the backup window available and the amount of data that you need to backup.

"What if" failure scenarios

Here are some theoretical assumptions:

- **ISL or switch failure**

  One path to all tape devices in the pair of switches lost. Access available through the other switch. Performance may be impacted depending on number of drives attached. Traditionally, tape failover is a manual operation. Multiple path devices are configured as several logical devices, one per path. Only one of these logical devices is made active. If there is a failure the application aborts and it can then be restarted using a different logical device. Latest levels of a tape device driver provide alternate pathing support and tape failover for Fibre Channel connections. With this support enabled if an error occurs the device driver will automatically initiate error recovery and the operation will continue using the next logical path.

- **Device link or device port failure**

  Only one tape drive is affected. Alternate path remains operational. Recovery may be manual or automatic depending on operating system and driver level as explained for ISL or switch failure.

- **Switch port failures**

  GBICs are hot swappable. H_Ports can be moved to a spare port.

- **User error trying to access more than two drives on the same link**

  Performance of all drives attached to the switch pair may be degraded. Switch or switches performance view may be used to find out what paths are carrying traffic.

- **Tape drive failure in a single tape zone**

  An alternate zone should be made active to get access to a working device.

11.6.1 Remote tape vaulting with disaster tolerance

Figure 11-7, shows an extension or variant of the tape vaulting solution. In this solution the primary site/local site has a tape library. Data can be written to one or both tape libraries, however in the event of failure on the tape library at the local site, data can be backed up and restored from the tape library at the remote site.

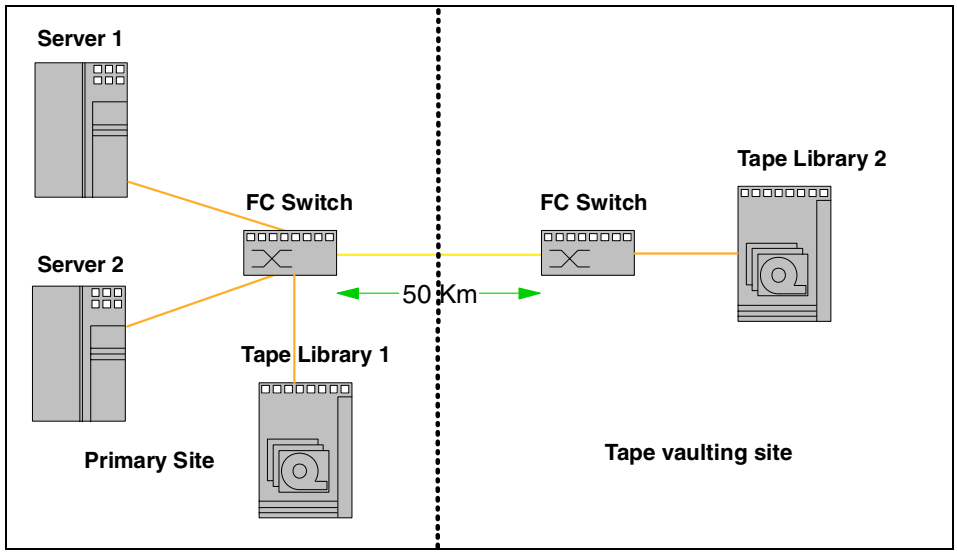


Figure 11-7 Remote tape vaulting with disaster tolerance

Checklist

In addition to the items already considered, we must now consider:

- Host HBA supported for tape drive/library attachment
- Host tape device driver levels
- Host operating system levels compatible with tape library fiber requirements
- Zone configuration allowing tape access to required servers
- Host software tape sharing capabilities
- Switches LAN connection, firmware and licence code, where appropriate

Performance

Depending upon the drive interface LTO or Magstar, the performance will vary. Performance will basically depend on the number of HBAs available on each server and the number of storage connections and most important the distance between storage and servers. It is important to consider application behavior over the distance, and this varies from application to application and we have assumed that the performance is adequate for applications running in the servers.

Scalability

Potentially we can scale the solution in terms of adding additional tape drives by also increasing the number of additional IBM 2109 switches. However, physical space and cost should be considered.

Security

Here are the security issues:

- Zoning can be used to restrict access to devices to specific servers when required. This may also be used to change workload profiles to tape drives, for example, over night may be when only the backup servers can see all the drives.
- Proper tape management procedures will avoid servers contending for the same tape device. You would need to size the backup window available and the amount of data that you need to backup.

"What if" failure scenarios

Here are some theoretical assumptions:

- **ISL or switch failure**

  One path to all tape devices in the pair of switches lost. Access available through the other switch. Performance may be impacted depending on number of drives attached. Traditionally, tape failover is a manual operation. Multiple path devices are configured as several logical devices, one per path. Only one of these logical devices is made active. If there is a failure the application aborts and it can then be restarted using a different logical device. Latest levels of a tape device driver provide alternate pathing support and tape failover for Fibre Channel connections. With this support enabled if an error occurs the device driver will automatically initiate error recovery and the operation will continue using the next logical path.

- **Device link or device port failure**

  Only one tape drive is affected. Alternate path remains operational. Recovery may be manual or automatic depending on operating system and driver level as explained for ISL or switch failure.

- **Switch port failures**

  GBICs are hot swappable. H_Ports can be moved to a spare port.

- **User error trying to access more than two drives on the same link**

  Performance of all drives attached to the switch pair may be degraded. Switch or switches performance view may be used to find what paths are carrying traffic.

- **Tape drive failure in a single tape zone**

  An alternate zone should be made active to get access to a working device.

11.7 Two sites: point-to-point DWDM

In Figure 11-8 we show a basic high availability SAN design for a multiple servers deployed with dual switch and redundant fabric at each site.

This design will support a small number of servers. These switches use a DWDM solution for the inter-site ISL connections, IBM 2109 extended fabric functionality, extended long wave GBIC's, and by implementing the DWDM infrastructure it is possible to achieve distances up to 100 km. However, testing your systems and applications in this architecture is highly recommended.

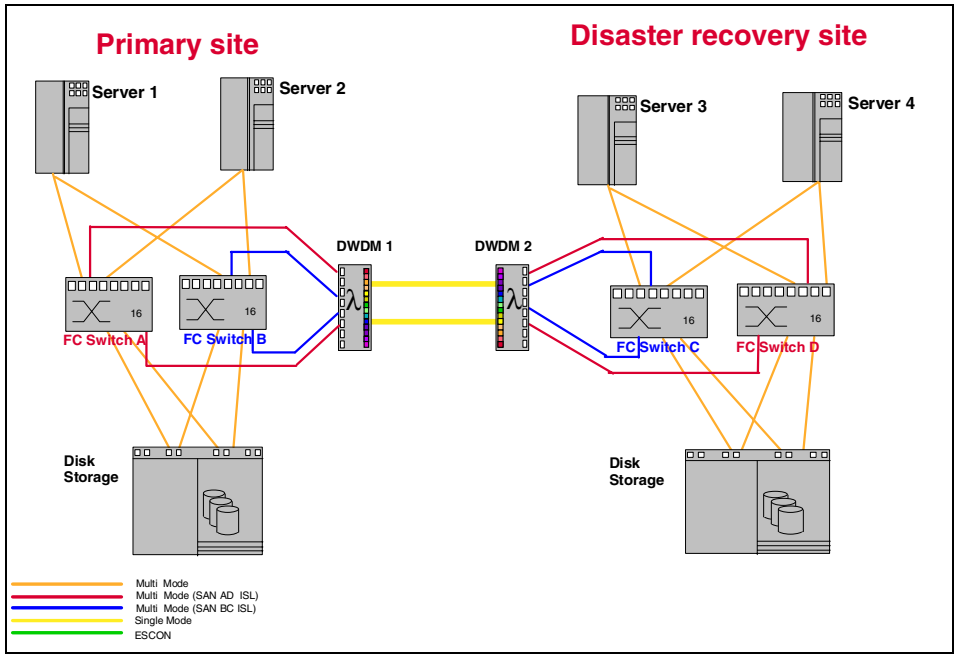


Figure 11-8 Dual switch with two redundant fabrics at two sites

A dual fabric SAN is a topology where you have two independent fabrics that connect the same hosts and storage devices, which are mutually exclusive. This design is not highly scalable as all hosts and storage must be connected to both switches to achieve high availability, however it does allow for inter site traffic. We have labeled each site here, one is the primary site, the other the DR site, as these names imply this infrastructure would give the transport mechanism to

enable off-site recovery. Maybe with hot standby servers to reduce recovery times. The method for replicating the data could be performed at a host level, for example, disk mirroring or at a storage level. The data would be replicated by host software in this example.

Within each site server to storage availability is provided with this dual switch, dual fabric design within each site. Dual HBAs are installed in each host and each storage device must have at least two ports. Failover for a failed path or even a failed switch is dependent on host failover software, namely, IBM Subsystem Device Driver (SDD).

In our example, SAN AD (switches A and D) will be discrete from SAN BC (switches B and C). There are two fabrics here and each switch is a single-switch fabric.

The components used are:

- SAN Fabric
  - Four 16-port IBM SAN Fibre Channel Switch Model 2109-S16
- DWDM
  - Two DWDM's linked via dark fiber
- Servers
  - Four IBM Netfinity Servers running on MS Windows NT, two Netfinity PCI Adapters and four HBA cards
- Storage
  - Two ESS 2105-F20 with native Fibre Channel Adapter
- Software
  - IBM Subsystem Device Driver (SDD)
  - IBM StorWatch Specialist
  - IBM ESS Expert (for historical disk performance data)

Checklist

We checked the following items:

- Installing Fibre Channel host bus adapters
  - Servers will need Fibre Channel adapters — adding and configuring
  - ESS may need Fibre Channel adapters depending on configuration
- Installing servers systems
  - Install SDD for dual path connectivity
- Installing ESS Software on appropriate workstations — not specified here

  - IBM StorWatch Specialist
  - IBM ESS Expert Software
- Configuring storage
- Define hosts
  - Allocate LUNs
  - Assign LUNs to hosts
- Connectivity
  - Attaching ESS to switches
  - Attaching servers to switch
  - Attach Switches to DWDM
- Validate connectivity
  - Test server to storage on both paths
  - Tests of failover/failback operations
  - Test inter-site connectivity via DWDM
  - Setup LUNs on back up ESS and test from primary server, to quantify latency
- List of firmware and driver revisions of servers, HBA, storage, switches and DWDM's update were necessary
- The IBM TotalStorage SAN Switch Specialist version must be compatible with the ESS firmware level

Performance

Typically, for a low performance server, the recommended server to storage connection ratio is 12 to 1; and for a high performance server, the server to storage ratio is 6 to 1. Low performance servers are typically made up of file and print servers, whereas high performance servers are application servers.

To increase the performance of the SAN, multiple connections may be added from the hosts to the switches and from the switches to the storage devices.

Scalability

This design is able to accommodate up to twelve dual path servers at each site. This leaves two spare ports on each switch for future storage additions or increased bandwidth to the existing storage. By adding more switches, it can scale to a larger fabric.

Security

Dual fabrics (the two discrete fabrics, where not all the switches are connected to each other) can protect you against user errors, such as a user erasing or changing the zoning information inappropriately.

The zoning information is separate for each fabric SAN AD and SAN BC, so when changing the zoning in one fabric, it is not automatically propagated into the other fabric.

Should you decide to add an ISL, ensure that all checks are done on the zone configuration changes from the management console.

"What if" failure scenarios

Here are the "what if" scenarios we considered:

- **Server**

  If one of the fibre paths fails within the server, the solution will failover to the second path dynamically.

- **HBA**

  If one of the HBA fails, IBM SDD software will automatically failover the workload to the alternate HBA.

- **Cable**

  If a cable between a server and the switch fails, IBM SDD 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 a performance loss in this solution.

- **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.

- **Switch**

  If a switch fails, the server will use the alternate path to the alternate fabric to connect to the storage.

- **Storage**

  If the ESS fails, the servers will not be able to access the storage. A redundant ESS may be added to mirror data from the primary ESS.

DWDM considerations

We have shown a point-to-point topology here. We would implement the DWDM solution as a protected fiber to ensure availability. In this example we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation and here we have assumed a light workload. We show an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one switch in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits; however, this is not unique to DWDM, more the general SAN solution over distance. The DWDM is a core architecture deployment and because of its independence from protocol it can be used for lots of other traffic, it extends beyond the SAN environment. The distance we have shown here is 80 km between nodes.

11.8 Two sites: point-to-point DWDM with ESS PPRC

In Figure 11-9 we show a basic high availability SAN design for multiple servers deployed with dual switch and redundant fabric at each site.

This design will support a small number of servers. These switches use a DWDM solution for the inter-site ISL connections. The DWDM shown here also facilitates the inter-site ESCON traffic for the PPRC.

.

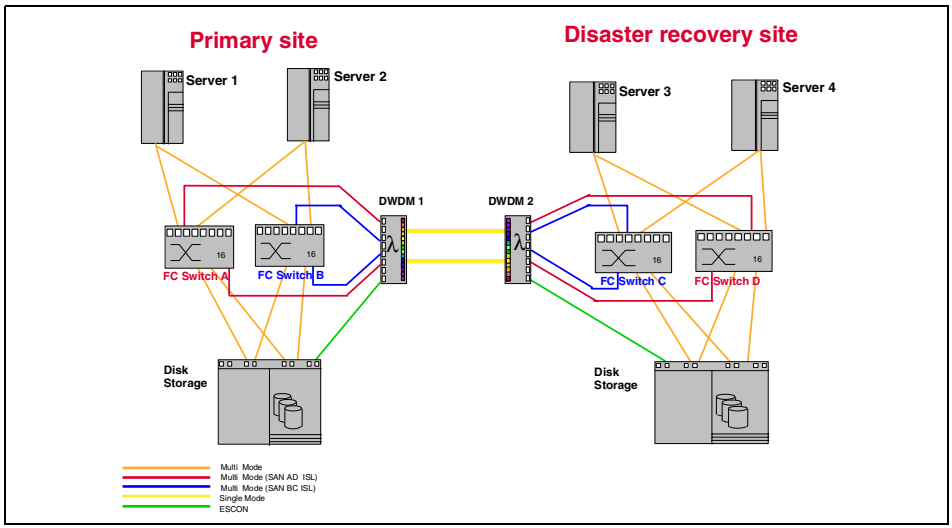


Figure 11-9 Dual switch with two redundant fabrics at two sites

A dual fabric SAN is a topology where you have two independent fabrics that connect the same hosts and storage devices, which are mutually exclusive. This design is not highly scalable as all hosts and storage must be connected to both switches to achieve high availability, however it does allow for inter-site traffic. We have labeled each site here, one is the primary site, the other the DR site. As these names imply, this infrastructure would give the transport mechanism to enable off-site recovery, maybe with hot standby servers to reduce recovery times. The method for replicating the data could be performed at a host level, for example, disk mirroring or at a storage level. This example shows data replication being ESS and PPRC. The PPRC connection is today via ESCON, and this can utilize IBM extended fabric functionality, extended long wave GBIC's, and the DWDM infrastructure. It is possible to achieve distances up to 100 km.

Within each site, server to storage availability is provided with this dual switch, dual fabric design within each site. Dual HBAs are installed in each host and each storage device must have at least two ports. Failover for a failed path or even a failed switch is dependent on host failover software, namely, IBM Subsystem Device Driver (SDD).

In our example SAN AD (switches A and D) will be discrete from SAN BC (switches B and C). Even though there are two fabrics here, each switch is a single-switch fabric.

The components used are:

- **SAN Fabric**

Four 16-port IBM TotalStorage SAN Switch Model 2109-S16

- **DWDM**

  Two DWDM's linked via dark fiber

- **Servers**

  Four IBM Netfinity Servers running on MS Windows NT 4.0/EE SP6a, two Netfinity PCI Adapters and four QLogic 2200 PCI HBAs

- **Storage**

  Two ESS 2105-F20 with native Fibre Channel Adapter, and ESCON adapters for PPRC connectivity

- **Software**

  – IBM Subsystem Device Driver (SDD)
  – IBM StorWatch Specialist
  – IBM ESS Expert (for historical disk performance data)

Checklist

We checked the following items:

- Installing Fibre Channel host bus adapters
  – Servers will need Fibre Channel adapters — adding and configuring
  – ESS may need Fibre Channel adapters and ESCON cards, depending on configuration
- Installing servers systems
  – Install SDD for dual path connectivity
- Installing ESS software on appropriate workstations — not specified here
  – IBM StorWatch Specialist
  – IBM ESS Expert Software
- Configuring storage
- Define hosts
  – Allocate LUNs
  – Assign LUNs to hosts
- Connectivity
  – Attaching ESS to switches
  – Attaching servers to switch
  – Attach Switches to DWDM
  – attaching ESS ESCON HBA to DWDM
- Validate connectivity

  - Test server to storage on both paths
  - Tests of failover/failback operations
  - Test inter-site connectivity via DWDM
  - Setup LUNs on back up ESS and test from primary server, to quantify latency
  - Test PPRC
    - Establish paths
    - Test PPRC functionary - qualify timings
- List of firmware and driver revisions of servers, HBA, storage, switches and DWDM's update were necessary
- The IBM StorWatch Specialist software version must be compatible with the ESS firmware level

Performance

Typically, for a low performance server, the recommended server to storage connection ratio is 12 to 1; and for a high performance server, the server to storage ratio is 6 to 1. Low performance servers are typically made up of file and print servers, whereas high performance servers are application servers.

To increase the performance of the SAN, multiple connections may be added from the hosts to the switches and from the switches to the storage devices.

Scalability

This design is able to accommodate up to twelve dual path servers at each site; this leaves two spare ports on each switch for future storage additions or increased bandwidth to the existing storage. By adding more switches, it can scale to a larger fabric.

Security

Dual fabrics (the two discrete fabrics, where not all the switches are connected to each other) can protect you against user errors, such as a user erasing or changing the zoning information inappropriately. The zoning information is separate for each fabric SAN AD and SAN BC, so when changing the zoning in one fabric, it does not automatically propagate into the other fabric.

Should you decide to add an ISL, ensure that all checks are done on the zone configuration changes from the management console.

"What if" failure scenarios

Here are the "what if" scenarios we considered:

- **Server**

  If one of the fibre paths fails within the server, the solution will failover to the second path dynamically.

- **HBA**

  If one of the HBA fails, IBM SDD software will automatically failover the workload to the alternate HBA.

- **Cable**

  If a cable between a server and the switch fails, IBM SDD 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 a performance loss in this solution.

- **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.

- **Switch**

  If a switch fails, the server will use the alternate path to the alternate fabric to connect to the storage.

- **Storage**

  If the ESS fails, the servers will not be able to access the storage. A redundant ESS may be added to mirror data from the primary ESS.

DWDM considerations

We have shown a point-to-point topology here. We would implement the DWDM solution as protected fibres to ensure availability. In this example we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation and here we have assumed a light workload. We show here an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one switch in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning, or bad level of firmware.

Latency will need to be taken into account, which is also related to buffer credits, however this is not unique to DWDM, more the general SAN solution over distance. The DWDM is a core architecture deployment and because of its independence from protocol, it can be used for lots of other traffic. It extends beyond the SAN environment. The distance we have shown here is 80 km between nodes.

The solution shown here also implements IBM's PPRC between the two ESS subsystems and this will, today, be transported over the ESCON protocol. The ESCON traffic can be transported over the already deployed DWDM architecture.

We show a single pair of 62.5 ESCON connections here, connecting the ESCON ports on the ESS to the DWDM, again this will be dependent upon traffic expectations and should follow the ESCON channel sizing guidelines for PPRC implementation.

In order to maintain performance at extended distances, we need to increase the number of buffers on each interconnecting port to compensate for the number of frames that are in transit.

Configuring the switch ports connected to the DWDM for 10 to 100 km provides 60 buffers and that is enough for this distance.

Scalability

The DWDM comes in "shelves". Each shelf provides four highly available channels. Up to eight shelves can be installed for a total of 32 highly available channels. For an additional number of channels, we would need to install another DWDM and also need another two pairs of fibers.

Security

The DWDM provides a Web based management software that can be accessed by any workstation connected to the same LAN. Different user levels are provided for administrators, operators or observers. Different passwords must be set to limit access. The physical security of fiber connections and patch panels should be considered.

"What if" failure scenarios

Here are some theoretical assumptions:

- **DWDM optical channel card failure**

  As we configured the channels for high availability, there are redundant cards in the DWDM. If we lose one, the traffic will be automatically switched to the other and the channel will remain available. The failed card can be replaced concurrently.

- **DWDM optical channel manager card failure**

  The optical channel manager card performs path high availability switching. There are two cards in each shelf, if one fails the other takes control and the operation is not affected. The failed card can be replaced concurrently.

- **DWDM Optical Multiplexer failure or shelf backplane failure**

  The entire shelf will be unavailable. As we spread connections in different shelves, we will have at least half the channels of each type available. Operation will continue although performance may be affected.

- **Dark fiber failure**

  Because we configured for high availability, operation will continue using the available pair with no performance impact.

12

INRANGE FC/9000 distance storage solutions

In this chapter we will discuss some solutions that can be implemented using the INRANGE FC/9000 (2042-001) included in the IBM portfolio.

We will combine these with some of IBM's storage and other vendors' DWDM solutions to describe high distance SAN solutions.

For product details of the INRANGE FC/9000 refer to:

- *IBM SAN Survival Guide*, SG24-6143
- *IBM SAN Survival Guide Featuring the INRANGE Portfolio*, SG24-6150
- *Implementing an Open IBM SAN,* SG24-6116

The solutions we will discuss are:

- Remote disk consolidation
- Two sites up to 10 km apart
- Two sites up to 80 km apart with native fibre connection
- Point-to-Point DWDM
- Point-to-Point DWDM with PPRC
- Ring topology DWDM
- SAN over WAN
- Remote tape vaulting
- Remote tape vaulting with redundancy

12.1 Remote disk consolidation

This remote disk consolidation solution is an extension of the most basic SAN solutions of disk consolidation.

Disk consolidation provides a means for storage administrators to take advantage of the capabilities of Fibre Channel fabric-based Storage Area Networks to organize storage in more flexible and efficient ways, easing many aspects of storage administration.

This is an attempt to extend the distance between the storage and the servers, in order to provide business benefits such as consolidation, or physical separation of disk storage from the server across sites. This solution extends the distance up to 80 km apart between two directors, using Extended Long wave GBICs (fc #2030) in the SAN fabric, and single mode fiber cable interconnect. However, it is recommended that the distance from server to storage should be considered for your application performance.

We show our solution in Figure 12-1.

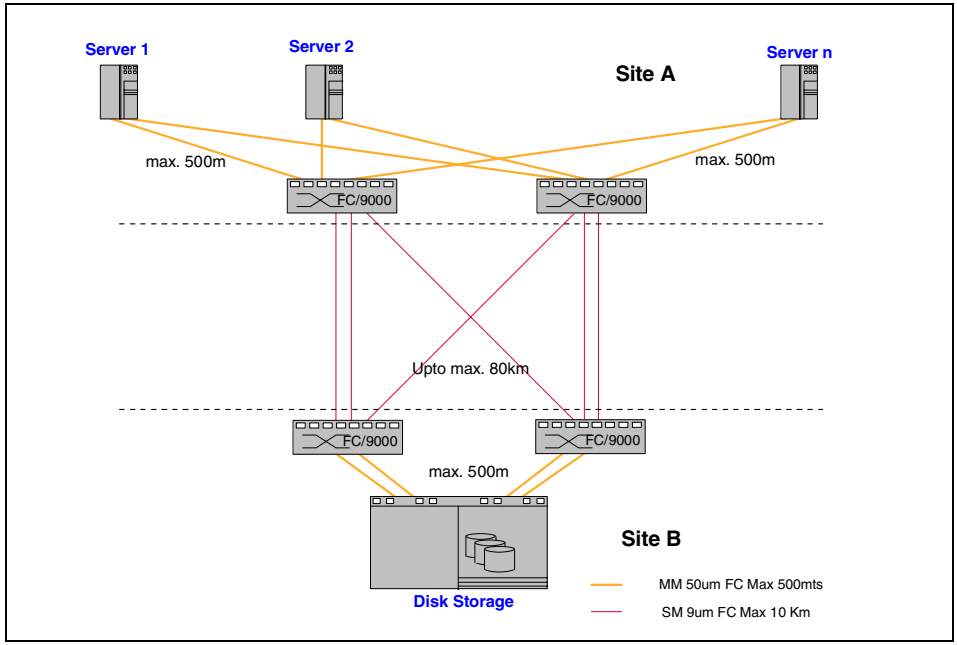


Figure 12-1 Remote disk solution

Checklist

We checked the following items:

- The host servers are connected to the SAN fabric via SW GBIC's using MM fibre cable. (LW GBICs using SM is also possible.)
- Port counts for the configuration
- Zones have meaningful names applied
- The WWNs of the HBAs Port names of servers have been defined to a name server zone
- Nickname assignments so we can quickly cross reference WWNs to devices
- Leave some ports spare for contingency
- Consider the impact of losing an Fibre Channel I/O (FIO) blade, balance the server groups to minimize impact
- Version of cluster is supported with ESS
- Storage capacity and LUN assignments to each server
- Disk storage features and microcode level for proposed configuration
- Dual HBAs have been used for load balancing and failover with SDD
- Version of cluster is supported with SDD
- The version of operating system is supported on device you will be connecting
- Servers that are connected to directors and storage have multiple paths
- All cables have been clearly labeled
- Optimized performance
- E_D_TOV, R_A_TOV, and BB_Credit settings equal on all directors
- Maximum distance for individual devices

Performance

The performance of the SAN will be determined on how much traffic will be moved through the T_Port or HBA. With detailed server profiles, it is possible to balance this accordingly.

