Draft Document for Review August 15, 2012 8:09 pm SG24-5470-04

Introduction to Storage Area Networks and System Networking

- Learn basic SAN and System Networking concepts
- Introduce yourself to the business benefits
- Discover the IBM System Networking portfolio

Jon Tate
Pall Beck
Hector Hugo Ibarra
Shanmuganathan Kumaravel
Libor Miklas

ibm.com/redbooks

Redbooks

International Technical Support Organization

Storage Area Networks and System Networking

March 2012

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

Fifth Edition (March 2012)

This edition applies to the products in the IBM System Networking portfolio.

This document was created or updated on August 15, 2012.

Contents

Notices

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US

registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

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

AFS®
AIX®
AS/400®
BladeCenter®
BNT®
DB2®
Domino®
DS4000®
DS6000™
DS8000®
Easy Tier®
ECKD™
Enterprise Storage Server®
ESCON®
Express Storage™
FICON®
FlashCopy®
GPFS™
HACMP™
HyperFactor®
i5/OS®
IBM Flex System™
IBM®
Informix®
iSeries®
Lotus®
OS/390®
OS/400®
Power Systems™
POWER6®
PowerHA®
PowerPC®
POWER®
ProtecTIER®
pSeries®
PureFlex™
Redbooks®
Redbooks (logo) ®
RMF™
RS/6000®
ServerProven®
Storwize®
System i®
System p®
System Storage®
System x®
System z9®
System z®
Tivoli®
VMready®
XIV®
xSeries®
z/OS®
z/VM®
z9®
zSeries®

The following terms are trademarks of other companies:

Intel Xeon, Intel, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

LTO, Ultrium, the LTO Logo and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

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

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

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

Summary of changes

This section describes the technical changes made in this edition of the book and in previous editions. This edition might also include minor corrections and editorial changes that are not identified.

Summary of Changes
for SG24-5470-04
for Storage Area Networks and System Networking
as created or updated on August 15, 2012.

March 2012, Fifth Edition

This revision includes amendments, deletions, and additions to support IBM® strategies and initiatives, and to add an introduction to IBM System Networking, as well as update those areas of Storage Area Networking as appropriate.

For any omissions or inaccuracies contact Jon Tate (tatej@uk.ibm.com).

Preface

The plethora of data created by the businesses of today is making storage a strategic investment priority for companies of all sizes. As storage takes precedence, three major initiatives have emerged:

- Flatten and converge your network

  IBM takes an open, standards-based approach to implement the latest advances in today's flat, converged data center network designs. IBM System Networking solutions enable clients to deploy a high-speed, low-latency Unified Fabric Architecture.

- Optimize and automate virtualization

  Advanced virtualization awareness reduces the cost and complexity of deploying physical and virtual data center infrastructure.

- Simplify management

  IBM data center networks are easy to deploy, maintain, scale and virtualize, delivering the foundation of consolidated operations for dynamic infrastructure management.

Storage is no longer an afterthought. Too much is at stake. Companies are searching for more ways to efficiently manage expanding volumes of data, and to make that data accessible throughout the enterprise; this is propelling the move of storage into the network. Also, the increasing complexity of managing large numbers of storage devices and vast amounts of data is driving greater business value into software and services.

With current estimates of the amount of data to be managed and made available increasing at 60 percent per annum, this is where a storage area network (SAN) enters the arena. Simply put, SANs are the leading storage infrastructure for the global economy of today. SANs offer simplified storage management, scalability, flexibility, availability, and improved data access, movement, and backup.

Welcome to the era of Smarter Networking for Smarter Data Centers!

The smarter data center with improved economics of IT can be achieved by connecting servers and storage with a high-speed and intelligent network fabric that is smarter, faster, greener, open and easy to manage. IBM System Networking solutions:

This book gives an introduction to the SAN, Ethernet Networking, and how these help achieve a smarter data center.