For this long distance solution it is of utmost importance to calculate the correct number of lines between sites. Undersizing the number of lines will dramatically effect the performance of the SAN.

From an availability aspect, if we configure a single connection from the ESS to each FIO blade (8 connections), in the event of a single FIO failure we have lost 12.5% of the available bandwidth to the ESS.

It is vital that detailed performance data is available prior to sizing the number of lines required.

Typically, latency will also increase over long distances, a good rule of thumb is 4.8 microseconds per kilometer. INRANGE has performed comprehensive testing over distances in excess of 100 km and has found no performance implications with the director.

The default E_D_TOV and R_A_TOV values do not need to be modified for this distance.

The following considerations are also recommended for better performance:

- Conduct a detailed server performance profile.
- Spread Fibre Channel adapters as evenly as possible across all of the bays.
- Monitor the performance using the IN-VSN software.
- Collect MIB information to determine busy ports.

Scalability

Based on our solution configuration, we have now created a high availability SAN that could support about 100 device connections attached, because the front-end director at site A has 64 ports.

Now we have six connections between directors at site A and directors at site B. You can add more connections between them depending on how many devices are attached and how many ports are available with consideration to performance.

The INRANGE director supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation.

The number of connections between directors at site B and storage can also be increased for more bandwidth.

You should consider that even though you can add more connections for better performance, it may be better to take this into account at the planning stage.

Security

Here are some security considerations:

- All fabrics and storage are in secure locations
- Work with the best carrier company for lines
- Any disk devices that do not support LUN masking are zoned to their respective servers
- The WWNs of the HBAs Port names of servers have been defined

- Disk Storage ESS LUN masking by WWN will allow each server access only to configured LUNs
- IN-VSN Enterprise Manager user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions
- Only SAN administrators have access to the IN-VSN userid and passwords
- As we have several name server zones defined, backups of the IN-VSN should be performed on a regular basis and at least when the information has changed.
- Remote access to Enterprise Manager configured to limit access to authorized workstations
- A maintenance window is available, when OS/390, UNIX, Windows 2000, and Windows NT will be unavailable, so the hard zone can be implemented
- When operating with primary and secondary sites we need to insure all related SAN documentation is in a secure location that can be accessed at the recovery site. We also need to ensure we have enough userids of the correct type that are able to make any required changes to the zones.

"What if" failure scenarios

Here are some theoretical assumptions:

- If one of the lines between site A and site B fails, an alternate route will be used.
- If all high performance profile servers are on the same FIO blade, the Director is a non-blocking device, so there should be no performance impact, although it would be sensible to spread the load.
- If a cable fails between the director and ESS, an alternate route will be used.
- If an FIO blade fails, we still have connectivity, as we have dual connections, but we would lose 50% bandwidth to any connected servers.
- If an Fibre Channel switch (FSW) blade fails, there would be no effect, as the spare FSW would be automatically invoked.
- If a Fibre Channel module (FCM) fails, there would be no effect, as the spare FCM module would be automatically invoked.
- If the backplane was damaged, we would lose connectivity to all servers at that site.
- If a server HBA fails, we lose up to 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.
- If an ESS HBA is unavailable, we have multiple other connections that will automatically be used.

- If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used.
- Director port failure: The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port.
- Fiber failure: Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.
- IN-VSN Enterprise Manager Server failure: No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.
- IN-VSN Server hard drive failure: Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.
- Director completely down, storage completely down, or site down (power, air conditioning, site damage): These will cause an interruption in normal operation.
- Physical damage to storage causing data loss (fire, flood): We will need to restore data from backup tapes.

12.2 Two sites up to 10 km apart

One can assume that a greater distance between IT sites results in greater security against wide spread disaster. However, increased distance has its price in terms of interconnection cost, business relocation effort, and network cost. Before you start to expand your storage network, you have to consider if the following solution can meet your business purpose.

The solution as shown in Figure 12-2 can be used for accessing databases at a remote site, as well as mirroring data at an alternative site up to 10 km between two directors at both sites.

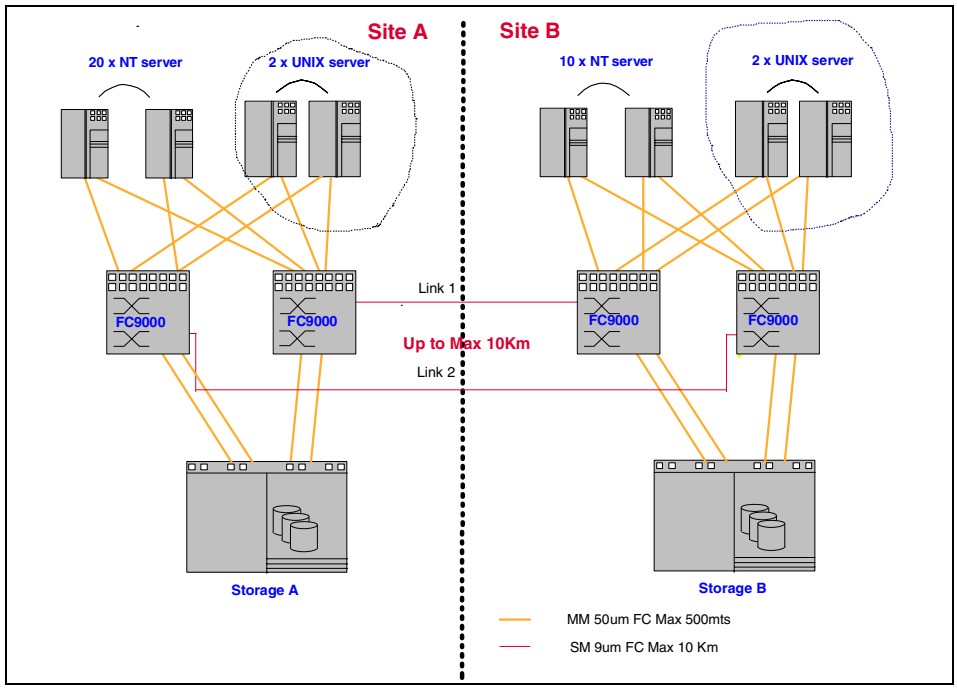


Figure 12-2 Solution for up to 10 km apart

We have assumed that for our solution:

- Both sites have SAN infrastructure.
- UNIX servers in production, which are being used for mission critical production, need data mirroring for disaster recovery.
- NT servers at site A need to access database in storage B at site B, in addition to using data in storage A. And NT servers' data are backed up by tape subsystem according to company's back-up procedure.
- UNIX servers at site B are in standby mode and being used for backup servers at ordinary times.
- No DWDM and Channel Extender proposed.
- No PPRC proposed.

Checklist

The following items should be considered to implement the solution:

- The host servers are connected to the SAN fabric via SW GBIC's using MM fibre cable. (LW GBICs using SM is possible.)
- Single mode FC cables are used between FC/9000s at each site.

- Operating systems supported S/W mirroring.
- Software mirroring is used in UNIX servers.
- Fibre Channel supported on servers.
- Longwave laser GBICs (enabling a transmission distance of up to 10 kilometers). Any application performance implications of longer distances.

Note: Eight GBICs are required for each FIO module: eight shortwave, eight long wave, eight extended longwave, or mixture of shortwave, longwave and/or extended longwave GBIC.

- Port counts for the configuration.
- Zones have meaningful names applied.
- The WWNs of the HBAs Port names of servers have been defined to a name server zone.
- Nickname assignments so we can quickly cross reference WWNs to devices.
- Leave some ports spare for contingency.
- Consider the impact of losing an FIO blade, balance the server groups to minimize impact.
- Version of cluster is supported with ESS.
- Storage capacity and LUN assignments to each server.
- Disk storage features and microcode level for proposed configuration.
- Dual HBAs have been used for load balancing and failover with SDD.
- Version of cluster is supported with SDD.
- The version of operating system is supported on device you will be connecting.
- Servers that are connected to directors and storage have multiple paths.
- All cables have been clearly labeled.
- Optimized performance.
- E_D_TOV, R_A_TOV, and BB_Credit settings equal on all directors.
- Maximum distance for individual devices.

Performance

The major performance consideration with a long distance solution is calculating the correct number of lines between sites. This number can only be determined by performing a detailed performance profile of the servers and storage that will be remote. Over estimating the number of lines will increase costs dramatically, under sizing the number of lines will dramatically effect the performance of the SAN.

It is vital that detailed performance data is available prior to sizing the number of lines required.

Typically, latency will increase over long distances, a good rule of thumb is 4.8 microseconds per kilometer. INRANGE has performed comprehensive testing over distances in excess of 100 km and has found no performance implications with the director.

The following considerations are also recommended for better performance:

- The INRANGE Director supports buffer to buffer credits. As we have mentioned previously, buffer to buffer credits allow commands to be queued up in the buffer of the switch, lessen the effect of the latency, and improve aggregate performance.
- Conduct a detailed server performance profile.
- Spread Fibre Channel adapters as evenly as possible across all of the bays.
- Monitor the performance using the IN-VSN software.
- Collect MIB information to determine busy ports.

Scalability

Based on our solution configuration, we have now created a high availability SAN that could support about 100 device connections attached.

Now we have two connections between site A and site B. You can add more connections between them depending on how many devices are attached and how many ports are available with consideration to performance.

The INRANGE director supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation.

The number of connections between directors and storage are also able to be increased for more bandwidth.

But if other protocol support is needed as well as more fibre cables between two sites, its scalability using this configuration would be limited. Then you would need consider using DWDM.

Security

Here are some security considerations:

- All fabrics and storage are in secure locations.
- Work with the best carrier company for lines.
- Any disk devices that do not support LUN masking are zoned to their respective servers.
- The WWNs of the HBAs Port names of servers have been defined.
- Disk Storage ESS LUN masking by WWN will allow each server access only to configured LUNs.
- IN-VSN Enterprise Manager user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions.
- Only SAN administrators have access to the IN-VSN userid and passwords.
- As we have several name server zones defined, backups of the IN-VSN should be performed on a regular basis and at least when the information has changed.
- Remote access to Enterprise Manager configured to limit access to authorized workstations.
- A maintenance window is available, when OS/390, UNIX, Windows 2000, and Windows NT will be unavailable, so the hard zone can be implemented.
- When operating with primary and secondary sites we need to ensure all related SAN documentation is in a secure location that can be accessed at the recovery site. We also need to ensure we have enough userids of the correct type that are able to make any required changes to the zones.

“What if” failure scenarios

Here are some theoretical assumptions:

- If one of the lines between site A and site B fails, an alternate route will be used.
- If all high performance profile servers are on the same FIO blade, the director is a non-blocking device, so there should be no performance impact, although it would be sensible to spread the load.
- If a cable fails between the director and ESS, an alternate route will be used.

- If an FIO blade fails, we still have connectivity, as we have dual connections, but we would lose 50% bandwidth to any connected servers.
- If an FSW blade fails, there would be no effect, as the spare FSW would be automatically invoked.
- If an FCM module fails, there would be no effect, as the spare FCM module would be automatically invoked.
- If the backplane was damaged, we would lose connectivity to all servers at that site.
- If a server HBA fails, we lose up to 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.
- If an ESS HBA is unavailable, we have multiple other connections that will automatically be used.
- If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used.
- Director port failure — The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port.
- Fiber failure — Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.
- IN-VSN Enterprise Manager Server failure. No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.
- IN-VSN Server hard drive failure — Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.
- Director completely down, storage completely down, or site down (power, air conditioning, site damage); It will be recovered from alternative site.

12.3 Two sites up to 100 km apart

The 2042 Extended Longwave Wave GBIC is an optional 2042 feature (FC 2030) that provides for Interswitch Links up to 80 km.

Note: With the use of the GBIC no repeaters are necessary to reach a distance up to 80 km. However, with the use of repeaters and by combining the extended longwave GBIC with repeaters, greater distances can be achieved. In this section we focus on the GBIC only.

We show an example of this configuration in Figure 12-3.

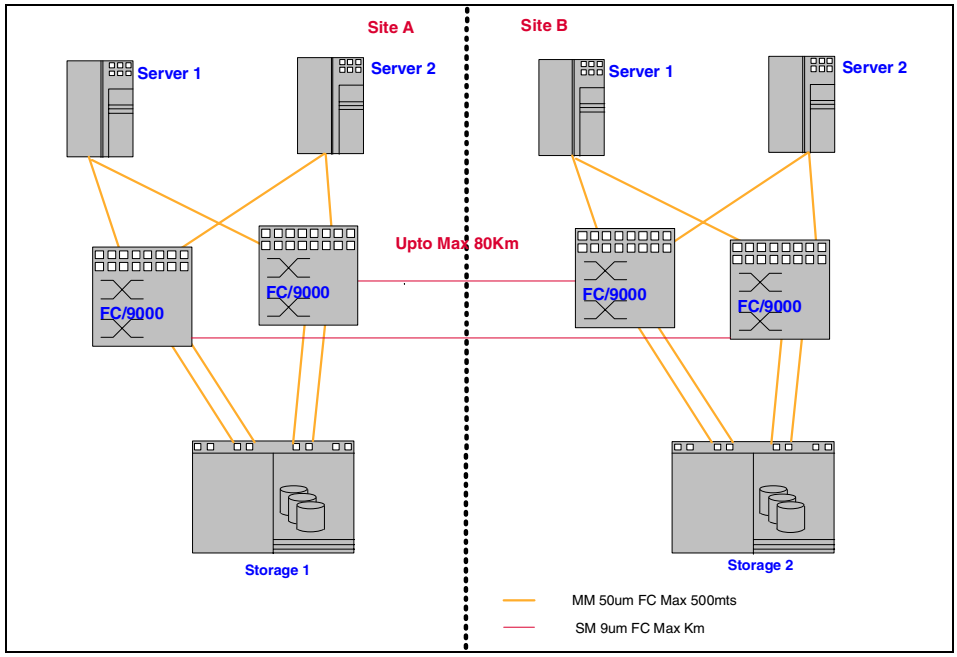


Figure 12-3 Solution for up to 80 km apart

We have assumed that for our solution:

- Both sites have SAN infrastructure.
- Connection of servers to FC/9000, FC/9000 to storage use multiple mode fibre.
- Multimode FC cables are used for connection between servers and FC/9000, and between FC/9000 and storages. And, single mode FC cables are used between FC/9000s at each site.
- No DWDM and Channel Extender proposed.
- No PPRC proposed.

Checklist

The following items should be considered to implement the solution:

- The 2042 Extended Longwave Wave GBIC is an optional 2042 feature (FC 2030).
- The host servers are connected to the SAN fabric via SW GBIC's using MM fibre cable. (LW GBICs using SM is possible.)
- Single mode FC cables are used between FC/9000s at each site.
- Operating systems supported S/W mirroring.
- Software mirroring is used in UNIX servers.
- Fibre Channel supported on servers.

Note: Eight GBICs are required for each FIO module: eight shortwave, eight long wave, eight extended longwave, or mixture of shortwave, longwave and/or extended longwave GBIC.

- Port counts for the configuration.
- Zones have meaningful names applied.
- The WWNs of the HBAs Port names of servers have been defined to a name server zone.
- Nickname assignments so we can quickly cross reference WWNs to devices.
- Leave some ports spare for contingency.
- Consider the impact of losing an FIO blade, balance the server groups to minimize impact.
- Version of clustering is supported with ESS.
- Storage capacity and LUN assignments to each server.
- Disk storage features and microcode level for proposed configuration.
- Dual HBAs have been used for load balancing and failover with SDD.
- Version of clustering is supported with SDD.
- The version of operating system is supported on device you will be connecting.
- Servers that are connect to directors and storage have multiple paths.
- All cables have been clearly labeled.
- Optimized performance.
- E_D_TOV, R_A_TOV, and BB_Credit settings equal on all directors.
- Maximum distance for individual devices.

Performance

The major performance consideration with a long distance solution is calculating the correct number of lines between sites. This number can only be determined by performing a detailed performance profile of the servers and storage that will be remote. Over-estimating the number of lines will increase costs dramatically, under sizing the number of lines will dramatically effect the performance of the SAN.

It is vital that detailed performance data is available prior to sizing the number of lines required.

Typically, latency will increase over long distances, a good rule of thumb is 4.8 microseconds per kilometer. INRANGE has performed comprehensive testing over distances in excess of 100 km and has found no performance implications with the director.

The following considerations are also recommended for better performance.

- The INRANGE FC/9000 supports buffer to buffer credits. As we have already mentioned, buffer to buffer credits allow commands to be queued up in the buffer of the switch and this lessens the effect of the latency and improves aggregate performance.
- Conduct a detailed server performance profile.
- Spread Fibre Channel adapters as evenly as possible across all of the bays.
- Monitor the performance using the IN-VSN software.
- Collect MIB information to determine busy ports.

Scalability

Based on our solution configuration, we have now created a high availability SAN that could support about 100 device connections attached.

Now we have two connections between site A and site B. You can add more connections between them depending on how many devices are attached and how many ports are available with consideration to performance. The INRANGE director supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation.

The number of connections between directors and storage are also to be increased for more bandwidth.

However, if other protocol support is needed as well as more fibre cables between two sites, its scalability, having this configuration, would be limited. Then you need to consider using DWDM.

Security

Here are some security considerations:

- All fabrics and storage are on the secure locations.
- Work with the best carrier company for lines.
- Any disk devices that do not support LUN masking are zoned to their respective servers.
- The WWNs of the HBAs Port names of servers have been defined.
- Disk Storage ESS LUN masking by WWN will allow each server access only to configured LUNs.
- IN-VSN Enterprise Manager user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions.
- Only SAN administrators have access to the IN-VSN userid and passwords.
- As we have several name server zones defined, backups of the IN-VSN should be performed on a regular basis and at least when the information has changed.
- Remote access to Enterprise Manager configured to limit access to authorized workstations.
- A maintenance window is available, when OS/390, UNIX, Windows 2000, and Windows NT will be unavailable, so the hard zone can be implemented.
- When operating with primary and secondary sites, we need to ensure all related SAN documentation is in a secure location that can be accessed at the recovery site. We also need to ensure we have enough userids of the correct type that are able to make any required changes to the zones.

“What if” failure scenarios

Here are some theoretical assumptions:

- If one of the lines between site A and site B fails, an alternate route will be used.
- If all high performance profile servers are on the same FIO blade, the director is a non-blocking device, so there should be no performance impact, although it would be sensible to spread the load.
- If a cable fails between the director and ESS, an alternate route will be used.
- If an FIO blade fails, we still have connectivity, as we have dual connections, but we would lose 50% bandwidth to any connected servers.
- If an FSW blade fails, there would be no effect, as the spare FSW would be automatically invoked.

- If an FCM module fails, there would be no effect, as the spare FCM module would be automatically invoked.
- If the backplane was damaged, we would lose connectivity to all servers at that site.
- If a server HBA fails, we lose up to 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.
- If an ESS HBA is unavailable, we have multiple other connections that will automatically be used.
- If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used.
- Director port failure — The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port.
- Fiber failure — Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.
- IN-VSN Enterprise Manager Server failure. No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.
- IN-VSN Server hard drive failure — Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.
- Director completely down, storage completely down, or site down (power, air conditioning, site damage); It will be recovered from alternative site.

12.4 Point-to-point DWDM

Figure 12-4 shows a basic high availability SAN design for multiple servers deployed with dual directors and redundant fabric at each site. This design also uses the DWDM equipment for Intersite ISL links between the directors.

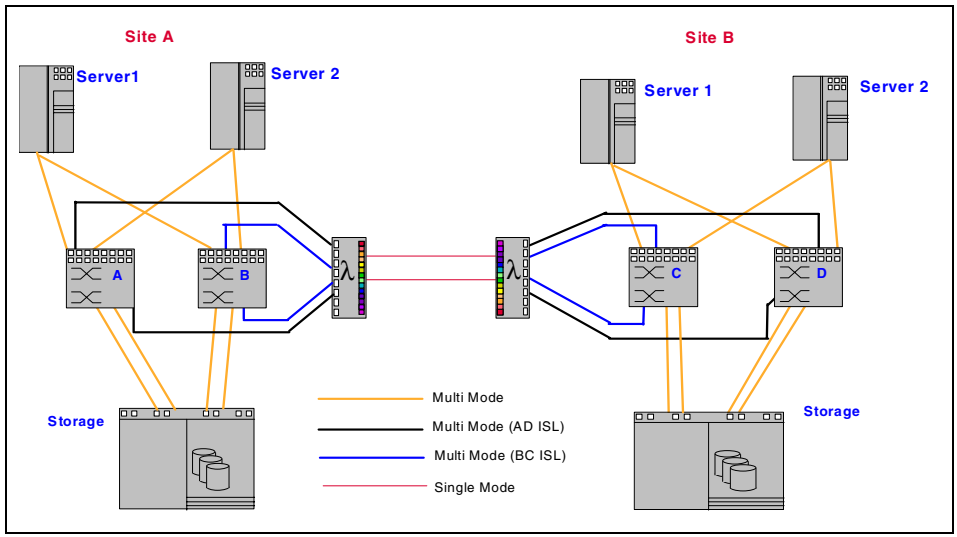


Figure 12-4 Point-to-point DWDM solution; two sites

We have shown a point-to-point topology here. We would implement the DWDM solution as a protected fiber to ensure availability. In this example, we have four multimode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the intersite traffic that is expected. This will be driven by the reasons for the implementation and here we have assumed a light workload.

Here we show an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fiber is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one director in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits, however this is not a DWDM consideration but more to do with general SAN design solutions over a distance.

The DWDM is a core architecture deployment and because of its independence from protocol it can be used for other traffic, that is to say it extends beyond the SAN environment. The distance we have shown here is up to 100 km between nodes.

Within each site, server to storage availability is provided with this dual director, dual fabric design. Dual HBAs are installed in each host and each storage device must have at least two ports. Failover for a failed path or even a failed switch is dependent on host failover software, namely, IBM SDD.

In Figure 12-4, AD ISL (Director A and D) will be a discrete fabric from BC ISL (Director B and C) fabric.

The components used are as follows:

- SAN fabric
  - IBM 2042-001(INRANGE FC/9000)
- DWDM
  - Two DWDM's linked via dark fiber
- Servers
  - Four servers with HBA cards
- Two ESS 2105-F20 with native Fibre Channel Adapter
- Software
  - IBM Subsystem Device Driver (SDD)
  - IBM StorWatch Specialist
  - IBM ESS Expert

Checklist

In addition to the items we have already considered we also checked the following items:

- The ISL link count to the DWDM equipment at each site
- Configuration and count of the client interface adapters in the DWDM equipment
- Configuration and count of the transport network adapters in the DWDM, ensuring same frequency band and channel
- Connectivity testing between
  - ESS and switches
  - Servers and switch
  - Switches and DWDM
- Pre-installation testing of the DWDM transport network for configuration and performance

Performance

The DWDM equipment is transparent to the subsystems that are using it. There are no internal queues or busy conditions, so it does not affect performance. The performance factors we need to consider in this solution are the number of ISLs and distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the director and the DWDM equipment.

Scalability

The DWDM comes in “shelves”. Each shelf provides four high availability channels. Up to eight shelves can be installed for a total of 32 high availability channels. For additional number of channels, we would need to install another DWDM and also need another two pairs of fibers.

Most DWDMs support multiple I/O interfaces for client equipment and support up to 32 separate channels. DWDM equipment also allows for mixing different protocol interfaces of client equipment to a single physical DWDM equipment.

Scalability rules for the SAN fabric of INRANGE Directors remain unchanged as in previous examples.

Security

Dual fabrics can protect you against user errors, such as a user erasing or changing the zoning information inappropriately. The zoning information is separate for each fabric SAN AD and SAN BC, so when changing the zoning in one fabric, it does not automatically propagated into the other fabric.

Should you decide to add an ISL, ensure that all checks are done on the zone configuration changes from the management console.

The DWDM equipment provides a Web based management software that can be accessed by any workstation connected to the same LAN. Different user levels are provided for administrators, operators or observers. Different passwords must be set to limit access.

The physical security of fiber connections and patch panels should be considered.

“What if” failure scenarios

In addition to the native solutions, this section has more considerations on “What if” failure scenarios about DWDM. Here are some theoretical assumptions:

- **DWDM optical channel card failure**

  As we configured the channels for high availability, there are redundant cards in the DWDM. If we lose one, the traffic will automatically be switched to the other and the channel will remain available. The failed card can be replaced concurrently.

- **DWDM optical channel manager card failure**

  The optical channel manager card performs path high availability switching. There are two cards in each shelf, and if one fails, the other takes control and the operation is not affected. The failed card can be replaced concurrently.

- **DWDM Optical Multiplexer failure or shelf backplane failure**

  The entire shelf will be unavailable. As we spread connections in different shelves, we will have at least half the channels of each type available. Operation will continue although performance may be affected.

- **Dark fiber failure**

  As we configured for high availability, operation will continue to use the available pair with no performance impact. Customers may have additional considerations based on their business recovery policy to source the redundant Dark fiber from another telco/source.

12.5 Point-to-point DWDM with PPRC

In Figure 12-5 we show a basic high availability SAN design for multiple servers deployed with dual directors and redundant fabric at each site.

This design also uses the DWDM equipment for Intersite ISL links between the directors and the PPRC for mirroring data on the remote site.

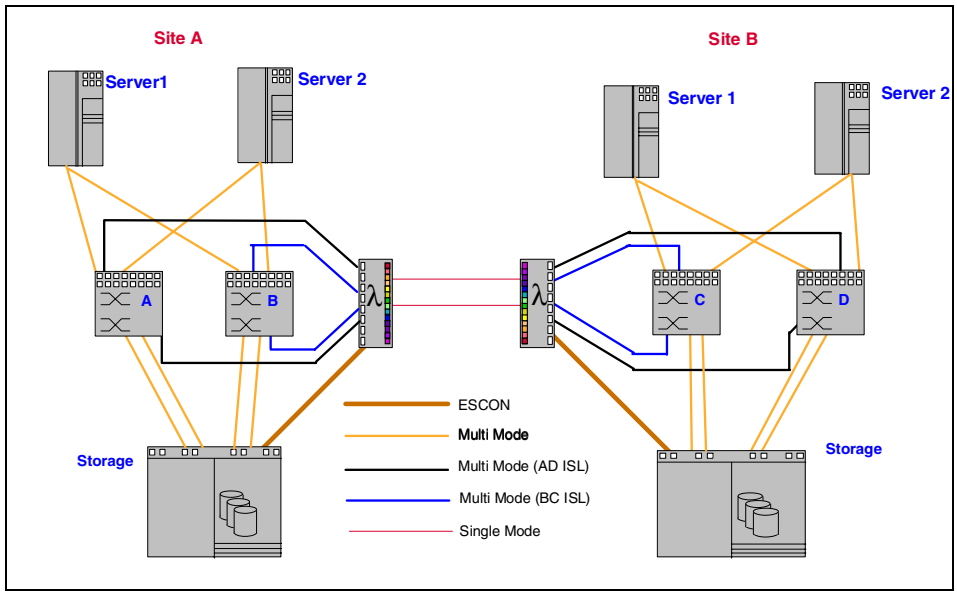


Figure 12-5 Point-to-point DWDM with PPRC solution

We have shown a point-to-point topology of DWDM here. But The PPRC connection today is via ESCON, so this can utilize the DWDM for intersite transport alongside the Fibre Channel.

We would implement the DWDM solution as a protected fiber to ensure availability. In this example, we have four multimode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the intersite traffic that is expected. This will be driven by the reasons for the implementation and here we have assumed a light workload.

Here we show an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one director in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits, however this is not a DWDM consideration but more to do with the general SAN solution over distance.

The DWDM is a core architecture deployment and because of its independence from protocol it can be used for lots of other traffic, it extends beyond the SAN environment. The distance we have shown here is up to 100 km between nodes.

The ESCON traffic can be transported over the already deployed DWDM architecture. We show a single pair of 62.5 ESCON connections here, connecting the ESCON ports on the ESS to the DWDM, again this will be dependent upon traffic expectations and should follow the ESCON channel sizing guidelines for PPRC implementation.

In order to maintain performance at extended distances, we need to increase the number of buffers on each interconnecting port to compensate for the number of frames that are in transit. Configuring the director ports connected to the DWDM for 10 to 100 km provides 60 buffers, and that is enough for this distance.

Within each site, server to storage availability is provided with this dual director, dual fabric design. Dual HBAs are installed in each host and each storage device must have at least two ports. Failover for a failed path or even a failed switch is dependent on host failover software, namely, IBM SDD.

In Figure 12-4, AD ISL(Director A and D) will be discrete fabric from BC ISL (Director B and C) fabric. The components used are as follows:

- SAN fabric
  - IBM 2042-001(INRANGE FC/9000)
- DWDM
  - Two DWDM's linked via dark fibre
- Servers
  - Four servers with HBA cards
- Two ESS 2105-F20 with native Fibre Channel Adapter
- Software
  - IBM Subsystem Device Driver (SDD)
  - IBM ESS StorWatch Specialist
  - IBM ESS Expert

Checklist

We also checked the following items:

- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration.
- Storage capacity and LUN assignments to each server.
- Disk storage features and microcode level for proposed configuration.

- INRANGE director LW GBIC count.
- The ISL link count to the DWDM equipment at each site.
- SM FC cable laying and termination.
- INRANGE Director high availability features.
- E_D_TOV, R_A_TOV, and BB_Credit settings equal on all directors.
- Maximum distance for individual devices.
- Nickname assignments so we can quickly cross reference WWNs to devices.
- Configuration and count of the client interface adapters in the DWDM equipment.
- Configuration and count of the transport network adapters in the DWDM, ensuring same frequency band and channel.
- Connectivity testing between:
  - ESS and directors
  - servers and directors
  - Directors and DWDM
  - ESS ESCON HBAs and the DWDM
- Validate connectivity:
  - Test Server to storage on both paths
  - Tests of failover/failback operations
  - Test inter-site connectivity via DWDM
  - Setup LUNs on back up ESS and test from primary server, to quantify latency
  - Test PPRC - establish paths and Test PPRC functioning - qualify timings
- Users and password defined in IN-VSN.
- Pre-installation testing of the DWDM transport network for configuration and performance.

Performance

The major performance consideration with a long distance solution is calculating the correct number of lines between sites. This number can only be determined by performing a detailed performance profile of the servers and storage that will be remote.

The DWDM equipment is transparent to the subsystems that are using it. There are no internal queues or busy conditions, so it does not affect performance. The performance factors we need to consider in this solution are the number of ISLs and distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the Director and the DWDM equipment.

From the performance point of view for PPRC implementation, more ESCON channels (up to eight) attached to DWDM equipment make for better performance. The number of ESCON channels are calculated by the data transfer rate at peak time.

The following considerations are also recommended for better performance:

- The INRANGE Director supports buffer to buffer credits. As we already mentioned, buffer to buffer credits allow commands to be queued up in the buffer of the switch and this lessens the effect of the latency and improves aggregate performance.
- Conduct a detailed server performance profile.
- Spread Fibre Channel adapters as evenly as possible across all of the bays.
- Monitor the performance using the IN-VSN software.
- Collect MIB information to determine busy ports.

Scalability

The DWDM comes in “shelves”. Each shelf provides four high availability channels. Up to eight shelves can be installed for a total of 32 high availability channels. For additional number of channels, we would need to install another DWDM and also need another two pairs of fibers.

Most DWDM support multiple I/O interfaces for client equipment and support up to 32 separate channels. DWDM equipment also allow for mixing different protocol interfaces of client equipment to a single physical DWDM equipment.

Scalability rules for the SAN fabric of INRANGE Directors remain unchanged as in previous examples.

Security

Dual fabrics can protect you against user errors, such as a user erasing or changing the zoning information inappropriately. The zoning information is separate for each fabric SAN AD and SAN BC, so when changing the zoning in one fabric it does not get automatically propagated into the other fabric.

Should you decide to add an ISL, ensure that all checks are done on the zone configuration changes from the management console.

The DWDM equipment provide a Web based management software that can be accessed by any workstation connected to the same LAN. Different user levels are provided for administrators, operators or observers. Different passwords must be set to limit access.

The physical security of fiber connections and patch panels should be considered.

Here are considerations for security:

- All fabrics and storage are on the secure locations.
- Work with the best carrier company for lines.
- Any disk devices that do not support LUN masking are zoned to their respective servers.
- The WWNs of the HBAs Port names of servers have been defined.
- Disk Storage ESS LUN masking by WWN will allow each server access only to configured LUNs.
- IN-VSN Enterprise Manager user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions.
- Only SAN administrators have access to the IN-VSN userid and passwords.
- As we have several name server zones defined, backups of the IN-VSN should be performed on a regular basis and at least when the information has changed.
- Remote access to Enterprise Manager configured to limit access to authorized workstations.
- A maintenance window is available, when OS/390, UNIX, Windows 2000, and Windows NT will be unavailable, so the hard zone can be implemented.
- When operating with primary and secondary sites we need to ensure all related SAN documentation is in a secure location that can be accessed at the recovery site. We also need to ensure we have enough userids of the correct type that are able to make any required changes to the zones.

“What if” failure scenarios

Here are some theoretical assumptions:

- DWDM optical channel card failure: As we configured the channels for high availability, there are redundant cards in the DWDM. If we lose one, the traffic will automatically be switched to the other and the channel will remain available. The failed card can be replaced concurrently.

- DWDM optical channel manager card failure: The optical channel manager card performs path high availability switching. There are two cards in each shelf, if one fails the other takes control and the operation is not affected. The failed card can be replaced concurrently.
- DWDM Optical Multiplexer failure or shelf backplane failure: The entire shelf will be unavailable. As we spread connections in different shelves, we will have at least half the channels of each type available. Operation will continue although performance may be affected.
- Dark fiber failure: As we configured for high availability, operation will continue to use the available pair with no performance impact. Customers may have additional considerations based on their business recovery policy to source the redundant dark fiber from another telco/source.
- If the storage in primary site is not recoverable, use the storage in alternative site which data are mirrored.
- If one of the lines between site A and site B fails, an alternate route will be used.
- If all high performance profile servers are on the same FIO blade, the Director is a non-blocking device, so there should be no performance impact, although it would be sensible to spread the load.
- If a cable fails between the director and ESS, an alternate route will be used.
- Regarding the director, if an FIO blade fails, we still have connectivity, as we have dual connections, but we would lose 50% bandwidth to any connected servers.
- If an FSW blade fails, there would be no effect, as the spare FSW would be automatically invoked.
- If an FCM module fails, there would be no effect, as the spare FCM module would be automatically invoked.
- If the backplane was damaged, we would lose connectivity to all servers at that site.
- If a server HBA fails, we lose up to 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.
- If an ESS HBA is unavailable, we have multiple other connections that will automatically be used.
- If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used.
- Director port failure: The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port.

- Fiber failure: Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.
- IN-VSN Enterprise Manager Server failure: No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.
- IN-VSN Server hard drive failure: Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.
- Director completely down, storage completely down, or site down (power, air conditioning, site damage; It will be recovered from alternative site.

12.6 Ring topology DWDM

In Figure 12-6, we show you a multimode DWDM configuration that spans four sites.

This gives a logical mesh, potentially giving any to any connectivity. The configuration has two rings, one ring connects to one director in each site, the other ring connects to the remaining director; this gives two discrete SAN fabrics. Each site is connected over a DWDM channel that includes dual paths for transmitting and receiving.

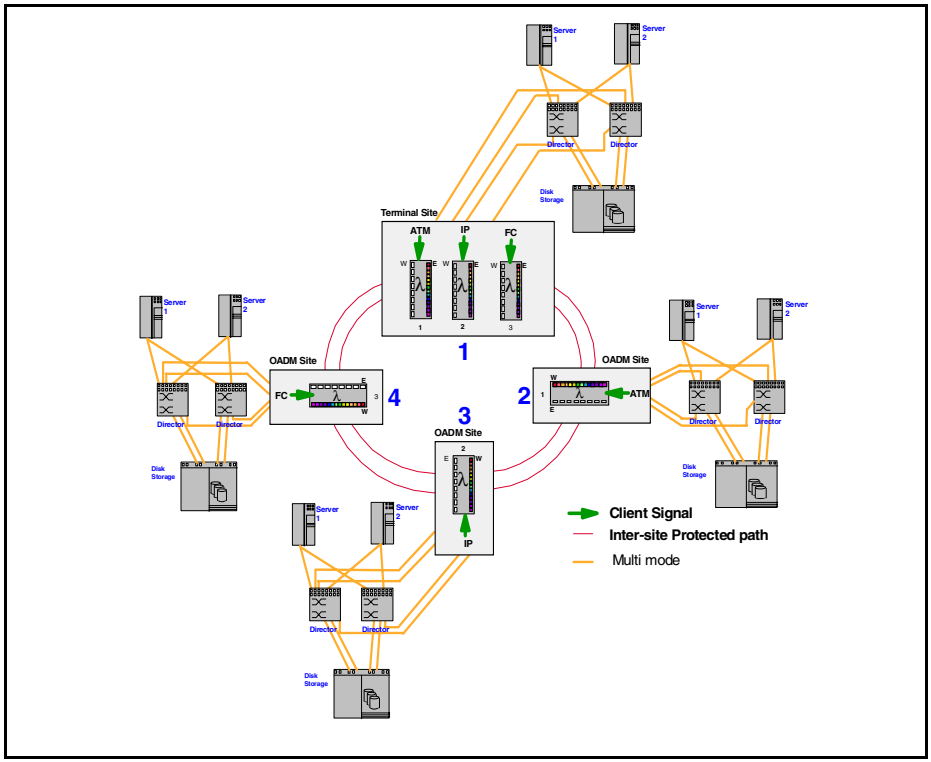


Figure 12-6 Multiple site ring topology DWDM solution

The number of channels that will be needed at each location is going to be dependent upon the intersite traffic that is expected. This will be driven by the reasons for the implementation, here we have assumed an light workload. Here we show an enterprise deployment that enable data to be made available across a metropolitan area network, or a company campus. It is likely that fiber is expensive here and that the DWDM reduces this overhead.

Each path has its own wavelength. The DWDM pass through feature enables non-contiguous sites to connect over an intermediate site as if they were directly connected. The only additional overhead of the pass through is the minimal latency (5 microseconds/km) of the second link. The pass through has no overhead since it is a passive device. This fabric would logically appear as a fully meshed topology.

Each of the links can operate in protected mode, which provides a redundant path in the event of a link failure. In most cases, link failures are automatically detected within 50 msec. In this case, the two wavelengths of the failed link reverse directions and reach the target port at the opposite side of the ring.

Calculating the distance between nodes in a ring depends on the implementation of the protected path scheme. For instance, if the link between DWDM 2 and 3 fails, the path from 1 to 3 would be 1 to 2, back from 2 to 1 (due to the failed link), 1 to 4, and finally 4 to 3.This illustrates the need to utilize the entire ring circumference (and more, in a configuration with over four nodes) for failover.

Another way to calculate distance between nodes is to set up the protected path in advance (in the reverse direction) so the distance is limited to the number of hops between the two nodes. In either case, the maximum distance between nodes determines the maximum optical reach.

Latency will need to be taken into account, which is also related to buffer credits, however this is not unique to DWDM, more the general SAN solution over distance. The DWDM is a core architecture deployment and because of its independence from protocol it can be used for other traffic and it extends beyond the SAN environment. The distance we have shown here is 25 km between nodes.

Within each site server to storage availability is provided with this dual director, dual fabric design. Dual HBAs are installed in each host and each storage device must have at least two ports. Failover for a failed path or even a failed switch is dependent on host failover software, namely, IBM Subsystem Device Driver (SDD). The components used are as follows:

- SAN fabric
  - IBM 2042-001(INRANGE FC/9000)
- DWDM
  - Two DWDM's linked via dark fibre
- Servers
  - Eight servers with HBA cards
- ESS 2105-F20 with native Fibre Channel Adapter
- Software
  - IBM Subsystem Device Driver (SDD)
  - IBM ESS StorWatch Specialist
  - IBM ESS Expert

Checklist

These are the additions to our checklist:

- Configuration and count of the client interface adapters in the DWDM equipment.

- Configuration and count of the transport network adapters in the DWDM, ensuring same frequency band and channel.
- Connectivity testing between:
  - ESS and directors
  - Servers and directors
  - Directors and DWDM
- Validate connectivity:
  - Test server to storage on both paths
  - Tests of failover/failback operations
  - Test inter-site connectivity via DWDM
  - Setup LUNs on back up ESS and test from primary server, to quantify latency
- Users and password defined in IN-VSN.
- Pre-installation testing of the DWDM transport network for configuration and performance.
- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration.
- Storage capacity and LUN assignments to each server.
- Disk storage features and microcode level for proposed configuration.
- INRANGE director LW GBIC count.
- The ISL link count to the DWDM equipment at each site.
- SM FC cable laying and termination.
- INRANGE Director high availability features.
- E_D_TOV, R_A_TOV, and BB_Credit settings equal on all directors.
- Maximum distance for individual devices.
- Nickname assignments so we can quickly cross reference WWNs to devices.

Performance

The major performance consideration with a long distance solution is calculating the correct number of lines between sites. This number can only be determined by performing a detailed performance profile of the servers and storage that will be remote.

The DWDM equipment is transparent to the subsystems that are using it. There are no internal queues or busy conditions, so it does not affect performance. The performance factors we need to consider in this solution are the number of ISLs and distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the director and the DWDM equipment.

The following considerations are also recommended for better performance:

- The 2042 supports buffer to buffer credits. As we already mentioned, buffer to buffer credits allow commands to be queued up in the buffer of the switch and this lessens the effect of the latency and improves aggregate performance.
- Conduct a detailed server performance profile.
- Spread Fibre Channel adapters as evenly as possible across all of the bays.
- Monitor the performance using the IN-VSN software.
- Collect MIB information to determine busy ports.

Scalability

The DWDM comes in “shelves”. Each shelf provides four high availability channels. Up to eight shelves can be installed for a total of 32 high availability channels. For an additional number of channels, we would need to install another DWDM and also need another two pairs of fibers.

Most DWDMs support multiple I/O interfaces for client equipment and support up to 32 separate channels. DWDM equipment also allow for mixing different protocol interfaces of client equipment to a single physical DWDM equipment.

Scalability rules for the SAN fabric of INRANGE Directors remain unchanged as in previous examples.

Security

Dual fabrics can protect you against user errors, such as a user erasing or changing the zoning information inappropriately. Should you decide to add an ISL, ensure that all checks are done on the zone configuration changes from the management console.

The physical security of fiber connections and patch panels should be considered.

Here are considerations for security:

- All fabrics and storage are on the secure locations.
- Work with the best carrier company for lines
- Any disk devices that do not support LUN masking are zoned to their respective servers.
- The WWNs of the HBAs Port names of servers have been defined.

- Disk Storage ESS LUN masking by WWN will allow each server access only to configured LUNs.
- IN-VSN Enterprise Manager user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions .
- Only SAN administrators have access to the IN-VSN userid and passwords.
- As we have several name server zones defined, backups of the IN-VSN should be performed on a regular basis and at least when the information has changed.
- Remote access to Enterprise Manager configured to limit access to authorized workstations.
- A maintenance window is available, when OS/390, UNIX, Windows 2000, and Windows NT will be unavailable, so the hard zone can be implemented.
- When operating with primary and secondary sites we need to ensure all related SAN documentation is in a secure location that can be accessed at the recovery site. We also need to ensure we have enough userids of the correct type that are able to make any required changes to the zones.

"What if" failure scenarios

Here are some theoretical assumptions:

- **DWDM optical channel card failure**

  As we configured the channels for high availability, there are redundant cards in the DWDM. If we lose one, the traffic will automatically be switched to the other and the channel will remain available. The failed card can be replaced concurrently.

- **DWDM optical channel manager card failure**

  The optical channel manager card performs path high availability switching. There are two cards in each shelf, if one fails the other takes control and the operation is not affected. The failed card can be replaced concurrently.

- **DWDM Optical Multiplexer failure or shelf backplane failure**

  The entire shelf will be unavailable. As we spread connections in different shelves, we will have at least half the channels of each type available. Operation will continue although performance may be affected.

- **Dark fiber failure**

  As we configured for high availability, operation will continue to use the available pair with no performance impact. Customers may have additional considerations based on their business recovery policy to source the redundant dark fiber from another telco/source.

- If one of the lines between site A and site B fails, an alternate route will be used.
- If all high performance profile servers are on the same FIO blade, the director is a non-blocking device, so there should be no performance impact, although it would be sensible to spread the load.
- If a cable fails between the director and ESS, an alternate route will be used.
- Regarding the director, if an FIO blade fails, we still have connectivity, as we have dual connections, but we would lose 50% bandwidth to any connected servers.
- If an FSW blade fails, there would be no effect, as the spare FSW would be automatically invoked.
- If an FCM module fails, there would be no effect, as the spare FCM module would be automatically invoked.
- If the backplane was damaged, we would lose connectivity to all servers at that site.
- If a server HBA fails, we lose up to 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.
- If an ESS HBA is unavailable, we have multiple other connections that will automatically be used.
- If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used.
- **Director port failure**

  The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port.
- **Fiber failure**

  Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.
- **IN-VSN Enterprise Manager Server failure**

  No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.
- **IN-VSN Server hard drive failure**

  Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.

- Director completely down, storage completely down, or site down (power, air conditioning, site damage); It will be recovered from alternative site.

12.7 SAN over WAN

Channel extenders usually use telecommunication lines for data transfer and therefore enable application and recovery sites to be located as far apart as possible. The use of channel extenders provides the separation for disaster recovery purposes and avoids some of the barriers imposed when customers do not have a “right of way” to lay fiber cable.

We show you an example of our SAN/WAN solution in Figure 12-7.

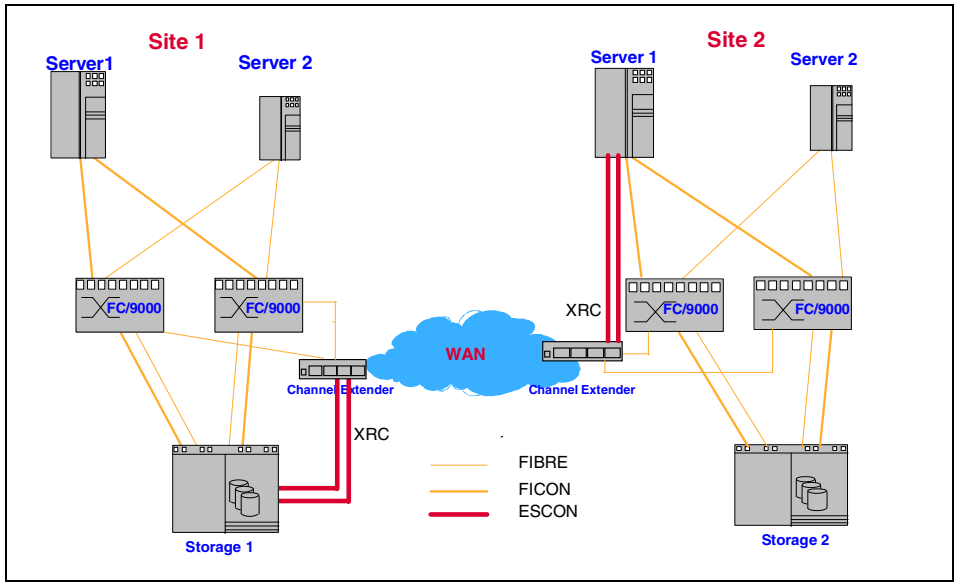


Figure 12-7 SAN/WAN solutions with channel extenders

We assumed the following:

- Server 1 is an IBM 9672 Parallel Enterprise G5, G6 and zSeries 900 servers with FICON channel cards.
- Server 2 is an Intel- and UNIX-based servers with Fibre Channel adapters.
- Storage is IBM ESS supporting FCP, FICON and XRC.
- FC/9000 supports intermixed fibre protocol such as FCP and FICON.
- Shortwave GBIC on FC/9000.

- XRC is used for disaster recovery solution for zSeries.
- All database log data and database update data are transmitted.
- SDM is running on server 1 and at site 2.
- The storage has intermixed FICON, ESCON, FCP adapters.
- Channel Extender has the capability to transmit OS/390 data and Open system data.

Checklist

The following items should be considered to implement the solution:

- XRC planning including ESCON and network interface specifications of channel extenders checking.
- Support of channel interface such as ESCON, SCSI, FCP.
- Support of network interface such as ATM, T3/E3, IP Ethernet, ESCON.
- Connector type between servers and FC/9000, FC/9000 and storage.
- Fibre cable length.
- Connectivity testing between:
  - ESS and directors
  - servers and directors
  - Directors and channel extender
- Validate connectivity:
  - Test Server to storage on both paths
  - Tests of failover/failback operations
  - Setup LUNs on back up ESS and test from primary server, to quantify latency
- Users and password defined in IN-VSN.
- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration.
- Storage capacity and LUN assignments to each server.
- Disk storage features and microcode level for proposed configuration.
- INRANGE director LW GBIC count.
- SM FC cable laying and termination.
- INRANGE Director high availability features.
- E_D_TOV, R_A_TOV, and BB_Credit settings equal on all directors.
- Maximum distance for individual devices.