For further reading, and a deeper dive into the SAN world, readers may find the following redbook especially useful to expand their SAN knowledge:

Implementing an IBM b-type SAN with 8 Gbps Directors and Switches, SG24-6116

Implementing the IBM System Storage SAN32B-E4 Encryption Switch, SG24-7922

IBM System Storage b-type Multiprotocol Routing: An Introduction and Implementation, SG24-7544

IBM Converged Switch B32, SG24-7935

Also be sure to visit the IBM Redbooks® System Networking portal for the latest material from the International Technical Support Organization:

http://www.redbooks.ibm.com/portals/networking

The team who wrote this book

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

Jon Tate is a Project Manager for IBM System Storage® SAN Solutions at the International Technical Support Organization, San Jose Center. Before joining the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2 support for IBM storage products. Jon has 27 years of experience in storage software and management, services, and support, and is both an IBM Certified IT Specialist and an IBM SAN Certified Specialist. He is also the UK Chairman of the Storage Networking Industry Association.

Pall Beck is a SAN Technical Lead in IBM Nordic. He has 15 years of experience working with storage, both for dedicated clients as well as very large shared environments. Those environments include clients from the medical and financial sector including some of the largest shared SAN environments in Europe. He is a member of the SAN and SVC best practice community in the Nordics and in EMEA. In his current job role he is a member of a SCE+ Storage Deployment specialists, responsible for SCE+ storage deployments around the globe. Pall is also a member of a team assisting in critical situations and doing root cause analyzes. He is co-author of the Implementing SVC 5.1 and SVC Advanced Copy Services 4.2 Redbooks. Pall has a diploma as an Electronic Technician from Odense Tekniske Skole in Denmark and IR in Reykjavik, Iceland and he is IBM Certified IT Specialist.

Hector Hugo Ibarra is an Infrastructure IT Architect specialized in cloud computing and storage solutions currently working at the IBM Argentina. Hector has been designated as ITA Leader for The VMware Center of Competence in 2006. He specializes in virtualization technologies and has assisted several global IBM clients to deploy virtualized infrastructures across the world. Since 2009 he has been working as the Leader for the Argentina Delivery Center Strategy and Architecture Services department from where major projects are driven.

Shanmuganathan Kumaravel is an IBM Technical Services Specialist for the ITD-SSO MR Storage team of IBM India. He supports SAN, and disk products of both IBM and Hewlett Packard since August 2008. Prior to this he worked for HP product support providing remote support on HP SAN storage products, servers and operating systems including HP Unix and Linux. Shan is a Brocade Certified SAN Designer (BCSD), Brocade Certified Fabric Administrator (BCFA) and an HP Certified System Engineer (HPCSE).

Libor Miklas is a Team Leader and an experienced IT Specialist working at the IBM Delivery Center Central Europe in Czech Republic. He demonstrates ten years of practical experience within the IT industry. During last six years, his main focus has been on backup and recovery and on storage management. Libor and his team support midrange and enterprise storage environments for various global and local clients, worldwide. He is an IBM Certified Deployment Professional of the Tivoli® Storage Manager family of products and holds a Masters Degree in Electrical Engineering and Telecommunications.

Figure 1 shows the writing team.

Figure 1 Left-to-right: Libor, Pall, Jon, Hector, and Shan

Thanks to the previous authors of the first, second, third, and fourth editions of this book:

Angelo Bernasconi
Rajani Kanth
Ravi Kumar Khattar
Fabiano Lucchese
Peter Mescher
Richard Moore
Mark S. Murphy
Kjell E. Nyström
Fred Scholten
Giulio John Tarella
Andre Telles

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

Sangam Racherla
International Technical Support Organization, San Jose Center

Special thanks to the Brocade staff for their unparalleled support of this residency in terms of equipment and support in many areas:

Brian Steffler
Silviano Gaona
Marcus Thordal
Steven Tong
Jim Baldyga
Brocade Communications Systems

John McKibben
Cisco Systems

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

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

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

  ibm.com/redbooks
- Send your comments in an email to:

  redbooks@us.ibm.com
- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks
- Follow us on Twitter:

  http://twitter.com/ibmredbooks
- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806
- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm
- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Introduction

Computing is based on information. Information is the underlying resource on which all computing processes are based; it is a company asset. Information is stored on storage media, and is accessed by applications executing on a server. Often the information is a unique company asset. Information is created and acquired every second of every day. Information is the currency of business.