- Nickname assignments so we can quickly cross reference WWNs.
- Any application performance implications of longer distances.

Performance

The major performance consideration with a long distance solution is calculating the correct number of lines between sites. This number can only be determined by performing a detailed performance profile of the servers and storage that will be remote.

The following considerations are also recommended for better performance.

- Data compression.
- Channel extenders buffer credits.
- The 2042 supports buffer to buffer credits. As we already mentioned, buffer to buffer credits allow commands to be queued up in the buffer of the switch and this lessens the effect of the latency and improves aggregate performance.
- Conduct a detailed server performance profile.
- Spread Fibre Channel adapters as evenly as possible across all of the bays.
- Monitor the performance using the IN-VSN software.
- Collect MIB information to determine busy ports.

Scalability

Here are considerations for scalability:

- A channel extender has 6-slot and 12-slot chassis.
- You can add interfaces to scale up.
- One to four WAN modules for a channel extender.
- Up to eight ESCON connections for a channel extender.

Scalability rules for the SAN fabric of INRANGE Directors remain unchanged as in previous examples.

Security

Dual fabrics can protect you against user errors, such as a user erasing or changing the zoning information inappropriately. If you decide to add an ISL, ensure that all checks are done on the zone configuration changes from the management console.

The physical security of fiber connections and patch panels should be considered.

Here are considerations for security:

- All fabrics and storage are on the secure locations.
- Work with the best carrier company for lines.
- Any disk devices that do not support LUN masking are zoned to their respective servers.
- The WWNs of the HBAs Port names of servers have been defined.
- Disk Storage ESS LUN masking by WWN will allow each server access only to configured LUNs.
- In-VSN Enterprise Manager user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions.
- Only SAN administrators have access to the IN-VSN userid and passwords.
- As we have several name server zones defined, backups of the IN-VSN should be performed on a regular basis and at least when the information has changed.
- Remote access to Enterprise Manager configured to limit access to authorized workstations.
- A maintenance window is available, when OS/390, UNIX, Windows 2000, and Windows NT will be unavailable, so the hard zone can be implemented.
- When operating with primary and secondary sites we need to ensure all related SAN documentation is in a secure location that can be accessed at the recovery site. We also need to ensure we have enough userids of the correct type that are able to make any required changes to the zones.

"What if" failure scenarios

Here are some theoretical assumptions:

- **Transport network failure**

  The telco link failure is not planned for in the current solution, however a redundant link can be acquired from the vendor or a separate telco may be considered depending on the business recovery policy.

- **Automatic failover of service monitor**

  Other rules for the SAN fabric of INRANGE Directors remain unchanged as in previous sections.

12.8 Remote tape vaulting

For an existing large corporation with multiple sites, tape library resources can be consolidated to a separate site. This simplifies data movement logistics and centralizes backup software configurations. By doing this the cost of doing business is reduced as the infrastructure is efficiently utilized and less IT personnel required for backup data management. In addition, in the event of a disaster, the data is already located on tape in remote location and there is no longer a need to ship the data to another site.

In Figure 12-8 we show the basic layout for a remote tape vaulting solution using INRANGE directors.

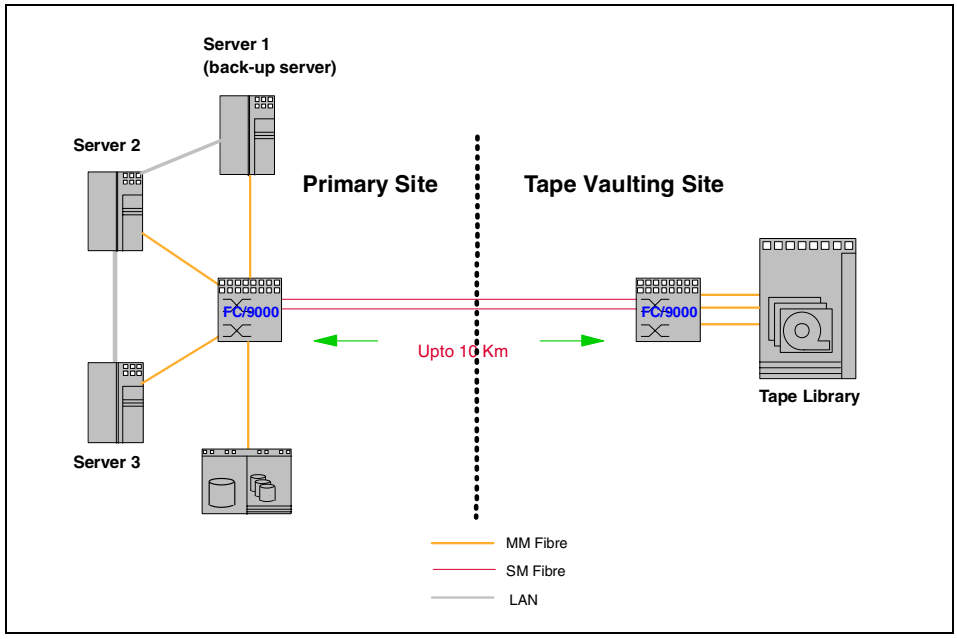


Figure 12-8 Remote tape vaulting

This solution extends the distance to up to 10 km apart, using long wave GBICs in the SAN fabric, and Single mode fiber cable interconnect. Distance up to 80 km are possible with extended long wave GBICs, but the solution is not recommended for large data movement.

Most tape libraries' Fiber Channel ports run in Arbitrated Loop (FC-AL) and they can be directly attached to the director as they will automatically sense the port type and it will be assigned as FC-AL.

Zoning can be established by WWN, so we can limit access of each server or each group of servers to specific tape drives.

Checklist

In addition to the items considered in the disk consolidation example, we must now consider:

- Host HBA supported for tape drive/library attachment
- Host tape device driver levels
- Host operating system levels compatible with tape library fiber requirements
- Backup S/W selection for types of data
- Backup hardware and software support fibre and this configuration
- Zone configuration allowing tape access to required servers
- Host software tape sharing capabilities
- Switches LAN connection, firmware and licence code where appropriate

Rules for the SAN fabric of INRANGE Directors remain unchanged as in previous sections.

Performance

Performance will basically depend on the number of HBAs available on each server and the number of links between sites. As with other local tape solutions, it is very important to consider data types, such as databases or small files to be transferred over the distance, and this varies from application to application.

And, you must consider which backup software is adequate for your solutions according to your backup policy and data types.

Depending on backup hardware and the drive interface, LTO or Magstar, the performance will also vary.

Scalability

Potentially we can scale the solution in terms of adding additional tape drives. However, physical space, cost, and performance should be considered.

Security

Zoning can be used to restrict access to devices on specific servers when required. This may also be used to change workload profiles to tape drives, for example, overnight may be the only time when the backup servers can see all the drives.

Proper tape management procedures will avoid servers contending for the same tape device. You would need to size the backup window available and the amount of data that you need to backup.

“What if” failure scenarios

These are some theoretical assumptions:

- DWDM optical channel card failure: As we configured the channels for high availability, there are redundant cards in the DWDM. If we lose one, the traffic will automatically be switched to the other and the channel will remain available. The failed card can be replaced concurrently.
- DWDM optical channel manager card failure: The optical channel manager card performs path high availability switching. There are two cards in each shelf, if one fails the other takes control and the operation is not affected. The failed card can be replaced concurrently.
- DWDM Optical Multiplexer failure or shelf backplane failure: The entire shelf will be unavailable. As we spread connections in different shelves, we will have at least half the channels of each type available. Operation will continue although performance may be affected.
- Dark fiber failure: As we configured for high availability, operation will continue to use the available pair with no performance impact. Customers may have additional considerations based on their business recovery policy to source the redundant dark fiber from another telco/source.
- If one of the lines between site A and site B fails, an alternate route will be used.
- ISL or switch failure: Traditionally tape failover is a manual operation. Multiple path devices are configured as several logical devices, one per path. Only one of these logical devices is made active. If there is a failure the application aborts and it can then be restarted using a different logical device. The latest levels of a tape device driver provide alternate pathing support and tape failover for Fibre Channel connections. With this support enabled if an error occurs the device driver will automatically initiate error recovery and the operation will continue to use the next logical path.
- Device link or device port failure: Only one tape drive is affected. Alternate path remains operational. Recovery may be manual or automatic depending on operating system and driver level as explained for ISL or switch failure.
- Switch port failures: GBICs are hot swappable. H_Ports can be moved to a spare port.
- User error trying to access more than two drives on the same link: Performance of all drives attached to the switch pair may be degraded. Switches performance view may be used to find what paths are carrying traffic.
- Tape drive failure in a single tape zone: An alternate zone should be made active to get access to a working device.
- If a cable fails between the director and ESS, an alternate route will be used.

- Regarding the director, if an FIO blade fails, we still have connectivity, as we have dual connections, but we would lose 50% bandwidth to any connected servers.
- If an FSW blade fails, there would be no effect, as the spare FSW would be automatically invoked.
- If an FCM module fails, there would be no effect, as the spare FCM module would be automatically invoked.
- If the backplane was damaged, we would lose connectivity to all servers at that site.
- If a server HBA fails, we lose up to 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.
- If an ESS HBA is unavailable, we have multiple other connections that will automatically be used.
- If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used.
- Director port failure: The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port.
- Fiber failure: Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.
- IN-VSN Enterprise Manager Server failure: No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.
- IN-VSN Server hard drive failure: Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.
- Director completely down, storage completely down, or site down (power, air conditioning, site damage); It will be recovered from alternative site.

12.9 Remote tape vaulting with redundancy

In Figure 12-9 we show an extension or variant of the tape vaulting solution.

In this solution the primary site and local site has a tape library. Data is written to both tape libraries, however in the event of failure on the tape library at the local site, data can be backed up and restored from the tape library at the remote site.

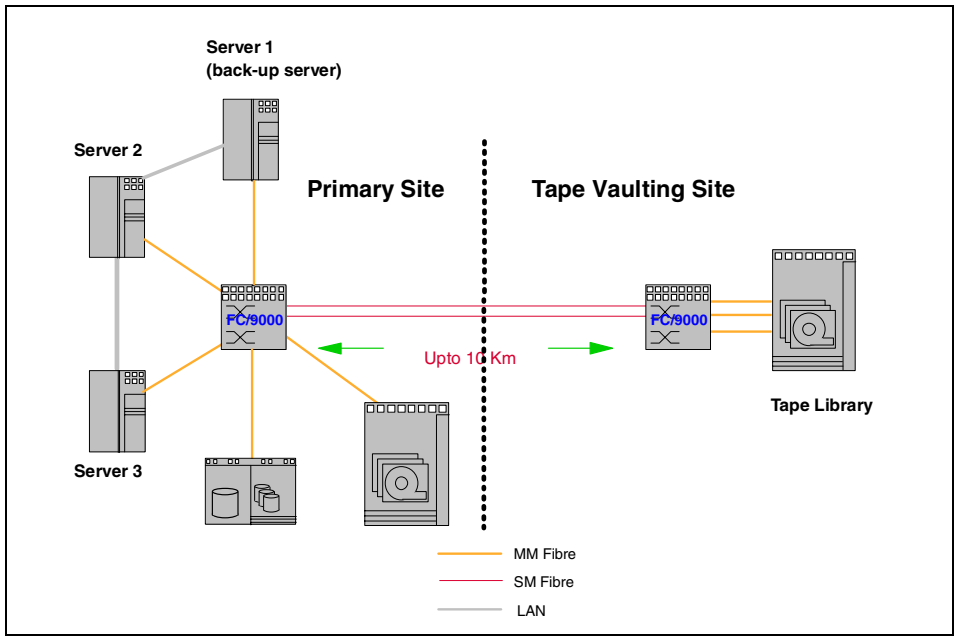


Figure 12-9 Remote tape vaulting with redundancy

Checklist

In addition to the items considered previously we must now consider:

- Host HBA supported for tape drive/library attachment
- Host tape device driver levels
- Host operating system levels compatible with tape library fiber requirements
- Backup S/W selection for types of data
- Backup hardware and software support fibre and this configuration
- Zone configuration allowing tape access to required servers
- Host software tape sharing capabilities
- Switches LAN connection, firmware and licence code where appropriate

Rules for the SAN fabric of INRANGE Directors remain unchanged as in previous sections.

Performance

Performance will basically depend on the number of HBAs available on each server and the number of links between sites. As with other local tape solutions, it is very important to consider data types such as databases or small files to be transferred over distance, and this varies from application to application.

You must consider which backup software is the most adequate for your solutions according to your backup policy and data types needs.

Depending on backup hardware and the drive interface, LTO or Magstar, the performance will also vary.

Scalability

Potentially we can scale the solution in terms of adding additional tape drives. However, physical space, cost, and performance should be considered.

Security

Zoning can be used to restrict access to devices to specific servers when required. This may also be used to change workload profiles to tape drives, for example, overnight may be the only time when the backup servers can see all the drives.

Proper tape management procedures will avoid servers contending for the same tape device. You would need to size the backup window available and the amount of data that you need to backup.

"What if" failure scenarios

Here are some theoretical assumptions:

- **DWDM optical channel card failure**

  As we configured the channels for high availability, there are redundant cards in the DWDM. If we lose one, the traffic will automatically be switched to the other and the channel will remain available. The failed card can be replaced concurrently.

- **DWDM optical channel manager card failure**

  The optical channel manager card performs path high availability switching. There are two cards in each shelf, if one fails the other takes control and the operation is not affected. The failed card can be replaced concurrently.

- **DWDM Optical Multiplexer failure or shelf backplane failure**

  The entire shelf will be unavailable. As we spread connections in different shelves, we will have at least half the channels of each type available. Operation will continue although performance may be affected.

- **Dark fiber failure**

  As we configured for high availability, operation will continue to use the available pair with no performance impact. Customers may have additional considerations based on their business recovery policy to source the redundant dark fiber from another telco/source.

- If one of the lines between site A and site B fails, an alternate route will be used.
- **ISL or switch failure**

  Traditionally tape failover is a manual operation. Multiple path devices are configured as several logical devices, one per path. Only one of these logical devices is made active. If there is a failure the application aborts and it can then be restarted using a different logical device. The latest levels of a tape device driver provide alternate pathing support and tape failover for Fibre Channel connections. With this support enabled if an error occurs the device driver will automatically initiate error recovery and the operation will continue to use the next logical path.
- **Device link or device port failure**

  Only one tape drive is affected. Alternate path remains operational. Recovery may be manual or automatic depending on operating system and driver level as explained for ISL or switch failure.
- **Switch port failures**

  GBICs are hot swappable. H_Ports can be moved to a spare port.
- **User error trying to access more than two drives on the same link**

  Performance of all drives attached to the switch pair may be degraded.
- **Tape drive failure in a single tape zone**

  An alternate zone should be made active to get access to a working device.
- If a cable fails between the director and ESS, an alternate route will be used
- Regarding the Director, if an FIO blade fails, we still have connectivity, as we have dual connections, but we would lose 50% bandwidth to any connected servers.
- If an FSW blade fails, there would be no effect, as the spare FSW would be automatically invoked.
- If an FCM module fails, there would be no effect, as the spare FCM module would be automatically invoked.
- If the backplane was damaged, we would lose connectivity to all servers at that site.
- If a server HBA fails, we lose up to 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.
- If an ESS HBA is unavailable, we have multiple other connections that will automatically be used.
- If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used.

- **Director port failure**

  The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port.

- **Fiber failure**

  Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.

- **IN-VSN Enterprise Manager Server failure**

  No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.

- **IN-VSN Server hard drive failure**

  Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.

- Director completely down, storage completely down, or site down (power, air conditioning, site damage); It will be recovered from alternative site.

13

McDATA distance storage solutions

In this chapter we will discuss some solutions that can be implemented using the McDATA products included in the IBM portfolio.

For product details of the McDATA portfolio products refer to:

- *IBM SAN Survival Guide*, SG24-6143
- *IBM SAN Survival Guide Featuring the McDATA Portfolio*, SG24-6149
- *Implementing an Open IBM SAN,* SG24-6116

This chapter presents various implementations and uses for Storage Area Network (SAN) in high availability graphically dispersed environments. The considerations for a high availability SAN design using IBM SAN Fibre Channel switch are covered in the following topics.

By reducing or eliminating single points of failure in the enterprise environment, SANs can help to improve overall availability of business applications. By utilizing highly available components and solutions as well as a fault-tolerant design, enterprises can achieve the availability needed to support 24x7 uptime requirements. High distances in these solutions, enable the SAN to cover a whole new host of IT availability issues, for example, we can deploy disaster tolerant solutions.

In network systems such as SANs, with their associated servers, fabric, and storage components, as well as software applications, downtime can occur even if parts of the system are highly available or fault tolerant. To improve business continuance under a variety of circumstances, SANs can incorporate redundant components, connections, software, and configurations to minimize or eliminate single points of failure.

With the emergence of the Internet and the proliferation of global e-business applications, more and more companies are implementing computing infrastructures specifically designed for continuous data and system availability. Today, even applications such as company e-mail have become mission critical for ongoing business operations. Faced with increased customer and internal user expectations, companies are currently striving to achieve at least 99.999 percent (the five “nines”) availability in their computing systems — a figure equivalent to less than 5.3 minutes of downtime a year. Additional downtime can severely impact business operations and cost valuable time, money, and resources.

To ensure the highest level of system uptime, companies are implementing reliable storage networks capable of boosting the availability of data for all the users and applications that need it. These companies typically represent the industries that demand the highest levels of system and data availability — the utilities and telecommunications sector, brokerages and financial service institutions, and a wide variety of service providers.

13.1 High-level availability objectives

Developing highly available networks involves identifying specific availability requirements and predicting what potential failures might cause outages. The first step is to clearly define availability objectives, which can vary widely from company to company and even within segments of the same company. In some environments, no disruption can be tolerated, while other environments might be only minimally affected by short outages. As a result, availability is a function of the frequency of outages (caused by unplanned failures or scheduled maintenance and upgrades) and the time to recover from such outages.

Many companies are addressing their availability requirements by implementing networks fabrics of Fibre Channel devices designed to provide high-performance storage environments. These flexible SANs are based on the following principles:

- A thorough understanding of availability requirements throughout the enterprise
- A flexible design that incorporates fault tolerance through redundancy and mirroring

- Simplified fault monitoring, diagnostics, and repair capabilities to ensure fast recovery
- A minimal amount of human intervention required during failover events
- A reliable backup and recovery plan to account for a wide variety of contingencies

13.2 Disk consolidation with a remote disk

This remote disk consolidation solution is an extension of the most basic SAN solutions of disk consolidation (Figure 13-1).

Disk consolidation provides a means for storage administrators to take advantage of the capabilities of Fibre Channel fabric-based Storage Area Networks to organize storage in more flexible and efficient ways, easing many aspects of storage administration.

In an attempt to extend the distance between the storage and the servers, and in order to provide business benefits like consolidation, physical separation of disk storage from the server across sites, is necessary. This solution extends the distance to up to 10 km apart, using Long wave GBICs in the SAN fabric, and Single mode Fiber cable interconnect.

The SAN fabric is comprised of:

- IBM 2031- 016 McDATA FC switch
- IBM 2031 - 032 McDATA FC switch
- IBM 2032 - 064 McDATA Enterprise FC director (used in this example)

The host servers are connected to the SAN fabric via SW GBIC's using MM Fibre cable.

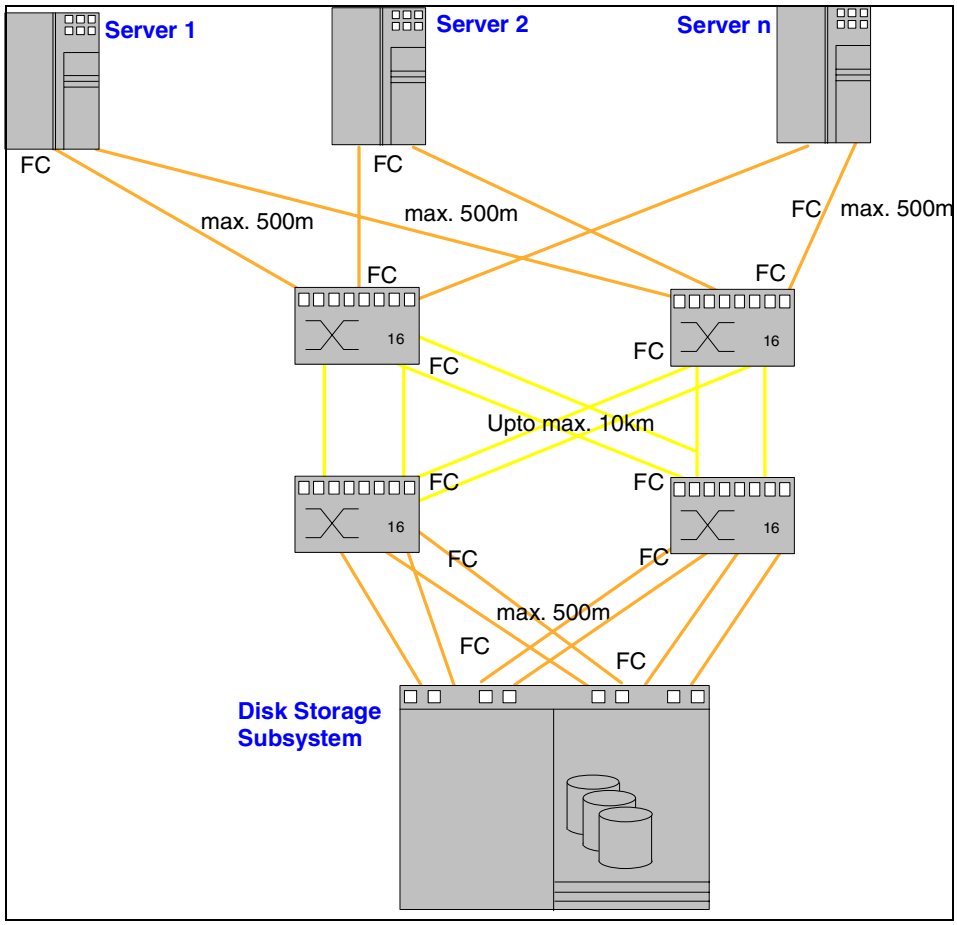


Figure 13-1 Remote disk consolidation

Checklist

We checked the following items:

- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and microcode level for proposed configuration
- McDATA switches and director LW GBIC count
- The ISL links — SM fiber cable
- SM FC cable laying and termination
- McDATA Director high availability features

- Nickname assignments so we can quickly cross reference WWNs to devices
- ESS LUN definitions done
- EFC Server and Manager Users and password defined

Performance

In this simple implementation, performance will basically depend on the number of HBAs available on each server and the number of storage connections and most important the distance between storage and servers. It is important to consider application behavior over the distance, and this varies from application to application and we have assumed that the performance is adequate for applications running in the servers.

The McDATA director supports any to any connectivity so it will not affect performance by itself. Latency in the director is around 2.6 microseconds, this is not relevant compared to storage devices response time in the millisecond range.

Knowing the requirements of our servers we can decide whether it is enough or if we need to add more connections or even more storage devices.

Without having the exact requirements we may consider that for a high profile server, a server to storage ratio of 6:1 is acceptable. This is only a starting point, we will then implement some measurement system or use statistical data to decide whether we need to add more connections or we have more bandwidth than required.

Scalability

The McDATA director supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation. With four ISL between the directors on the host side and the directors on the storage side, we can have 26 servers with two HBAs each, and still keep four spare ports in a fully populated director.

Security

Here are some security considerations:

- Disk Storage ESS LUN masking by WWN will allow each server access only to configured LUNs.
- EFC Manager user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions .
- Remote access to EFC Manager configured to limit access to authorized workstations.

- Physical director security — locked cabinet, restricted access site.

“What if” failure scenarios

These are some theoretical assumptions:

- **Host HBA failure**

  SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. EFC Manager user with Product Administrator rights and ESS Specialist access are required.

- **Storage host adapter failure**

  The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce available bandwidth by 25%. For an average workload and five servers as shown it should not impact performance. When the host adapter is replaced we need to update zoning with the new WWN. We also need to reconfigure the OS and SDD to pickup the new path information. EFC Manager user with Product Administrator rights and OS root access are required.

- **Director port failure**

  The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port. Physical access to director and EFC Manager user with maintenance rights is required. OS root access may be required.

- **Fiber failure**

  Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.

- **EFC Server failure**

  No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.

- **EFC Server hard drive failure**

  Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.

- **Director completely down, storage completely down, or site down** (power, air conditioning, site damage)

  This will cause an interruption in normal operation.

- **Physical damage to storage causing data loss (fire, flood)**

  We will need to restore data from backup tapes.

13.3 Two sites up to 10 km apart

In Figure 13-2, we show a basic SAN fabric distance extension solution between two sites up to 10 km apart. Each site has a fully redundant infrastructure at the SAN fabric level comprising of two IBM 2032 Enterprise directors, and the directors are interlinked across the sites via ISL link using the LW GBICs and SM fiber cables. The host servers are connected to the SAN fabric via SW GBIC's using MM Fibre cable.

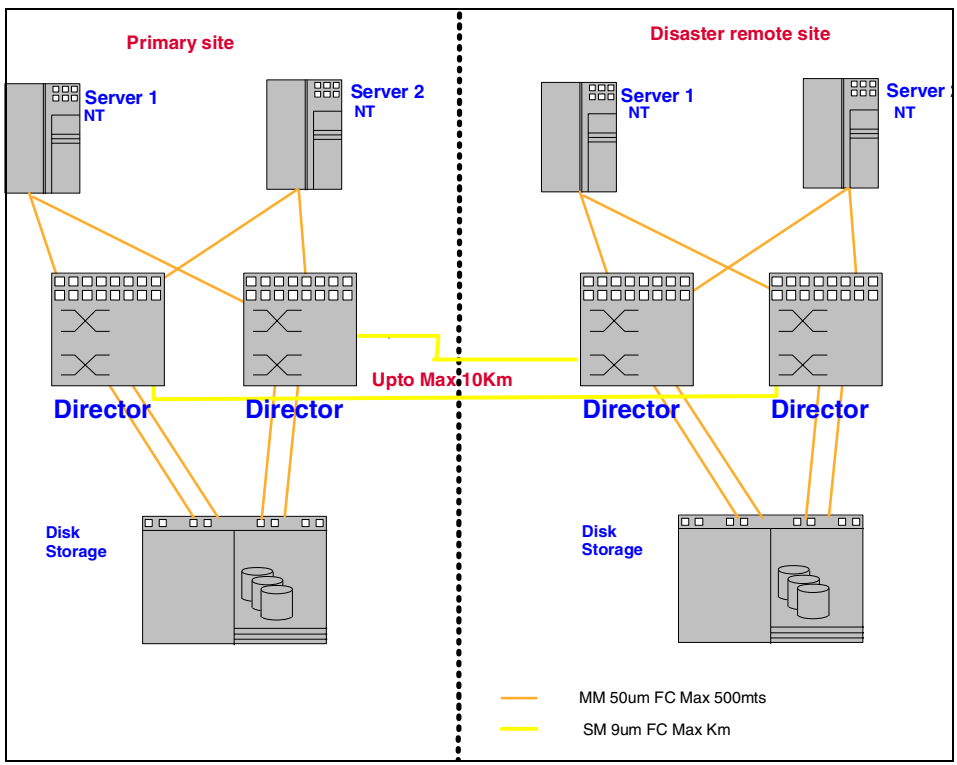


Figure 13-2 SAN distance extension up to 10 km

The solution can be used for accessing a database at remote site as well as mirroring data at an alternative site up to 10 km.

One can assume that a greater distance between IT sites results in greater security against wide spread disaster. However, increased distance has its price in terms of interconnection cost, business relocation effort, and network cost. Before you start to expand your storage network, you have to consider if the following solution can meet your business purpose.

The SAN fabric is comprised of:

- IBM 2031- 016 McDATA FC switch
- IBM 2031 - 032 McDATA FC switch
- IBM 2032 - 064 McDATA Enterprise FC director (We have used this in the example)

Note: McDATA has tested a distance of up to 20 km using the LW GBICs between the directors.

We have assumed the following for our solution:

- Both sites have SAN infrastructure.
- Server to IBM 2032 Director and IBM 2032 Director to storage use multiple mode fibre.
- UNIX servers in production which are being used for mission critical production need data mirroring for disaster recovery.
- However, NT servers at site A need to access database in storage B at site B, in addition to using data in storage A.
- UNIX servers at site B are in standby mode, and being used for backup servers at ordinary times.
- Single mode FC cables are used as ISL links between IBM 2032 at each site.
- Software mirroring is used in UNIX servers.
- No PPRC proposed.

Checklist

The following items should be considered to implement the solution:

- Fibre Channel HBA supported on servers
- Operating systems supported S/W mirroring
- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration

- Storage capacity and LUN assignments to each server
- Disk storage features and microcode level for proposed configuration
- McDATA switches and director LW GBIC count
- The ISL links - SM fiber cable
- SM FC cable laying and termination
- McDATA Director high availability features
- Nickname assignments so we can quickly cross reference WWNs to devices
- ESS LUN definitions done
- EFC Server and Manager Users and password defined

Performance

The major performance consideration with a long distance solution is calculating the correct number of lines between sites. This number can only be determined by performing a detailed performance profile of the servers and storage that will be remote. Over estimating the number of lines will increase costs dramatically, under sizing the number of lines will dramatically effective the performance of the SAN.

It is vital that detailed performance data is available prior to sizing the number of lines required.

Typically, latency will increase over long distances, a good rule of thumb is 4.8 microseconds per kilometer. The default E_D_TOV and R_A_TOV values do not need to be modified for this distance.

Scalability

The McDATA director supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation. With four ISLs between the directors on the host side and the directors on the storage side, we can have 26 servers with two HBAs each, and still keep four spare ports in a fully populated director.

Security

Here are some security considerations:

- Disk Storage ESS LUN masking by WWN will allow each server access only to configured LUNs.
- EFC Manager user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions.

- Remote access to EFC Manager configured to limit access to authorized workstations.
- Physical director security — locked cabinet, restricted access site.
- Our solution assumes we own private lines between sites so encryption is not required.
- For leased lines or managed services where lines are shared, encryption is normally an option available from the service provider.

"What if" failure scenarios

Here are some theoretical assumptions:

- **Host HBA failure**

  SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. EFC Manager user with Product Administrator rights and ESS Specialist access are required.

- **Storage host adapter failure**

  The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce available bandwidth by 25%. For an average workload and five servers as shown it should not impact performance. When the host adapter is replaced we need to update zoning with the new WWN. We also need to reconfigure the OS and SDD to pickup the new path information. EFC Manager user with Product Administrator rights and OS root access are required.

- **Director port failure**

  The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port. Physical access to director and EFC Manager user with maintenance rights is required. OS root access may be required.

- **Fiber failure**

  Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.

- **EFC Server failure**

  No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.

- **EFC Server hard drive failure**

  Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.

- **Director completely down, storage completely down, or site down** (power, air conditioning, site damage)

  This will cause an interruption in normal operation.

- **Physical damage to storage causing data loss (fire, flood)**

  We will need to restore data from backup tapes.

13.4 Two sites up to 100 km apart

In Figure 13-3 we show a basic SAN fabric distance extension solution between two sites up to 10 km apart. Each site has a fully redundant infrastructure at the SAN fabric level comprising of two IBM 2032 Enterprise directors, the directors are interlinked across the sites via ISL link using the LW GBICs and SM fiber cables and repeaters. The host servers are connected to the SAN fabric via SW GBIC's using MM Fibre cable.

One can assume that a greater distance between IT sites results in greater security against wide spread disaster. However, increased distance has its price in terms of interconnection cost, business relocation effort, and network cost. Before you start to expand your storage network, you have to consider if the following solution can meet your business purpose.

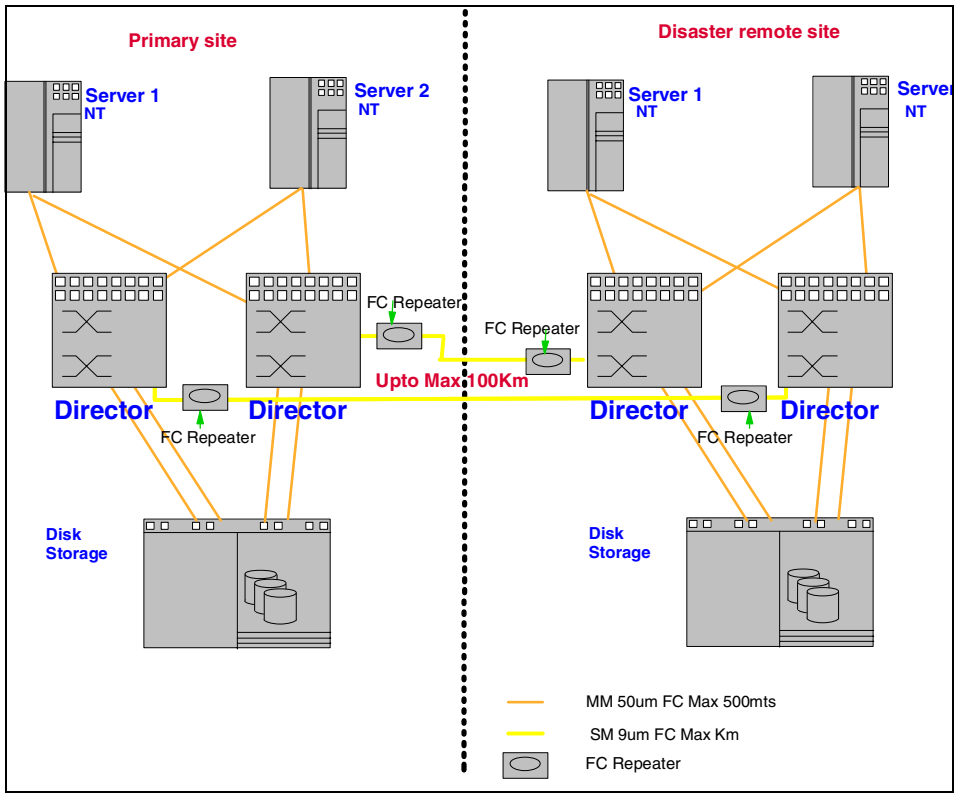


Figure 13-3 SAN distance extension up to 100 km with repeaters

The SAN fabric is comprised of:

- IBM 2031- 016 McDATA FC switch
- IBM 2031 - 032 McDATA FC switch
- IBM 2032 - 064 McDATA Enterprise FC director (We have used this in the example)

Note: McDATA has tested a distance of up to 100 using the LW GBICs and repeaters at every 20 km between the directors.

We have assumed the following for our solution:

- Both sites have SAN infrastructure.
- Server to IBM 2032 Director and IBM 2032 Director to storage use multiple mode fibre.
- SAN repeaters are installed and connected every 20 km.

- And single mode FC cables are used as ISL links between IBM 2032 and repeaters across sites.
- Software mirroring is used in UNIX servers.
- No PPRC proposed.

Checklist

The following items should be considered to implement the solution.

- Fibre Channel HBA supported on servers.
- Operating systems supported S/W mirroring.
- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration.
- Storage capacity and LUN assignments to each server.
- Disk storage features and microcode level for proposed configuration.
- McDATA switches and director LW GBIC count.
- Repeater count verified for the distance in every ISL link.
- Repeaters installed connected and tested.
- The ISL links - SM fiber cable.
- SM FC cable laying and termination.
- McDATA Director high availability features.
- Nickname assignments so we can quickly cross reference WWNs to devices.
- ESS LUN definitions done.
- EFC Server and Manager Users and password defined.

Performance

The major performance consideration with a long distance solution is calculating the correct number of lines between sites. This number can only be determined by performing a detailed performance profile of the servers and storage that will be remote. Over estimating the number of lines will increase costs dramatically, under sizing the number of lines will dramatically effective the performance of the SAN.

It is vital that detailed performance data is available prior to sizing the number of lines required.

Typically, latency will increase over long distances, a good rule of thumb is 4.8 microseconds per kilometer. The default E_D_TOV and R_A_TOV values do not need to be modified for this distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the Directors across sites (the ISL links).

Scalability

The McDATA director supports the concurrent addition of port cards, so we can scale this solution by adding more servers or storage devices without disrupting operation. With four ISL between the directors on the host side and the directors on the Storage side, we can have 26 servers with two HBAs each, and still keep four spare ports in a fully populated director.

Security

The following are some security considerations:

- Disk Storage ESS LUN masking by WWN will allow each server access only to configured LUNs
- EFC Manager user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions
- Remote access to EFC Manager configured to limit access to authorized workstations
- Physical director security — locked cabinet, restricted access site.
- Our solution assumes we own private lines between sites so encryption is not required.
- For leased lines or managed services where lines are shared, encryption is normally an option available from the service provider.
- Physical repeater security — locked cabinet, restricted access site

“What if” failure scenarios

Here are some theoretical assumptions:

- **Host HBA failure**

  SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. EFC Manager user with Product Administrator rights and ESS Specialist access are required.

- **Storage host adapter failure**

  The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce available bandwidth by 25%. For an average workload and

five servers as shown it should not impact performance. When the host adapter is replaced we need to update zoning with the new WWN. We also need to the reconfigure OS and SDD to pickup the new path information. EFC Manager user with Product Administrator rights and OS root access are required.

- **Director port failure**

  The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port. Physical access to director and EFC Manager user with maintenance rights is required. OS root access may be required.

- **Repeater failure**

  Impact will depend on whether if alternate ISL links are available or not. Only action required is repeater replacement. Physical access to repeater and repeater site required.

- **Fiber failure**

  Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.

- **EFC Server failure**

  No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.

- **EFC Server hard drive failure**

  Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.

- **Director completely down, Storage completely down, or site down** (power, air conditioning, site damage)

  This will cause an interruption in normal operation.

- **Physical damage to Storage causing data loss (fire, flood)**

  We will need to restore data from backup tapes.

13.5 Two sites - point-to-point DWDM solution

In Figure 13-4 we show a basic high availability SAN design for a multiple servers deployed with dual directors and redundant fabric at each site. This design also uses the DWDM equipment for inter-site ISL links between the directors.

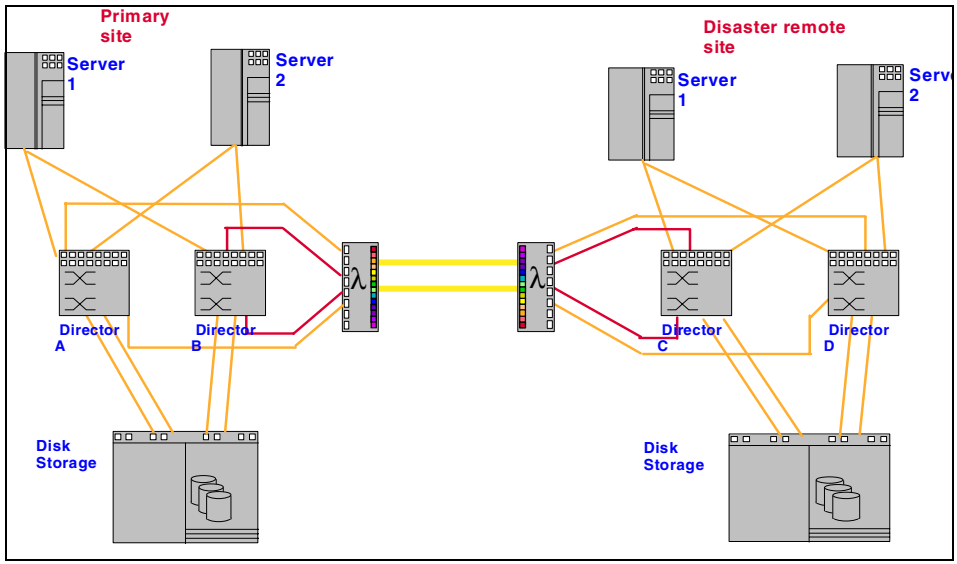


Figure 13-4 Point-to-point DWDM solution - two sites

The SAN fabric is comprised of:

- IBM 2031- 016 McDATA FC switch
- IBM 2031 - 032 McDATA FC switch
- IBM 2032 - 064 McDATA Enterprise FC director (We have used this in the example)

A dual fabric SAN is a topology where you have two independent fabrics that connect the same hosts and storage devices, which are mutually exclusive. This design is not highly scalable with 16 port switches as all hosts and storage must be connected to both switches to achieve high availability, however it does allow for inter site traffic, so we have used directors in this example. We have labeled each site here, one is the primary site, the other the DR site, as these names imply this infrastructure would give the transport mechanism to enable offsite recovery. Maybe with hot standby servers to reduce recovery times. The method for replicating the data could be performed at a host level, for example disk mirroring or at a storage level. The data would be replicated by host software in this example.

Within each site server to storage availability is provided with this dual director, dual fabric design within each site. Dual HBAs are installed in each host and each storage device must have at least two ports. Failover for a failed path or even a failed switch is dependent on host failover software, namely, IBM Subsystem Device Driver (SDD).

In our example, SAN AD (Directors A and D) will be discrete fabric from SAN BC (Director B and C) fabric. Each fabric comprises of 2 director, one at each site.

The components used are as follows:

- SAN Fabric
  - IBM 2032 - 064 McDATA Enterprise FC directors
- DWDM
  - Two DWDM's linked via dark fibre
- Servers
  - Four IBM Netfinity Servers running on MS Windows NT, two Netfinity PCI Adapters and four HBA cards
- Storage
  - Two ESS 2105-F20 with native Fibre Channel Adapter
- Software
  - IBM Subsystem Device Driver (SDD)
  - IBM StorWatch Specialist
  - IBM ESS Expert (for historical disk performance data)

Checklist

We checked the following items:

- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and microcode level for proposed configuration
- McDATA switches and director LW GBIC count
- The ISL link count to the DWDM equipment at each site
- SM FC cable laying and termination
- McDATA Director high availability features
- E_D_TOV, R_A_TOV, and BB_Credit settings equal on all directors
- Maximum distance for individual devices
- Nickname assignments so we can quickly cross reference WWNs to devices
- Configuration and count of the client interface adapters in the DWDM equipment
- Configuration and count of the transport network adapters in the DWDM, ensuring same frequency band and channel

- Connectivity testing between
- ESS to switches
- Servers to switch
- Switches to DWDM
- EFC Server and Manager Users and password defined
- Pre-installation testing of the DWDM transport network for configuration and performance

Performance

Typically, for a low performance server, the recommended server to storage connection ratio is 12 to 1, and for a high performance server, the server to storage ratio is 6 to 1. Low performance servers are typically made up of file and print servers whereas high performance servers are application servers.

The DWDM equipment is transparent to the subsystems that are using it. There are no internal queues or busy conditions, so it does not affect performance. The performance factors we need to consider in this solution are the number of ISLs and distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the director and the DWDM equipment.

Scalability

Most DWDM support multiple I/O interfaces for client equipment and support up to 32 separate channels. DWM equipment also allow for mixing different protocol interfaces of client equipment to a single physical DWDM equipment.

Scalability rules for the SAN fabric of McDATA switches and directors remain unchanged as in previous examples.

Security

Dual fabrics can protect you against user errors, such as a user erasing or changing the zoning information inappropriately. The zoning information is separate for each fabric SAN AD and SAN BC, so when changing the zoning in one fabric it does not automatically propagated into the other fabric.

If you decide to add an ISL, ensure that all checks are done on the zone configuration changes from the management console.

The DWDM equipment provide a Web based management software that can be accessed by any workstation connected to the same LAN. Different user levels are provided for administrators, operators or observers. Different passwords must be set to limit access.

The physical security of fiber connections and patch panels should be considered.

"What if" failure scenarios

Here are some theoretical assumptions:

- **Host HBA failure**: SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. EFC Manager user with Product Administrator rights and ESS Specialist access are required.
- **Storage host adapter failure:** The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce available bandwidth by 25%. For an average workload and five servers as shown it should not impact performance. When the host adapter is replaced we need to update zoning with the new WWN. We also need to reconfigure the OS and SDD to pickup the new path information. EFC Manager user with Product Administrator rights and OS root access are required.
- **Director port failure:** The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port. Physical access to director and EFC Manager user with maintenance rights is required. OS root access may be required.
- **Fiber failure:** Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.
- **EFC Server failure:** No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.
- **EFC Server hard drive failure**: Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.
- **Director completely down, storage completely down, or site down** (power, air conditioning, site damage): Will cause an interruption in normal operation.

- **Physical damage to Storage causing data loss (fire, flood)**: We will need to restore data from backup tapes.
- **DWDM optical channel card failure:** As we configured the channels for high availability, there are redundant cards in the DWDM. If we lose one, the traffic will automatically be switched to the other and the channel will remain available. The failed card can be replaced concurrently.
- **DWDM optical channel manager card failure**: The optical channel manager card performs path high availability switching. There are two cards in each shelf, if one fails the other takes control and the operation is not affected. The failed card can be replaced concurrently.
- **DWDM Optical Multiplexer failure or shelf backplane failure**: The entire shelf will be unavailable. As we spread connections in different shelves, we will have at least half the channels of each type available. Operation will continue although performance may be affected.
- **Dark fiber failure:** As we configured for high availability, operation will continue using the available pair with no performance impact. Customers may have additional considerations based on their business recovery policy to source the redundant dark fiber from another telco/source.

DWDM considerations

We have shown a point-to-point topology here. We would implement the DWDM solution as a protected fiber to ensure availability. In this example we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation, here we have assumed an light workload. We show here an enterprise deployment that enable data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one switch and or director in each location. This is an excellent approach for availability and redundancy. It gives you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits, however this is not a DWDM consideration but more to do with the general SAN solution design over a distance. The DWDM is a core architecture deployment and because of it's independence from protocol it can be used lots of other traffic, it extends beyond the SAN environment. The distance we have shown here is up to 100 km between nodes.

13.6 Two sites: Point-to-point DWDM with ESS PPRC

In Figure 13-5 we show a basic high availability SAN design for multiple servers deployed with dual directors and redundant fabric at each site.

This design also uses the DWDM equipment for inter-site ISL links between the directors. The DWDM shown here also facilitates the inter-site ESCON traffic for the PPRC.

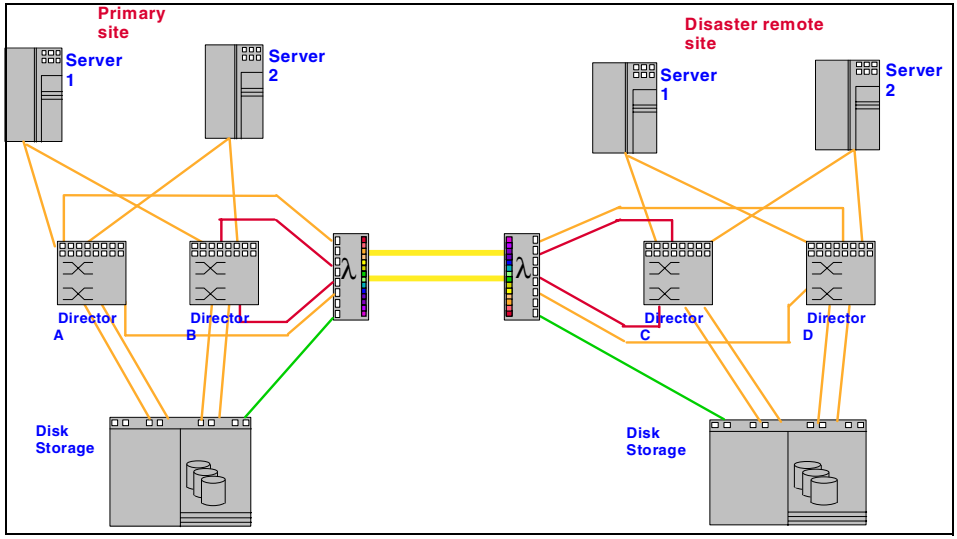


Figure 13-5 Point-to-point DWDM with PPRC solution

The SAN fabric is comprised of:

- IBM 2031- 016 McDATA FC switch
- IBM 2031 - 032 McDATA FC switch
- IBM 2032 - 064 McDATA Enterprise FC director (We have used this in the example)

A dual fabric SAN is a topology where you have two independent fabrics that connect the same hosts and storage devices, which are mutually exclusive. This design is not highly scalable with 16 port switches as all hosts and storage must be connected to both switches to achieve high availability, however it does allow for inter site traffic, so we have used directors in this example. We have labeled each site here, one is the primary site, the other the DR site, as these names imply this infrastructure would give the transport mechanism to enable offsite recovery. Maybe with hot standby servers to reduce recovery times. The method

for replicating the data could be performed at a host level, for example disk mirroring or at a storage level. This example shows data replication being ESS and PPRC. The PPRC connection today is via ESCON, this can utilize the DWDM for intersite transport alongside the Fibre Channel.

Within each site server to storage availability is provided with this dual director, dual fabric design within each site. Dual HBAs are installed in each host and each storage device must have at least two ports. Failover for a failed path or even a failed switch is dependent on host failover software, namely, IBM Subsystem Device Driver (SDD).

In our example, SAN AD (Directors A and D) will be discrete fabric from SAN BC (Director B and C) fabric. Each fabric comprises of two directors, one at each site.

The components used are as follows:

- SAN Fabric
  - IBM 2032 - 064 McDATA Enterprise FC directors
- DWDM
  - Two DWDM's linked via dark fibre
- Servers
  - Four IBM Netfinity Servers running on MS Windows NT, two Netfinity PCI Adapters and four HBA cards
- Storage
  - Two ESS 2105-F20 with native Fibre Channel Adapter and ESCON adapters for PPRC connectivity
- Software
  - IBM Subsystem Device Driver (SDD)
  - IBM StorWatch Specialist
  - IBM ESS Expert (for historical disk performance data)

Checklist

We checked the following items:

- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and microcode level for proposed configuration
- McDATA switches and director LW GBIC count
- The ISL link count to the DWDM equipment at each site

- SM FC cable laying and termination
- McDATA Director high availability features
- E_D_TOV, R_A_TOV, and BB_Credit settings equal on all directors
- Maximum distance for individual device Nickname assignments so we can quickly cross reference WWNs to devices
- Configuration and count of the client interface adapters in the DWDM equipment
- Configuration and count of the transport network adapters in the DWDM, ensuring same frequency band and Channel
- Connectivity testing between
- ESS to switches
- Servers to switch
- Switches to DWDM
- ESS Escon HBAs and the DWDM
- Validate connectivity
- Test Server to storage on both paths
- Tests of failover/fail back operations
- Test inter-site connectivity via DWDM
- Setup LUNs on back up ESS and test from primary server, to quantify latency
- Test PPRC - Establish paths and Test PPRC funtionality - qualify timings
- EFC Server and Manager Users and password defined
- Pre-installation testing of the DWDM transport network for configuration and performance

Performance

Typically, for a low performance server, the recommended server to storage connection ratio is 12 to 1, and for a high performance server, the server to storage ratio is 6 to 1. Low performance servers are typically made up of file and print servers whereas high performance servers are application servers.

The DWDM equipment is transparent to the subsystems that are using it. There are no internal queues or busy conditions, so it does not affect performance. The performance factors we need to consider in this solution are the number of ISLs and distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the director and the DWDM equipment.

Scalability

Most DWDM support multiple I/O interfaces for client equipment and support up to 32 separate channels. DWM equipment also allow for mixing different protocol interfaces of client equipment to a single physical DWDM equipment.

Scalability rules for the SAN fabric of McDATA switches and Directors remain unchanged as in previous examples.

Security

Dual fabrics can protect you against user errors, such as a user erasing or changing the zoning information inappropriately. The zoning information is separate for each fabric SAN AD and SAN BC, so when changing the zoning in one fabric it does not automatically propagated into the other fabric.

If you decide to add an ISL, ensure that all checks are done on the zone configuration changes from the management console.

The DWDM equipment provide a Web based management software that can be accessed by any workstation connected to the same LAN. Different user levels are provided for administrators, operators or observers. Different passwords must be set to limit access.

The physical security of fiber connections and patch panels should be considered.

“What if” failure scenarios

Here are some theoretical assumptions:

- **Host HBA failure**: SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. EFC Manager user with Product Administrator rights and ESS Specialist access are required.
- **Storage host adapter failure:** The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce available bandwidth by 25%. For an average workload and five servers as shown it should not impact performance. When the host adapter is replaced we need to update zoning with the new WWN. We also need to reconfigure the OS and SDD to pickup the new path

information. EFC Manager user with Product Administrator rights and OS root access are required.

- **Director port failure**: The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port. Physical access to director and EFC Manager user with maintenance rights is required. OS root access may be required.
- **Fiber failure**: Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.
- **EFC Server failure**. No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.
- **EFC Server hard drive failure:** Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.
- **Director completely down, Storage completely down, or site down** (power, air conditioning, site damage): Will cause an interruption in normal operation.
- **Physical damage to Storage causing data loss (fire, flood)**: We will need to restore data from backup tapes.
- **DWDM optical channel card failure:** As we configured the channels for high availability, there are redundant cards in the DWDM. If we lose one, the traffic will automatically be switched to the other and the channel will remain available. The failed card can be replaced concurrently.
- **DWDM optical channel manager card failure**: The optical channel manager card performs path high availability switching. There are two cards in each shelf, if one fails the other takes control and the operation is not affected. The failed card can be replaced concurrently.
- **DWDM Optical Multiplexer failure or shelf backplane failure:** The entire shelf will be unavailable. As we spread connections in different shelves, we will have at least half the channels of each type available. Operation will continue although performance may be affected.
- **Dark fiber failure:** As we configured for high availability, operation will continue using the available pair with no performance impact. Customers may have additional considerations based on their business recovery policy to source the redundant dark fiber from another telco/source.

DWDM considerations

We have shown a point-to-point topology here. We would implement the DWDM solution as a protected fiber to ensure availability. In this example we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation, here we have assumed a light workload.

We show here an enterprise deployment that enables data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one switch and or director in each location. This is an excellent approach for availability and redundancy. It give you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits, however this is not a DWDM consideration but more to do with the general SAN solution design over a distance. The DWDM is a core architecture deployment and because of its independence from protocol it can be used for other traffic, as it extends beyond the SAN environment. The distance we have shown here is up to 100 km between nodes.

The solution shown here also implements IBM's PPRC between the two ESS subsystems, this will today be transported over the ESCON protocol. The ESCON traffic can be transported over the already deployed DWDM architecture. We show a single pair of 62.5 ESCON connections here, connecting the ESCON ports on the ESS to the DWDM, again this will be dependent upon traffic expectations and should follow the ESCON channel sizing guidelines for PPRC implementation.

In order to maintain performance at extended distances, we need to increase the number of buffers on each interconnecting port to compensate for the number of frames that are in transit. Configuring the director ports connected to the DWDM for 10 to 100 km provides 60 buffers, that is enough for this distance. Scalability The DWDM comes in "shelves". Each shelf provides four high available channels. Up to eight shelves can be installed for a total of 32 high available channels. For additional number of channels, we would need to install another DWDM and also need another two pairs of fibers.

13.7 Multiple site ring DWDM solution

Here we describe the IBM 2032 with a multi node DWDM configuration that spans four sites and provisions optical services. There are four sites, with each connected over a DWDM channel that includes dual paths for transmitting and receiving. This is shown in Figure 13-6.

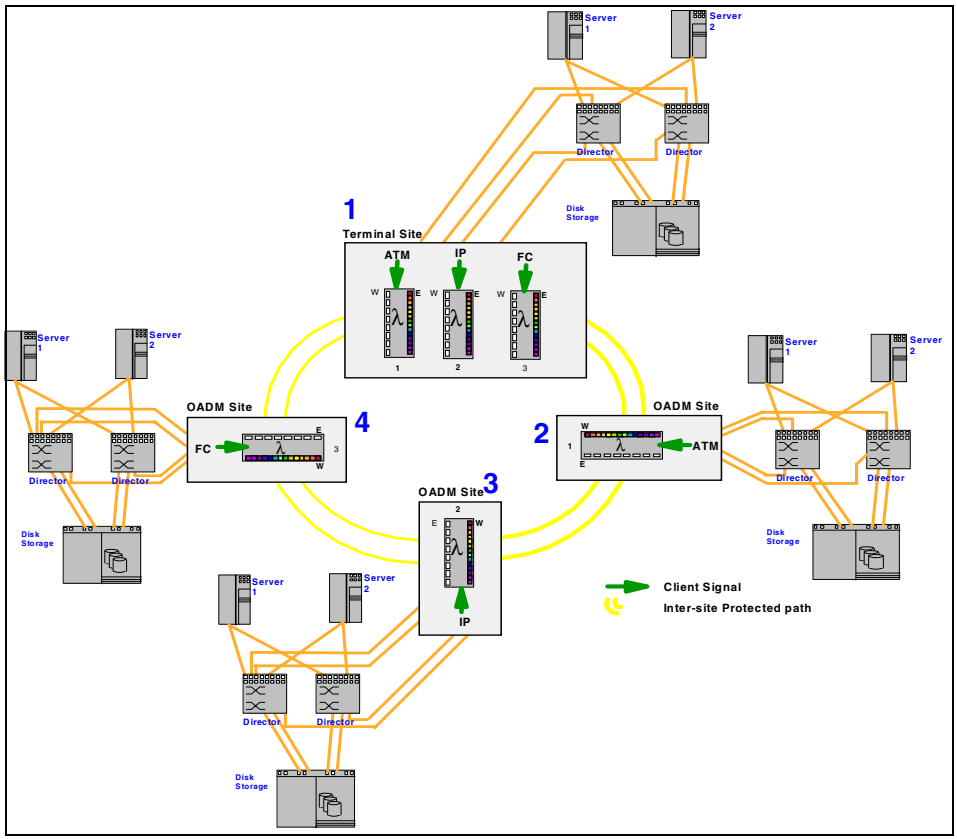


Figure 13-6 Multiple site: Ring topology DWDM solution

Each path has its own wavelength. The DWDM pass through feature enables non-contiguous sites to connect over an intermediate site as if they were directly connected. The only additional overhead of the pass through is the minimal latency (5 microseconds/km) of the second link. The pass through has no overhead since it is a passive device. This fabric would logically appear as a fully meshed topology.

Each of the links can operate in protected mode, which provides a redundant path in the event of a link failure. In most cases, link failures are automatically detected within 50 msec. In this case, the two wavelengths of the failed link reverse directions and reach the target port at the opposite side of the ring. If the link between DWDM 1 and 4 fails, the transmitted wavelength from 4 to 1 would reverse direction and reach 1 through 3 and 2. The transmitted wavelength from 1 to 4 would also reverse direction and reach 4 through 2 and 3.

Calculating the distance between nodes in a ring depends on the implementation of the protected path scheme. For instance, if the link between DWDM 2 and 3 fails, the path from 1 to 3 would be 1 to 2, back from 2 to 1 (due to the failed link), 1 to 4, and finally 4 to 3. This illustrates the need to utilize the entire ring circumference (and more, in a configuration with over four nodes) for failover.

Another way to calculate distance between nodes is to set up the protected path in advance (in the reverse direction) so the distance is limited to the number of hops between the two nodes. In either case, the maximum distance between nodes determines the maximum optical reach. An example of this specification is 80 to 100 km for a maximum distance between nodes and 160 to 400 km for maximum ring size. These distances should continue to increase as fiber optic technology advances.

The SAN fabric is comprised of:

- IBM 2031- 016 McDATA FC switch
- IBM 2031 - 032 McDATA FC switch
- IBM 2032 - 064 McDATA Enterprise FC director (We have used this in the example)

A dual fabric SAN is a topology where you have two independent fabrics that connect the same hosts and storage devices, which are mutually exclusive. This design is not highly scalable with 16 port switches as all hosts and storage must be connected to both switches to achieve high availability, however it does allow for inter site traffic, so we have used directors in this example. We have labeled each site here, one is the primary site, the other the DR site, as these names imply this infrastructure would give the transport mechanism to enable offsite recovery. Maybe with hot standby servers to reduce recovery times. The method for replicating the data could be performed at a host level, for example disk mirroring or at a storage level. The data would be replicated by host software in this example.

Within each site server to storage availability is provided with this dual director, dual fabric design within each site. Dual HBAs are installed in each host and each storage device must have at least two ports. Failover for a failed path or even a failed switch is dependent on host fail-over software, namely, IBM Subsystem Device Driver (SDD). In our example Figure 13-4, SAN AD (Directors A and D) will be discrete fabric from SAN BC (Director B and C) fabric. Each fabric comprises of two directors, one at each site.

The components used are as follows:

- SAN Fabric
  - IBM 2032 - 064 McDATA Enterprise FC directors
- DWDM
  - Two DWDM's linked via dark fibre
- Servers
  - Four IBM Netfinity Servers running on MS Windows NT, two Netfinity PCI Adapters and four HBA cards
- Storage
  - Two ESS 2105-F20 with native Fibre Channel Adapter
- Software
  - IBM Subsystem Device Driver (SDD)
  - IBM ESS StorWatch Specialist
  - IBM ESS Expert (for historical disk performance data)

Checklist

We checked the following items at each site:

- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and microcode level for proposed configuration
- McDATA switches and director LW GBIC count
- The ISL link count to the DWDM equipment at each site
- SM FC cable laying and termination
- McDATA Director high availability features
- E_D_TOV, R_A_TOV, and BB_Credit settings equal on all directors
- Maximum distance for individual devices
- Nickname assignments so we can quickly cross reference WWNs to devices

- Configuration and count of the client interface adapters in the DWDM equipment
- Configuration and count of the transport network adapters in the DWDM, ensuring same frequency band and Channel
- Connectivity testing between
- ESS to switches
- Servers to switch
- Switches to DWDM
- EFC Server and Manager Users and password defined
- Pre-installation testing of the DWDM transport network for configuration and performance

Performance

Typically, for a low performance server, the recommended server to storage connection ratio is 12 to 1, and for a high performance server, the server to storage ratio is 6 to 1. Low performance servers are typically made up of file and print servers whereas high performance servers are application servers.

The DWDM equipment is transparent to the subsystems that are using it. There are no internal queues or busy conditions, so it does not affect performance. The performance factors we need to consider in this solution are the number of ISLs and distance.

To increase the performance of the SAN, multiple connections may be added from the hosts to the directors, from the directors to the storage devices and also between the director and the DWDM equipment.

Scalability

- Most DWDM support multiple I/O interfaces for client equipment and support up to 32 separate channels. DWM equipment also allow for mixing different protocol interfaces of client equipment to a single physical DWDM equipment.
- Scalability rules for the SAN fabric of McDATA switches and Directors remain unchanged as in previous examples.

Security

Dual fabrics can protect you against user errors, such as a user erasing or changing the zoning information inappropriately. The zoning information is separate for each fabric SAN AD and SAN BC, so when changing the zoning in one fabric it does not automatically propagated into the other fabric.

Should you decide to add an ISL, ensure that all checks are done on the zone configuration changes from the management console.

The DWDM equipment provide a Web based management software that can be accessed by any workstation connected to the same LAN. Different user levels are provided for administrators, operators or observers. Different passwords must be set to limit access.

The physical security of fiber connections and patch panels should be considered.

"What if" failure scenarios

Here are some theoretical assumptions:

- **Host HBA failure:** SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. EFC Manager user with Product Administrator rights and ESS Specialist access are required.
- **Storage host adapter failure:** The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce available bandwidth by 25%. For an average workload and five servers as shown it should not impact performance. When the host adapter is replaced we need to update zoning with the new WWN. We also need to reconfigure the OS and SDD to pickup the new path information. EFC Manager user with Product Administrator rights and OS root access are required.
- **Director port failure**: The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. OS and SDD will have to be reconfigure to pickup the new path information if it was an storage port. Physical access to director and EFC Manager user with maintenance rights is required. OS root access may be required.
- **Fiber failure:** Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.
- **EFC Server failure:** No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.
- **EFC Server hard drive failure**: Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.

- **Director completely down, Storage completely down, or site down (power, air conditioning, site damage):** will cause an interruption in normal operation.
- **Physical damage to Storage causing data loss (fire, flood):** We will need to restore data from backup tapes.
- **DWDM optical channel card failure**: As we configured the channels for high availability, there are redundant cards in the DWDM. If we lose one, the traffic will automatically be switched to the other and the channel will remain available. The failed card can be replaced concurrently.
- **DWDM optical channel manager card failure**: The optical channel manager card performs path high availability switching. There are two cards in each shelf, if one fails the other takes control and the operation is not affected. The failed card can be replaced concurrently.
- **DWDM Optical Multiplexer failure or shelf backplane failure:** The entire shelf will be unavailable. As we spread connections in different shelves, we will have at least half the channels of each type available. Operation will continue although performance may be affected.
- **Dark fiber failure:** As we configured for high availability, operation will continue using the available pair with no performance impact. Customers may have additional considerations based on their business recovery policy to source the redundant Dark fiber from another telco/source.

DWDM considerations

We have shown a ring topology here, this gives a logical mesh, potentially giving any to any connectivity. We have implemented this as two rings, one ring connects to one switch in each site, the other ring connects to the remaining switch, this give two discrete SAN fabrics.

We would implement the DWDM solution as a protected ring to ensure availability. In this example we have four multi-mode fibre connections into each DWDM at each site. This can be changed and the number of channels that will be needed at each location is going to be dependent upon the inter-site traffic that is expected. This will be driven by the reasons for the implementation, here we have assumed an light workload. We show here an enterprise deployment that enable data to be made available across a metropolitan area network, or a company campus. It is likely that fibre is expensive here and that the DWDM reduces this overhead. This can be implemented as two discrete SANs, each with one switch in each location. This is an excellent approach for availability and redundancy. It give you a high level of protection and an example of this would be against human error in zoning.

Latency will need to be taken into account, which is also related to buffer credits, however this is not unique to DWDM, more the general SAN solution over distance. The DWDM is a core architecture deployment and because of it's independence from protocol it can be used lots of other traffic, it extends beyond the SAN environment. The distance we have shown here is 25 km between nodes.

13.8 Two sites: Channel extender and WAN extension

Channel extenders typically use telecommunication lines for data transfer and therefore enable application and recovery sites to be located over longer distances apart. The use of channel extenders provides the separation for disaster recovery purposes and avoids some of the barriers imposed when customers do not have a "right of way" to lay fiber cable.

In Figure 13-7 we show a typical SAN distance extension using Optical Channel extenders at both primary and secondary sites and the ATM network.

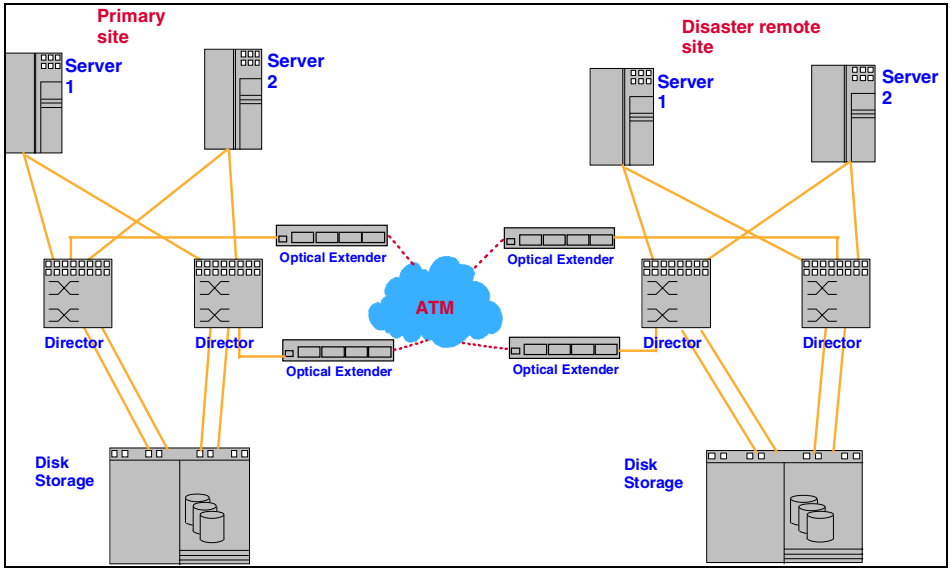


Figure 13-7 SAN extension over (ATM) WAN using channel extenders

The SAN fabric is comprised of:

- IBM 2031- 016 McDATA FC switch
- IBM 2031 - 032 McDATA FC switch

- IBM 2032 - 064 McDATA Enterprise FC director (We have used this in the example)

The host servers and storage are connected to the SAN fabric via SW GBIC's using MM Fibre cable.

We assumed some considerations as follows:

- The data movement/copy from primary to secondary site is done at the OS level in the host or by storage subsystem if it supports it.
- The ATM network is installed, configured and available. The ATM network technical details are beyond the scope of this example.
- As with any extended solution "distance = some performance sacrifice". Any individual transaction will be effected by the latency. Latency typically needs to be validated based on the specific end-to-end devices deployed. An extended distance configuration that worked with one Server/Application might fail with time-outs with another.

Checklist

We checked the following items:

- Host operating system, dual pathing software (IBM SDD) and adapter firmware levels checked for compatibility with proposed configuration
- Storage capacity and LUN assignments to each server
- Disk storage features and microcode level for proposed configuration
- McDATA switches and director SW GBIC count
- MM FC cable laying and termination
- McDATA Director high availability features and firmware levels
- Nickname assignments so we can quickly cross reference WWNs to devices
- EFC Server and Manager Users and password defined
- I/O interface (for example, FC or ESCON) selection and installation on the Optical channel extender
- Network interface (for example, ATM OC3) selection and installation on the Optical channel extender
- Pre-installation testing of the network interfaces (ATM network)

Performance

The major performance consideration with a long distance solution is calculating the correct number of interconnecting links between sites. This number can only be determined by performing a detailed performance profile of the servers, storage that will be remote. It is vital that detailed performance data is available prior to sizing the number of interconnecting links required. Channel extender generally compress the data before sending it over the transport network, however the compression ratio needs to be determined based on the application characteristics and the distance.

You must especially consider an amount of the updated data for a period of time (peak time), and reflect this on calculating the number of interconnecting links and the number of data volumes for SAN/WAN solution.

Scalability

Most optical channel extenders support multiple I/O interfaces for client equipment. However in a installation, it is recommended to have a single protocol interfaces on a physical channel extender.

Scalability rules for the SAN fabric of McDATA switches and directors remain unchanged as in previous examples.

Security

These are some security considerations we took into account:

- Disk Storage LUN masking by WWN will allow each server access only to configured LUNs
- EFC Manager user IDs, passwords and rights are defined and defaults are removed, so only authorized personnel can perform management functions
- Remote access to EFC Manager configured to limit access to authorized workstations
- Physical director security — locked cabinet, restricted access site
- The physical security of fiber connections and patch panels should be considered

“What if” failure scenarios

Here are some theoretical assumptions:

- **Host HBA failure:** SDD will move all load to remaining path. Available bandwidth to the specific server will be reduced to 50%. When the HBA is replaced the zoning information and ESS host definition will have to be updated with the new WWN. EFC Manager user with Product Administrator rights and ESS Specialist access are required.

- **Storage host adapter failure:** The available paths to storage will be reduced, impacting the server to storage ratio, and performance of all servers sharing that path may be affected. In this example if we installed four host adapters a single adapter failure will reduce available bandwidth by 25%. For an average workload and five servers as shown it should not impact performance. When the host adapter is replaced we need to update zoning with the new WWN. We also need to reconfigure the OS and SDD to pickup the new path information. EFC Manager user with Product Administrator rights and OS root access are required.
- **Director port failure**: The impact will depend on whether it is a server or storage port or the ISL. It will be similar to Host HBA or Storage Host adapter failure. The cable can be moved to a spare port. The OS and SDD will have to be reconfigured to pickup the new path information if it was a storage port. Physical access to director and EFC Manager user with maintenance rights is required. OS root access may be required.
- **Fiber failure**: Impact will depend on whether it is a host attachment or storage attachment fiber. The only action required is fiber replacement. Physical access to director and attached device are required.
- **EFC Server failure**: No management access unless we are using inbound management. Operation is not affected until we need to alter zoning information, for example.
- **EFC Server hard drive failure:** Operation with current zoning definition is not affected. Configuration and zone definition information can be restored from zip drive backup.
- **Director completely down, Storage completely down, or site down** (power, air conditioning, site damage): Will cause an interruption in normal operation.
- **Physical damage to Storage causing data loss (fire, flood)**: We will need to restore data from backup tapes.
- **Channel extenders failure:** In the current solution channel extenders are not configured in a redundant mode, however they can be.
- **Transport network failure:** The telco link failure is not planned for in the current solution however a redundant link can be acquired from the vendor or a separate telco may be considered depending on the business recovery policy.

13.9 Remote tape vaulting

For an existing large corporation with multiple sites, tape library resources can be consolidated to a separate site. This simplifies data movement logistics and centralizes backup software configurations. By doing this the cost of doing business is reduced as the infrastructure is efficiently utilized and less IT personnel required for backup data management. In addition, in the event of a disaster, the data is already located on tape in remote location and there is no longer a need to ship the data to another site. Another application is for outsourcing services companies like Storage service providers (SSP), that are interested in providing backup and Backup management services to smaller companies who may not have the infrastructure or need it based on their business recovery policy.

The Figure 13-8, shows the basic layout for a remote tape vaulting solution, using McDATA directors.

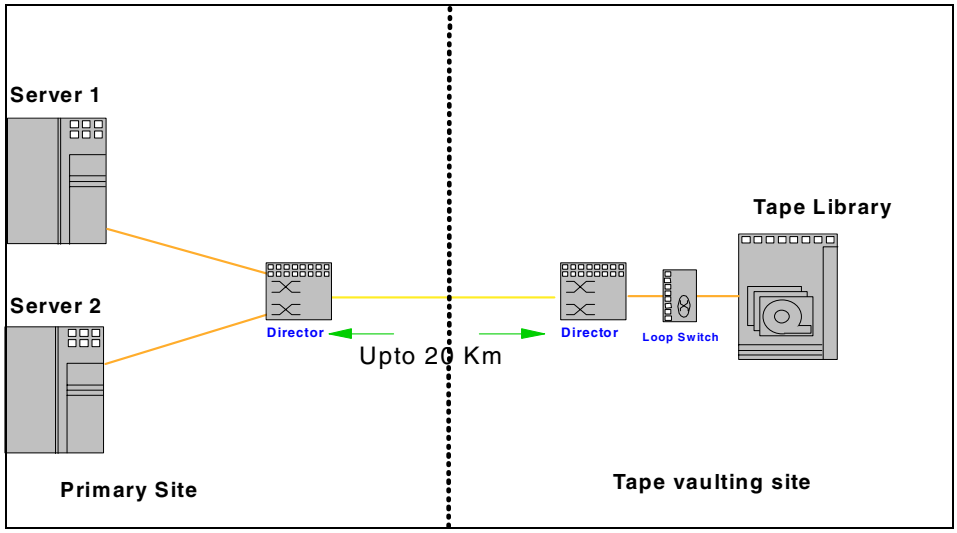


Figure 13-8 Remote tape vaulting

This solution extends the distance to up to 20 km apart, using Long wave GBICs in the SAN fabric, and Single mode Fiber cable interconnect. Distance up to 80 km are possible with Finisar repeaters.

Note: Two 2032s connected via longwave GBICs can be placed 20 km apart.

There will be no change to the longwave GBIC distances when the 2Gb SAN technology is available.

Architecturally, the fibre-channel connection extensions could be provided by “dark fibre”, Fibre Channel repeaters, or dense wave division multiplexing (DWDM) devices. Dark fibre is fibre provided by one of the telcos with the telco providing the repeaters.

2032 - IBM has not done any testing of the 2032 in a MAN configuration. McDATA, the 2032's manufacturer, has tested the 2032 with the Finisar repeaters at 100 km.

The SAN fabric is comprised of:

- IBM 2031- L00 McDATA Loop switch
- IBM 2031- 016 McDATA FC switch
- IBM 2031 - 032 McDATA FC switch
- IBM 2032 - 064 McDATA Enterprise FC director (We have used this in the example)

The host servers are connected to the SAN fabric via SW GBIC’s using MM Fibre cable. In case of McDATA switches 2032 and 2032, the distance is limited to 10 km.

Since most tape libraries fiber channel ports run in Arbitrated Loop, they cannot be directly attached to the McDATA Director, so we have attached them to McDATA ES-1000 Switches and connected the ES-1000 B_Ports to the McDATA Director.

The ES-1000s are used only for tape consolidation; there are no other devices attached to them.

Zoning can be established by WWN so we can limit access of each server or each group of servers to specific tape drives.

In order to avoid human errors that can affect operation of other servers, zone changes should only be performed by designated personnel, and proper procedures must be in place to make sure operation personnel are aware of the available devices to each server according to the zones currently active.

Checklist

In addition to the items considered in the disk consolidation example, we must now consider:

- Host HBA supported for tape drive / library attachment
- Host tape device driver levels
- Host operating system levels compatible with tape library fiber requirements
- ES-1000s unique switch IDs
- ES-1000s priority values higher than director
- ES-1000s E_D_TOV, R_A_TOV, and BB_Credit compatible with Director
- Zone configuration allowing tape access to required servers
- Host software tape sharing capabilities
- Switches LAN connection to EFC Server and remote workstations

Performance

All drives connected to a single ES-1000 switch will share the 100 MB/s bandwidth of the single ISL connection.

Depending upon the drive interface LTO or Magstar, the performance may vary.

In order to be able to use all drives simultaneously without impacting their performance, we will connect only two drives to each ES-1000 switch. We may add additional ES-1000 switch, as there is a possibility of attaching up to four drives to the two switches.

Note: The dual ES1000 can also be configured for switch redundancy and path redundancy only incase of AIX and Magstar 3590 FC drives.

Performance will basically depend on the number of HBAs available on each server and the number of storage connections and most important the distance between storage and servers. It is important to consider application behavior over the distance, and this varies from application to application and we have assumed that the performance is adequate for applications running in the servers.

Note: The 2032, 2042, and 2109 with the Extended Fabric feature all support Buffer to Buffer credits. Buffer to Buffer credits allows commands to be queued up in the buffer of the switch. This mechanism lessens the effect of the latency and improves the aggregate performance.

Any individual transaction will be effected by the latency. Latency typically needs to be validated based on the specific end-to-end devices deployed. An extended distance configuration that worked with one Server/Application might fail with time-outs with another.

IBM Global Services offers a wide array of SAN services to help your customer. A description of IGS's Fibre Transport Services can be found at:

http://www-1.ibm.com/services/its/us/drmkbb04.html

Scalability

Potentially we can scale the solution in terms of adding additional tape drives by also increasing the number of additional ES-1000 switches. However physical space and cost should be considered, also a loop switch experiencing a loop initialization (LIP), could potentially disable access to the tape drives attached to it and also may in some cases require a reset on the tape library.

Security

Here are some security considerations:

- Zoning can be used to restrict access to devices to specific servers when required.
- Proper tape management procedures will avoid servers contending for the same tape device.
- EFC manager and ES-1000 Web access users and passwords configured and defaults removed.

“What if” failure scenarios

Here are some theoretical assumptions:

- Performance may be impacted depending on number of drives attached. Traditionally tape failover is a manual operation. Multiple path devices are configured as several logical devices, one per path. Only one of these logical devices is made active. If there is a failure the application aborts and it can then be restarted using a different logical device. For tape drives with dual paths, the latest levels of a tape device driver provide alternate pathing support and tape failover for Fibre Channel connections. With this support enabled if an error occurs the device driver will automatically initiate error recovery and the operation will continue using the next logical path.

- Device link or device port failure — only one tape drive affected. Recovery may be manual or automatic depending on operating system and driver level as explained for ISL or switch failure.
- Switch port failures — GBICs are hot swappable. H_Ports can be moved to a spare port. B_Port cannot be moved, and if it fails and it is not the GBIC, the switch has to be replaced.
- User error trying to access more than two drives on the same link — performance of all drives attached to the switch pair may be degraded. Director or switches performance view may be used to find what paths are carrying traffic.
- Tape drive failure in a single tape zone — an alternate zone should be made active to get access to a working device.

13.9.1 Remote tape vaulting with disaster tolerance

In Figure 13-9 we show an extension or variant of the tape vaulting solution. In this solution the primary site / local site has a tape library. Data is written to both tape libraries, however in the event of failure on the tape library at the local site, data can be backed up and restored from the tape library at the remote site.

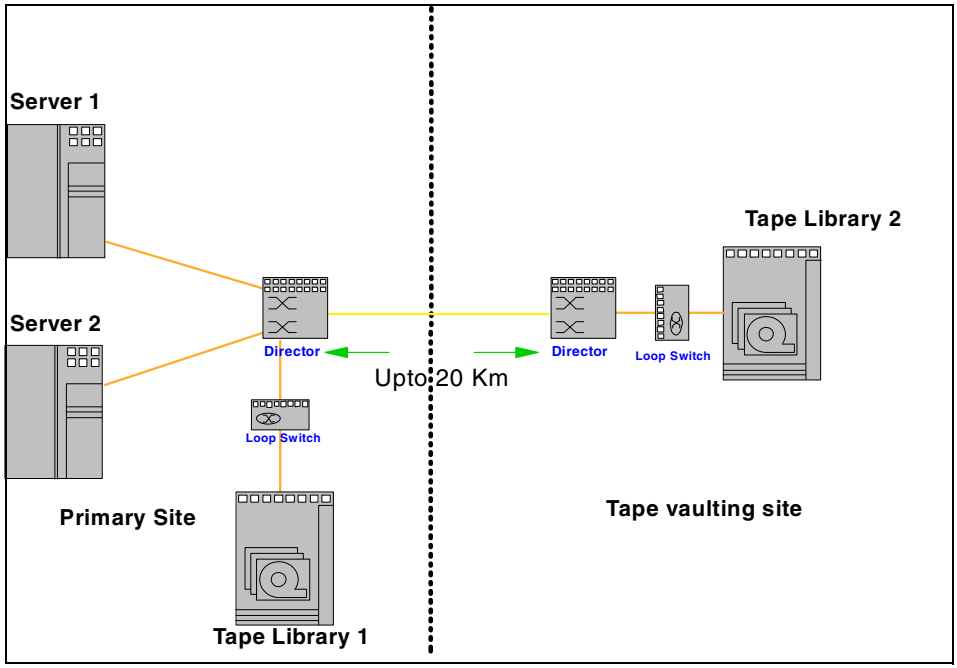


Figure 13-9 Remote tape vaulting with disaster tolerance

We show the basic layout for a remote tape vaulting with disaster tolerance solution, using McDATA directors. This solution extends the distance to up to 20 km apart, using Long wave GBICs in the SAN fabric, and Single mode Fiber cable interconnect.

Distances up to 80 km are possible with Finisar repeaters.

Note: Two 2032s connected via longwave GBICs can be placed 20km apart.

There will be no change to the longwave GBIC distances when the 2Gb SAN technology is available.

Architecturally, the Fibre Channel connection extensions could be provided by “dark fibre”, fibre-channel repeaters, or dense wave division multiplexing (DWDM) devices. Dark fibre is fibre provided by one of the telcos with the telco providing the repeaters.

2032 - IBM has not done any testing of the 2032 in a MAN configuration. McDATA, the 2032's manufacturer, has tested the 2032 with the Finisar repeaters at 100 km

The SAN fabric is comprised of:

- IBM 2031- L00 McDATA Loop switch
- IBM 2031- 016 McDATA FC switch
- IBM 2031 - 032 McDATA FC switch
- IBM 2032 - 064 McDATA Enterprise FC director (We have used this in the example)

The host servers are connected to the SAN fabric via SW GBIC’s using MM Fibre cable. In the case of the McDATA switches 2032 and 2032, the distance is limited to 10 km.

Since most tape libraries fiber channel ports run in Arbitrated Loop, they cannot be directly attached to the McDATA Director, so we have attached them to McDATA ES-1000 Switches and connected the ES-1000 B_Ports to the McDATA Director.

The ES-1000s are used only for tape consolidation, there are no other devices attached to them.

Zoning can be established by WWN so we can limit access of each server or each group of servers to specific tape drives.

In order to avoid human errors that can affect operation of other servers, zone changes should only be performed by designated personnel, and proper procedures must be in place to make sure operation personnel are aware of the available devices to each server according to the zones currently active.

Checklist

In addition to the items considered in the disk consolidation example, we must now consider:

- Host HBA supported for tape drive / library attachment
- Host tape device driver levels
- Host operating system levels compatible with tape library fiber requirements
- ES-1000s unique switch IDs
- ES-1000s priority values higher than director
- ES-1000s E_D_TOV, R_A_TOV, and BB_Credit compatible with Director
- Zone configuration allowing tape access to required servers
- Host software tape sharing capabilities
- Switches LAN connection to EFC Server and remote workstations

Performance

All drives connected to a single ES-1000 switch will share the 100 MB/s bandwidth of the single ISL connection.

Depending upon the drive interface LTO or Magstar, the performance may vary.

In order to be able to use all drives simultaneously without impacting their performance we will connect only two drives to each ES-1000 switch. We may add an additional ES-1000 switch as there is a possibility of attaching up to four drives to the two switches.

Note: The dual ES1000 can also be configured for switch redundancy and path redundancy only incase of AIX and Magstar 3590 FC drives.

Performance will basically depend on the number of HBAs available on each server and the number of storage connections and most important the distance between storage and Servers. Its important to consider application behavior over the distance, and this varies from application to application and we have assumed that the performance is adequate for applications running in the servers.

Note: The 2032, 2042, and 2109 with the Extended Fabric feature all support Buffer to Buffer credits. Buffer to Buffer credits allows commands to be queued up in the buffer of the switch. This mechanism lessens the effect of the latency and improves the aggregate performance.

Any individual transaction will be effected by the latency. Latency typically needs to be validated based on the specific end-to-end devices deployed. An extended distance configuration that worked with one Server/Application might fail with time-outs with another.

IBM Global Services offers a wide array of SAN services to help your customer. A description of IGS's Fibre Transport Services can be found at:

http://www-1.ibm.com/services/its/us/drmkbb04.html

Scalability

Potentially we can scale the solution in terms of adding additional tape drives by also increasing the number of additional ES-1000 switches. However physical space and cost should be considered, also a loop switch experiencing a LIP (Loop initilization), could potential disable access to the tape drives attached to it and also may in some cases require a reset on the tape library.

Security

- Zoning can be used to restrict access to devices to specific servers when required.
- Proper tape management procedures to avoid servers contending for the same tape device.
- EFC manager and ES-1000 Web access users and passwords configured and defaults removed.

“What if” failure scenarios

Here are some theoretical assumptions:

- Performance may be impacted depending on number of drives attached. Traditionally tape failover is a manual operation. Multiple path devices are configured as several logical devices, one per path. Only one of these logical devices is made active. If there is a failure the application aborts and it can then be restarted using a different logical device. for tape drives with dual paths latest levels of a tape device driver provide alternate path support and tape failover for Fibre Channel connections. With this support enabled if an error occurs the device driver will automatically initiate error recovery and the operation will continue using the next logical path.

- Device link or device port failure — only one tape drive affected. Recovery may be manual or automatic depending on operating system and driver level as explained for ISL or switch failure.
- Switch port failures — GBICs are hot swappable. H_Ports can be moved to a spare port. B_Port cannot be moved, and if it fails and it is not the GBIC, the switch has to be replaced.
- User error trying to access more than two drives on the same link — performance of all drives attached to the switch pair may be degraded. Director or switches performance view may be used to find what paths are carrying traffic.
- Tape drive failure in a single tape zone — an alternate zone should be made active to get access to a working device.

A

SAN distance solutions using IP

In this chapter we look at some of the new upcoming options to extending the SAN storage over IP networks. Some of the options enable customers to use their existing IP infrastructure and enable the SAN FC protocols over these via gateway routers or via IP tunneling.

We look at some of the offerings that are supported by IBM Global Services.

Storage over TCP/IP

The storage over TCP/IP technology takes advantage of the already existing networking skills available in the customer environment and the low cost of the TCP/IP based networking implementation. Storage can then use the TCP/IP network to access storage the same way it can be accessed through a Fiber Channel based SAN environment today.

Thus, storage over TCP/IP is very important for environments where you have a need to access storage from locations away from the data center, where customers usually have the SANs. For those locations, it is easier to use the TCP/IP infrastructure instead of replicating the Fibre-Channel based network in parallel to the Ethernet network.

In storage over TCP/IP there are some different solutions that implement different topologies. The next topics will show each one of these new solutions.

Additionally, the Internet Engineering Task Force is considering the following protocols as standards in the storage networking area:

- **iSCSI**

  Native SCSI commands over the TCP/IP network

- **iFCP**

  Storage networking gateway enabling coexistence and interoperability of storage and server devices in a multi-protocol environment including Fibre Channel, SCSI and iSCSI

- **FCIP**

  Fibre channel E_Port tunneling over the TCP/IP network

It is not a question of “which protocol is best?” any longer — but more a question of “which other protocols will work with mine”?

Storage with native TCP/IP interface

This solution has the same characteristics of a SAN and the only difference is that the Fibre Channel network is replaced by a TCP/IP network.

In Figure A-1, we show the topology of the solution:

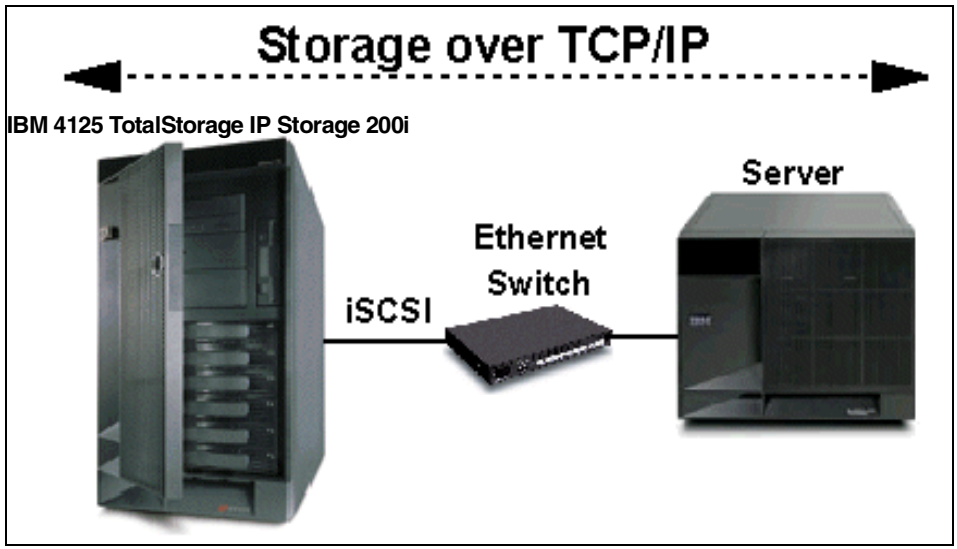


Figure A-1 IBM 4125 TotalStorage IP Storage 200i

This solution enables customers to build SANs without the costly Fibre Channel infrastructure.

IBM announced in February 2001, the IBM 4125 TotalStorage IP Storage 200i models 100 and 200 which are connected directly to the TCP/IP network.

The protocol supported by the IBM 4125 TotalStorage IP Storage 200i is iSCSI, a protocol for storage over TCP/IP jointly developed by IBM and Cisco. This protocol is being considered by the Internet Engineering Task Force (IETF) to be one of the standards in the storage over TCP/IP arena.

iSCSI is the existing SCSI protocol being intercepted by an agent within the host that then packages the SCSI I/O in an IP format and it is then transported over IP.

This process is reversed at the receiving device and the package is unwrapped and the SCSI I/O sent to the native storage device. This method is attractive to companies that already have deployed IP backbone infrastructure although this is often already over subscribed.

The transport of these packets can be enhanced by larger frame sizes and by implementing discrete routes, for example, by using Virtual Private Networks (VPNs) which will reduce the unpredictable nature of the performance. There is obviously an overhead in the packaging and unpackaging process and this may prove inefficient.

When scoping an iSCSI implementation it is sometimes found that the organization will have to deploy a much larger IP infrastructure to cope with the expected iSCSI traffic,

In this instance is it often a more considered approach to deploy a dedicated infrastructure, (maybe a Storage Area Network), with known dedicated bandwidth. This bring many other benefits with it, such as security and so on.

While organizations may choose to deploy a core SAN strategy this does leave a place for the iSCSI architecture, by utilizing the IBM 200i it is possible to enable iSCSI devices to communicate with the mission critical Fiber Channel Storage Area Network. This preserves the investment in SAN and may increase its penetration within the enterprise.

iSCSI has a large potential for customers due to savings in cost. One initial area is the cost of the equipment itself where Fibre Channel based components are usually more expensive initially than the correspondent Gigabit Ethernet ones. The higher difference in costs are actually related to the additional management and support infrastructure that the customer will need to built for the Fibre Channel protocol based products.

This cost is justifiable if you are in a data center, but for enterprise customers with several branches or large campuses it may not be desirable to replicate this expense in all points of the network.

Technically, the iSCSI solution has also a high potentiality. Although today some overhead is included through the use of a TCP stack over the TCP/IP network, there are several works under way to overcome that limitation:

- One of the bottlenecks are the server CPUs which have to process not only the application itself but also the TCP/IP and the iSCSI protocols. Network interface adapters are being developed by several companies with an off-loaded TCP/IP and iSCSI protocols leaving the server with its actual function of running applications.
- Usage of the network can be optimized also with the use of jumbo frames allowing you to send block I/O access to storage on one only network packet of data. This procedure allows you to send full blocks without breaking them to send through the network.
- The rate of throughput increase is also changing with Fibre Channel now making available 2 Gbps Fibre Channel switches and Ethernet now making available 10 Gbps switches.

All these improvements justify the use of iSCSI solutions even if you need to built a separate subnet for storage or a higher priority VPN for storage, once increasing the bandwidth of the TCP/IP network is much less costly than building a parallel Fibre Channel network.

SAN to iSCSI gateway

SANs are already a widely available technology and products that integrate the SAN environment to the storage over TCP/IP environment are a needed solution for many customers.

The topology of this solution is shown in the Figure A-2 below:

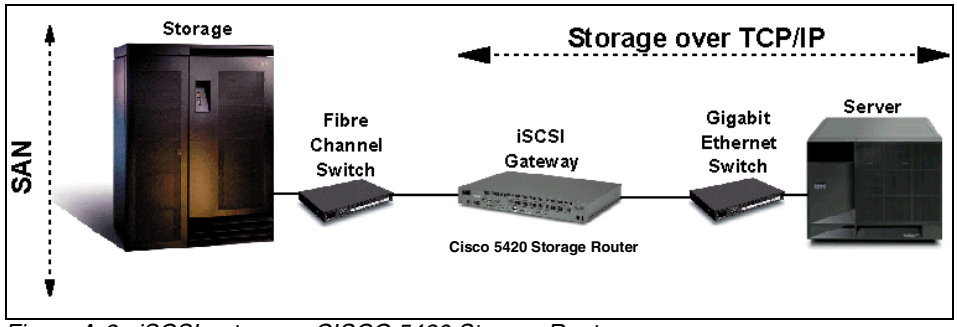


Figure A-2 iSCSI gateway - CISCO 5420 Storage Router

Cisco announced last March, the Cisco SN-5420 Storage Router, enabling servers on the TCP/IP environment to access storage devices on the SAN through the conversion of protocols from Fibre Channel to Gigabit Ethernet and TCP/IP.

The Cisco SN-5420 Storage Router implements the same iSCSI protocol implemented in the IBM products for storage over TCP/IP.

Storage to IP gateway

A different approach to enable storage on the TCP/IP environment is the tunneling function which extends the reach of a central SAN.

The topology of this solutions is shown in the following Figure A-3:

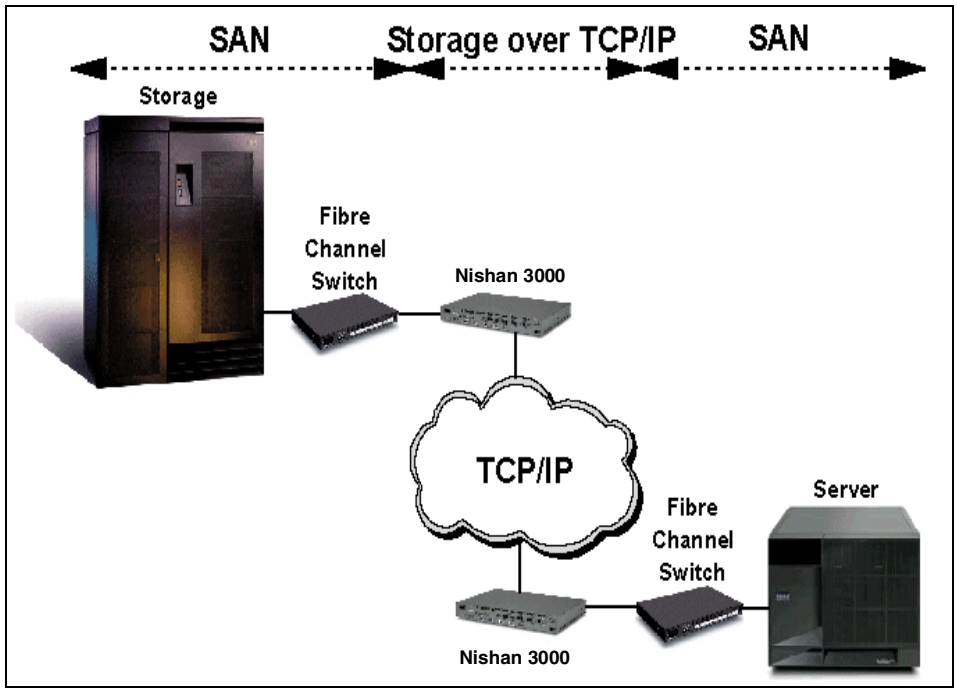


Figure A-3 Storage over TCP/IP Tunneling - Nishan 3000

These solutions are implemented by products like the Nishan 3000. In fact, the Nishan 3000 is a hybrid Fibre Channel/Gigabit Ethernet Switch enabling these products to act as the Fibre-Channel Switch and the Tunnel at the same time. It has up to eight ports that can be configured to be Fibre-Channel or Gigabit Ethernet.

Nishan offers the same solution to extend the reach of SCSI interfaces. The Nishan 2000 has four SCSI interfaces and two Gigabit Ethernet interfaces. Nishan 2000 and Nishan 3000 can work simultaneously in the same network enabling a mix of environments.

Both products from Nishan implements a protocol called mFCP. It works on UDP which means they have better performance but they are limited to the Local Area Network (LAN).

Nishan has just announced the Nishan 1000 which converts mFCP to iFCP which then would allow customer topologies that go through a Wide Area Network (WAN).

Nishan has also just announced the Nishan 3300 that consists of a two port Gigabit Ethernet adapter for the Nishan 3000. These two ports are capable of supporting iSCSI and converting them, through the gateway function, to enable access to storage or servers in the fibre-channel environment.

Storage over TCP/IP tunneling

CNT's UltraNet Edge Storage Router allows you to leverage the Fibre Channel SAN fabrics you already have in place, as well as your existing IP network infrastructure to deliver a high-performance storage networking environment. The Edge Storage Router provides an FCIP-based SAN interconnection capability over unlimited distances. FCIP, currently before the IETF standards body, encapsulates FC frames in an IP packet for delivery over an IP network. The Edge Storage Router utilizes FC connections on the storage/SAN side, and then extends them over an Ethernet/IP network, which is likely already in place. CNT provides compressed 10/100 Ethernet connectivity today, with Gigabit Ethernet to follow shortly. CNT offers a 1x1 or a 2x2 chassis, meaning 1 or 2 FC storage connections in, and 1 or 2 network connections out.

CNT's FCIP implementation is standards-based, so it can work with the industry-standard IP routers already in place in your network. However, CNT has also included significant value-add for critical storage applications:

- A store-and-forward architecture delivers optimal performance over any distance
- FC and IP payload matching for maximum efficiency
- Incremental buffer credit management for effective, high performance flow control
- Sophisticated recovery management, including load balancing and packet reordering and retransmission
- Patent-pending network-level hardware CRC, for full data integrity

The Edge Storage Router is Brocade Fabric Aware, and utilizes a Fibre Channel E_Port (switch-to-switch) connection, which delivers a single, extended SAN fabric spanning across multiple physical sites. This platform is user-installable, -configurable and maintainable, to ensure that the storage network is up and running with the least possible effort.

To this end, CNT provides a browser-based monitoring application called UltraNet Webview, and a simple, point and click configuration application called UltraNet ConfigManager.

The Edge Storage Router supports all the key storage networking applications you require, and extends them throughout your enterprise. For business continuity purposes, CNT supports remote disk mirroring or remote tape backup. You can also interconnect remote FC SAN islands, so they can be managed from a central location, or so that you can efficiently access or distribute information throughout your organization.

Through these applications, you can improve your resilience to disasters, increase the timeliness of the decision making throughout your enterprise, and even deploy new revenue-generating applications.

In Figure A-4 we show the Edge Storage Router deployed.

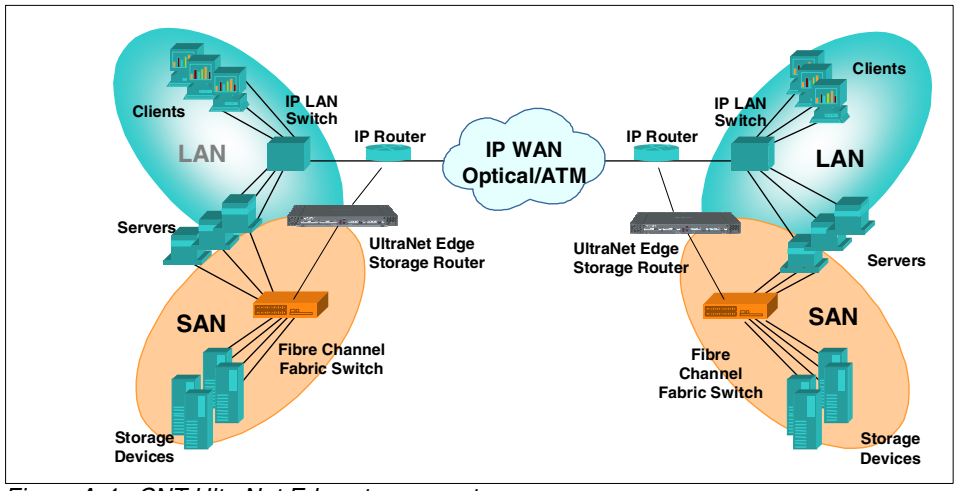


Figure A-4 CNT UltraNet Edge storage router

Finally, CNT is also adding ATM WAN connections to the Edge, for when the customer has in place, or prefers to use a dedicated telco link. Because the Edge Storage Router is a flexible, software-based architecture, it will also support other important storage networking standards, like iSCSI and InfiniBand, as they mature.

B

Finisar optical link extenders

Finisar OptiLinx-2000 is an Optical Link extender. The OptiLinx-2000 is available in two models:

- OptiLinx-2000 FC - For Fibre Channel interface
- OptiLinx-2000 GE - For Gigabit Ethernet interface

In this topic we will overview the OptiLinx-2000 FC, which is used for SAN applications.

OptiLinx-2000 FC

The OptiLinx-2000 FC allows you to extend the distance of FC (Fiber Channel) links beyond the standard distance of 10 km up to 120 km.

A pair of OptiLinx-2000 FCs are used at each end to convert the multi-mode FC interface to single mode FC interface.

We show a picture of the OptiLinx-2000 FC in Figure 13-10.

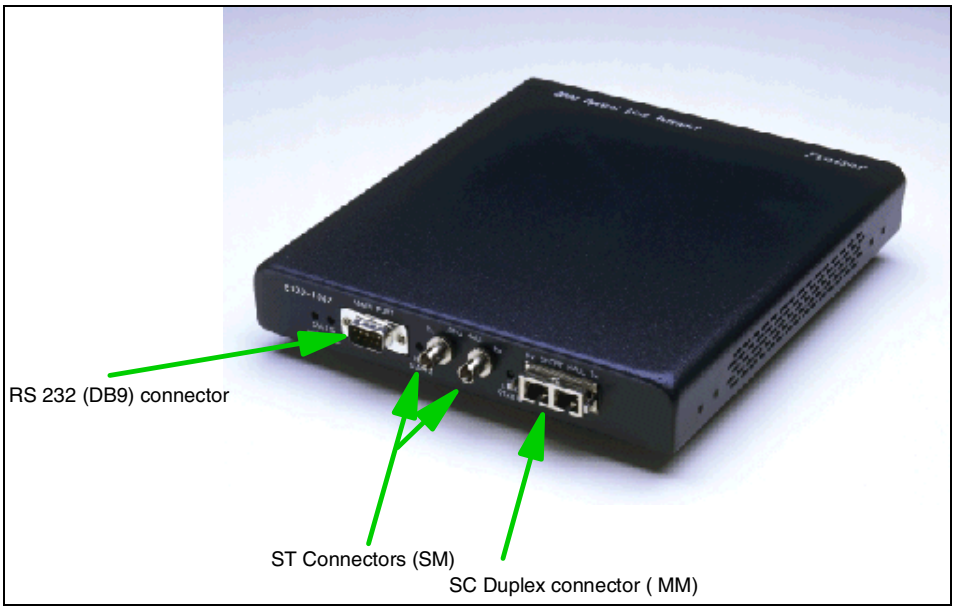


Figure 13-10 OptiLinx-2000 FC

The OptiLinx-2000 FC offers the following features:

- Long haul distances up to 120 km
- Quick and easy installation
- Remote network management
- Built-in diagnostics
- Low jitter
- Low bit error rate
- Redundant power supply
- Rack mounting kit chassis option (with up to 6 OptiLinx-2000 FC)

The OptiLinx-2000 extends the distance of high-speed fiber links well beyond the standard limits. Finisar's Optical Link Extenders facilitate new applications such as remote storage, disaster recovery, extended campus LANs and Gigabit-rate data services.

A pair of OptiLinx-2000s extend a link by converting the short-haul copper or multi-mode optical signal to a long-haul, single-mode signal, and vice versa. Its internal Digi-tal Signal Conditioner re-times the signal over the long-haul connection to eliminate jitter. Data integrity is pre-served over long distances with a bit error rate of 10 or better.

Simple to install, OptiLinx-2000 link extenders include comprehensive system status monitoring and link diagnostics (BERT—bit error rate test). The link monitoring and diagnostics provide quick fault detection and isolation. Integration with network management is provided via an RS232 serial port attached to a local terminal or modem to allow remote access.

A typical installation will involve a minimum of two OptiLinx-2000 FC's, one at each of the sites. We show an example of this in Figure 13-11.

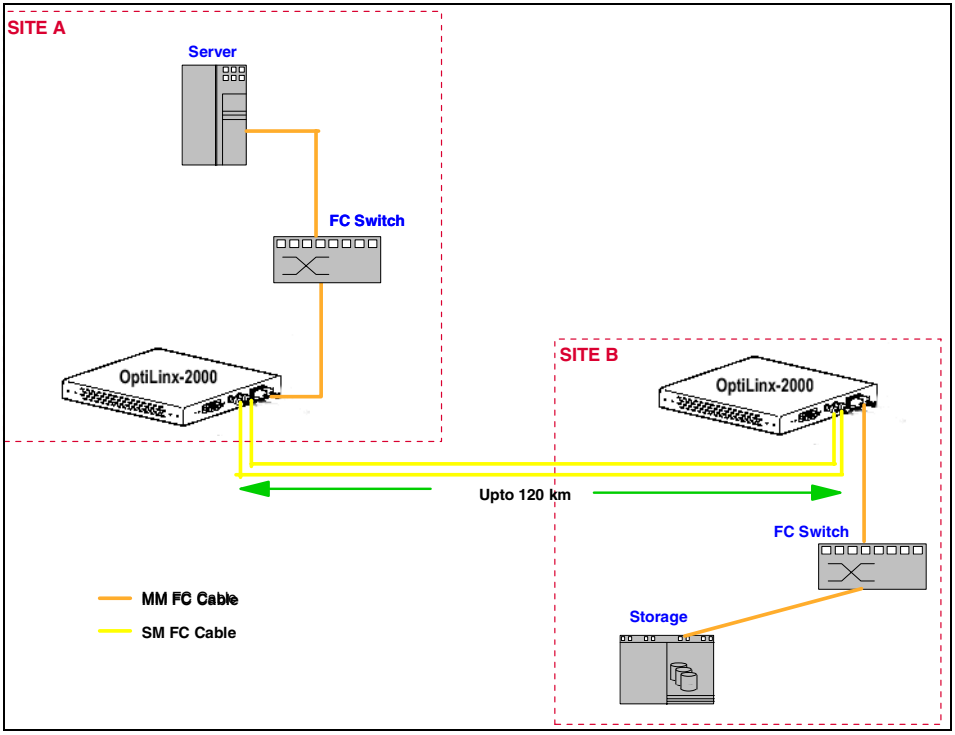


Figure 13-11 OptiLinx-2000 typical installation

Typical uses of the OptiLinx-2000 FC can include remote disk and tape vaulting solutions.

Note: Finisar is likely to replace or phase out the OptiLinx-2000 product and replace it with a GBIC size transceiver, model FTR-1619-XX (where XX denotes the wavelength of the laser). It uses a 1550 nm DFB laser and an Avalanche Photo Diode (APD).

For additional information refer to:

http://www.finisar.com

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 450.

- *IBM SAN Survival Guide*, SG24-6143
- *IBM SAN Survival Guide Featuring the IBM 2109*, SG24-6127
- *IBM SAN Survival Guide Featuring the INRANGE Portfolio*, SG24-6150
- *IBM SAN Survival Guide Featuring the McDATA Portfolio*, SG24-6149
- *Designing an IBM Storage Area Network*, SG24-5758
- *Implementing an Open IBM SAN,* SG24-6116
- *Introduction to Storage Area Network, SAN*, SG24-5470
- *IBM Storage Solutions for Server Consolidation*, SG24-5355
- *Implementing the Enterprise Storage Server in Your Environment*, SG24-5420
- *Storage Area Networks: Tape Future In Fabrics*, SG24-5474
- *IBM Enterprise Storage Server*, SG24-5465
- *Implementing ESS Copy Services on S/390*, SG24-5680
- *Implementing ESS Copy Services on UNIX and Windows NT/2000*, SG24-5757
- *The IBM LTO Ultrium Tape Libraries Guide*, SG24-5946

Other resources

These publications are also relevant as further information sources:

- *Building Storage Networks*, by Marc Farley. McGraw-Hill Professional Publishing, 2nd edition, May 2001. ISBN: 0072130725

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 2109 Model S08 User's Guide*, SC26-7349
- *IBM 2109 Model S08 Switch Service Guide*, SC26-7350
- *IBM 2109 S16 Switch User's Guide*, SC26-7351
- *IBM 2109 S16 Switch Service Guide*, SC26-7352
- *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 SAN Fibre Channel Managed Hub 3534 User's Guide,* GC26-7391
- *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
- *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

These Cisco publications are also relevant as further information sources:

- *Cisco ONS 15540 ESP Configuration Guide and Command Reference*, 78-12669-01
- *Cisco ONS 15540 ESP Planning and Design Guide*, 78-13102-01
- *Cisco ONS 15540 ESP Troubleshooting Guide*, 78-13022-01
- *Cisco ONS 15540 ESP Hardware Installation Guide*, 78-12591-01

These CNT publications are also relevant as further information sources:

- *UltraNet Wave Multiplexer/Optimizer User Guide*, 30200512/03
- *UltraNet Wave Optimizer User Guide*, 30200513/01
- *UltraNet Wave Multiplexer User Guide*, 30200512/03

These INRANGE publications are also relevant as further information sources:

- *IN-VSN FC/9000 Fibre Channel Director Installation Manual*, 9110461-102
- *FC/9000 Fibre Channel Director Site Planning Guide*, 9110460-101
- *FC/9000 Fibre Channel Director Maintenance Manual*, 9110774-307
- *IN-VSN Enterprise Manager (IN-VSN EM) Software Installation and Operation Guide*, 9110509-203

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 Nortel publications are also relevant as further information sources:

- *Technical Specifications: Rel.3.2 (08/01)*, 323-1701-180
- *System Description: Rel.3.2 (08/01)*, 323-1701-100
- *Building a Network: Rel.3.2 (08/01)*, 323-1701-280
- *Installing Shelves and Components: Rel.3.2 (08/01)*, 323-1701-201

- *Hardware Description:Rel.3.2 (08/01)*, 323-1701-102
- *Software and User Interface:Rel.3.2 (08/01)*, 323-1701-101
- *Planning Guide: Rel 3.2 (06/01)*, NTY410AB

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 Sorrento Network publications are also relevant as further information sources:

- *GigaMux Planning Guide*, 42-02011-01
- *GigaMux Installation Guide*, 42-02022-01
- *GigaMux Operations Guide*, 42-02033-01

Referenced Web sites

These Web sites are also relevant as further information sources:

- http://www.itu.int
  International Telecommunication Union Web site
- http://www.ibm.com/storage/ess
  IBM TotalStorage Enterprise Storage Server Web site
- http://www.storage.ibm.com/hardsoft/tape/index.html
  IBM tape and optical storage Web site
- http://www.ibm.com/storage/lto
  Linear Tape-Open Web site
- http://www.storage.ibm.com/hardsoft/tape/3590/3590opn.html
  IBM TotalStorage™ Enterprise Tape Drive 3590 — List of Supported Servers
- http://www-1.ibm.com/services/its/us/drmkbb04.html
  Structured Cabling Web site
- http://www.storage.ibm.com/ibmsan/index.html
  IBM Enterprise SAN
- http://www.storage.ibm.com/hardsoft/products/fchub/fchub.htm
  IBM Fibre Channel Storage HUB

- http://www.pc.ibm.com/ww/netfinity/san
  IBM Storage Area Networks: Nefinity Servers
- http://www.storage.ibm.com/hardsoft/products/fcswitch/fcswitch.htm
  IBM SAN Fibre Channel Switch
- http://www.storage.ibm.com/hardsoft/products/sangateway/supserver.htm
  IBM SAN Data Gateway
- http://www.storage.ibm.com/hardsoft/products/tape/ro3superserver.htm
  IBM SAN Data Gateway Router
- http://www.storage.ibm.com/hardsoft/products/fcss/fcss.htm
  IBM Fibre Channel RAID Storage Server
- http://www.storage.ibm.com/hardsoft/products/ess/ess.htm
  Enterprise Storage Server
- http://www.brocade.com
  Brocade Communications Systems, Inc.
- http://www.cisco.com
  Cisco Systems
- http://www.cdp.com
  Columbia Data Procucts, Inc.
- http://www.cnt.com
  Computer Network Technology
- http://www.emulex.com
  Emulex Corporation
- http://www.fibrechannel.com
  Fibre Channel Industry Association
- http://www.finisar.com
  Finisar Corporation
- http://www.jni.com
  JNI Corporation
- http://www.inrange.com
  INRANGE Technologies Corporation
- http://www.mcdata.com
  McDATA Corporation
- http://www.nortel.com
  Nortel Networks
- http://www.pathlight.com
  Pathlight

- http://www.qlogic.com
  QLogic Corporation
- http://www.sanergy.com
  Tivoli SANergy
- http://www.snia.org
  Storage Networking Industry Association
- http://www.sorrentonetworks.com
  Sorrento Networks, Inc.
- http://www.tivoli.com
  Tivoli
- http://www.t11.org
  Technical Committee T11
- http://www.vicom.com
  Vicom Systems
- http://www.vixel.com
  Vixel
- http://www.scsita.org
  SCSI Trade Association
- http://www.futureio.org
  InfiniBand (SM) Trade Association
- http://www.nsic.org
  National Storage Industry Consortium
- http://www.ietf.org
  Internet Engineering Task Force
- http://www.ansi.org
  American National Standards Institute
- http://www.standards.ieee.org
  Institute of Electrical and Electronics Engineers
- http://www.pc.ibm.com/us
  US Personal Systems Group

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.

Special notices

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

IBM may have patents or pending patent applications covering subject matter 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 the IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

The following terms are trademarks of other companies:

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.

Add/drop filter (ADF) A bandwidth-limited filter that passes only a specified wavelength or wavelength band between a device and a transmission path.

ADM Add/drop multiplexer digital multiplexing equipment that provides interfaces between different signals in a network.

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.

AIS Alarm Indication Signal.

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.

Asynchronous Data transmission in which the instant each character, or block of characters, starts is arbitrary. Once started, the time of occurrence of each signal (representing a bit within the character, or block) has the same relationship to significant instants of a fixed time frame.

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.

C-band Conventional band. In optical networks, a range of wavelengths between 1535 nm and 1565 nm.

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-delivery ability.

Class-2 A class of service providing a frame switching service between two ports with confirmed delivery or notification of non-delivery ability.

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-delivery ability.

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-delivery ability.

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.

CO Central office. A major equipment center designed to serve the communication traffic of a specific geographical area.

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.

Cross-over cable A cable that reverses Tx data and Rx data pin contacts. Cross-over cables are used to interconnect Ethernet hubs at a site.

DASD Direct Access Storage Device. Any online 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.

DEMUX Demultiplexer. A device at the receiving end of a transmission path that recovers the two or more signals originally combined by a multiplexer.

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.

Duplex A duplex cable contains two optical fibers; a duplex connector links two pairs of optical fibers.

DWDM Dense Wavelength Division Multiplexing. Similar to WDM but allows for information to be multiplexed over specific wavelengths.

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.

ECT Equalizer coupler tray. A module that is installed in the optical fiber amplifier (OFA) shelf and that separates, equalizes, and combines conventional and long band traffic (C-band and L-band).

EDFA Erbium-doped fiber amplifier. Optical fibers doped with the rare earth element erbium, which can amplify light in the 1550 nm region when pumped by an external light source.

EIA Electronic Industries Association. A standards association that specifies electrical transmission standards. The EIA and TIA have developed numerous well-known communications standards, including EIA/TIA-232 and EIA/TIA-449.

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.

E/O Electrical to optical conversion.

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 also refers to Ferrule connector A keyed, locking type of fiber-optic connector with a round barrel and threaded retaining ring.

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. A Fibre Channel industry association that works to promote awareness and understanding of the Fibre Channel technology and its application and provides a means for implementers to support the standards committee activities.

FCLC Fibre Channel Loop Association. An independent working group of the Fibre Channel Association focused on the marketing aspects of the Fibre Channel Loop technology.

FCP Fibre Channel Protocol. The mapping of SCSI-3 operations to Fibre Channel.

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, for example, 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.

Hz Hertz. A measure of frequency or bandwidth; one Hz equals one cycle per second.

In-Band Signaling Signaling that is carried in the same channel as the information. Also referred rto 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.

ISL Interswitch link.

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.

ITU-T International Telecommunication Union — Telecommunication Standardization Sector. The specialized agency of the United Nations for telecommunications. ITU is also the organization in which governments, private companies, and scientific and industrial institutions cooperate to improve the rational use of telecommunications.

ITU recommendations ITU standard wavelength designation. It is based on optical frequency with 100 GHz spacing. The anchor optical frequency is 193.1 THz (terahertz) corresponding to 1552.52 nm wavelength.

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-band Long band. In optical networks, a range of wavelengths between 1570 nm and 1620 nm.

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.

Laser Light amplification by stimulated emission of radiation. One of the wide range of devices that generates light by that principle. Laser light is directional, covers a narrow range of wavelengths, and is more coherent than ordinary light. Laser diodes are the standard light sources in fiber-optic systems.

Laser Diode An electro-optic device that produces coherent light within a narrow range of wavelengths commonly centered around 780 nm, 1310 nm, and 1550 nm. The wavelengths most commonly used in communications systems are 1310 nm and 1550 nm.

Latency A measurement of the time it takes to send a frame between two locations.

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.

Long-haul communications. Long-distance telecommunications links such as cross-country or transoceanic.

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.

Loss budget An accounting of overall attenuation in an optical system.

LVD Low Voltage Differential

MAN Metropolitan area network. A MAN consists of LANs interconnected within a radius of approximately 80 km (50 mi). MANs typically use fiber-optic cable to connect LANs.

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.

MAR Media Access Rules.

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).

MUX Multiplexer. A device that combines two or more signals into a signal composite data stream for transmission on a single channel.

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.

OADM Optical add/drop multiplexer. An add/drop multiplexer that allows the selective add and drop of optical channels at a particular site while passing the remaining channels through the optical fiber. The OADM acts as an optical router.

OC Optical carrier. Series of physical protocols (OC-1, OC-2, OC-3, and so on), defined for SONET optical signal transmissions. OC signal levels put STS frames onto fiber-optic line at a variety of speeds. The base rate is 51.84 Mbit/s (OC-1); each signal level thereafter operates at a speed divisible by that number (thus, OC-3 runs at 155.52 Mbit/s).

OCI Optical-channel interface. The circuit pack that interfaces with the customer equipment (subtending equipment) in an OPTera Metro 5300 network. A non-WDM circuit pack in the OPTera Metro 5200 node that provides an interface to the wavelength division multiplexing (WDM) fiber. Other types of OCI cards are OC-3, OC-12, Gigabit Ethernet, and FDDI.

OCLD Optical channel laser and detector. The optical channel transmitter and receiver circuit pack that interfaces with the WDM ring through the OMX in an OPTera Metro 5200 network.

OCM Optical channel manager. The circuit pack that does protection switching for the OPTera Metro 5200 network.

OFA Optical-fiber amplifier. An all-optical amplifier that uses erbium or other doped fibers and pump lasers to increase signal output power without electronic conversion.

OMX Optical multiplexer. A module that does optical add/drop operations.

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-electrical-optical conversion (O-E-O). The conversion of optical signals to electrical form, followed by reconversion to an optical signal. The conversion to electrical form is required to boost the signal, which fades over distance.

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 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.

Ring topology. A network topology in which terminals are connected serially point-to-point in an unbroken circle.

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 Subscriber connector. A push-pull type of fiber-optic connector with a square barrel.

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 De-serializer.

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.

Short-haul communications In common usage, this term is ordinarily applied to traffic between points less than 100 km (60 miles) apart.

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.

SONET Synchronous optical network. An interface standard for synchronous optical fiber transmission.

SSA Serial Storage Architecture - A high speed serial loop-based interface developed as a high speed point-to-point connection for peripherals, particularly high speed storage arrays, RAID and CD-ROM storage by IBM.

Star The physical configuration used with hubs in which each user is connected by communications links radiating out of a central hub that handles all communications.

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 (for example, 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.

Survivable network. A network that is capable of restoring traffic in the event of a failure condition.

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.

TIA Telecommunications Industry Association. Organization that develops standards relating to telecommunications technologies. Together, the TIA and the EIA have formalized standards, such as EIA/TIA-232, for the electrical characteristics of data transmission.

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.

U (vertical) unit. One U is 1.75 inches. Standard equipment racks have bolt holes spaced evenly on the mounting rails to permit equipment that is sized in multiples of this vertical unit to be mounted in the same rack.

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.

Wavelength The distance an electromagnetic wave travels in the time it takes to oscillate through a complete cycle. Wavelengths of light are measured in nanometers or microns.

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.

Index

D

E

F

G

H

I

J

L

M

N

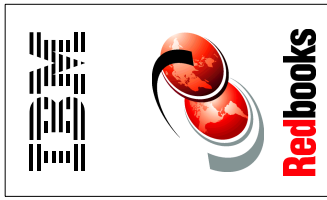
O

P

R

S

Y



Introduction to SAN Distance Solutions

(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

Introduction to SAN Distance Solutions

Learn the latest DWDM solutions that can be employed in a SAN

Build a SAN that puts distance between you and your data

Understand and design SAN distance solutions

The objective of this IBM Redbook is to provide information on the best configurations for distance solutions in a SAN environment. We show the particular business problems that distance solutions solve now; and with the future in mind, how these solutions can be expanded as the SAN world evolves.

We demonstrate the advantages that the IBM SAN and its OEM partners' and resellers' solutions bring to the marketplace. In addition, we provide information on the key factors to consider when choosing one particular solution over another, in order to protect and maximize your return on investment.

The distributed environment, particularly SAN, has resulted in a significant increase in communications. Data storage requirements have exploded. With these two developments in mind, it is vital to know how, where, and when data should be sent over distances quickly, and how to design and configure new and legacy systems while shaping them for the future.

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

SG24-6408-00 ISBN 0738424056