To ensure that any business delivers the expected results, they must have access to accurate information, and without delay. The management and protection of business information is vital for the availability of business processes.

This chapter introduces the concept of a Network, Storage, and the Storage Area Network, which has been regarded as the ultimate response to all these needs.

1.1 What is a network

A computer network, often simply referred to as a network, is a collection of computers and devices interconnected by communication channels that allows for the efficient sharing of resources, services and information among it.

Even though this definition is simple, understanding how to make a network work might be complicated for those who are not familiar with IT, or who are just starting out in the IT world. Because of this we will explain the basic concepts that need to be understood to facilitate our understanding of the networking world.

1.1.1 The importance of communication

It is impossible to imagine the human world as standalone humans, with nobody talking to each other, nobody doing anything for each other, and much more important trying to imagine how a human can work without using their senses. In our human world, we are sure you would agree with us that communication between individuals has made a significant difference in all aspects of life.

First of all, communication in any form is not easy, and we need a number of components: a common language, something to be communicated, a medium where the communication will flow, and finally we need to be sure that whatever was communicated was received and understood. In order to do that in the human world we use language as a communication protocol, and sounds and writing are the communication medium.

Similarly, a computer network needs almost the same components as our human example, but a difference is that all needs to be governed in some way to ensure effective communications. This is achieved by the use of industry standards, and companies will adhere to those standards to ensure that communication can take place.

There is a wealth of information devoted to networking history and its evolution, and we do not intend to give a history lesson in this book. In this book we will focus on the prevalent interconnection models, storage, and networking concepts.

1.2 Interconnection Models

An interconnection model is a standard used to connect sources and targets in a network, and there are some well known models in the IT industry such as OSI, DOD, TCP/IP protocol suite, Fibre Channel, and so on. Each model has its advantages and disadvantages and its model is applied where its has the maximum benefit in terms of performance, reliability, availability, cost benefits, and so on.

1.2.1 The OSI Model

The Open Systems Interconnection model (OSI model) was a product of the Open Systems Interconnection effort at the International Organization for Standardization. It is a way of sub-dividing a communications system into smaller parts called layers. Similar communication functions are grouped into logical layers. A layer provides services to its upper layer while receiving services from the layer below. At each layer, an instance provides service to the instances at the layer above and requests service from the layer below.

For the purpose of this book we well focus on the Physical, Data Link, Network and Transport layers.

Layer 1: Physical Layer

The Physical Layer defines electrical and physical specifications for devices. In particular, it defines the relationship between a device and a transmission medium, such as a copper or optical cable. This includes the layout of pins, voltages, cable specifications and more.

Layer 2: Data Link Layer

The Data Link Layer provides the functional and procedural means to transfer data between network entities and to detect and possibly correct errors that may occur in the Physical Layer.

Layer 3: Network Layer

The Network Layer provides the functional and procedural means of transferring variable length data sequences from a source host on one network to a destination host on a different network, while maintaining the quality of service requested by the Transport Layer (in contrast to the data link layer which connects hosts within the same network). The Network Layer performs network routing functions, and might also perform fragmentation and reassembly, and report delivery errors. Routers operate at this layer-sending data throughout the extended network and making the Internet possible.

Layer 4: Transport Layer

The Transport Layer provides transparent transfer of data between end users, providing reliable data transfer services to the upper layers. The Transport Layer controls the reliability of a given link through flow control, segmentation and desegmentation, and error control. Some protocols are state- and connection-oriented. This means that the Transport Layer can keep track of the segments and retransmit those that fail. The Transport layer also provides the acknowledgement of the successful data

Now that you know what an interconnection model is, what it does and how important it is in a network we can compare the OSI model with other models as shown in Figure 1-1.

OSI layer #	name	TCP/IP	Fibre Channel
5-7	application	telnet, ftp, SCSI-3 (iSCSI)	IP, SCSI-3 (FCP)
4	transport	TCP, UDP	FC-4
3	network	IP, ICMP, IGMP	FC-3
2	data link	Ethernet, Token Ring	FC-2, most of FC-1
1	physical	media	FC-0

Figure 1-1 OSI, TCP/IP and FC Models Comparison Table

The Fibre Channel Model will be covered later in this book.

1.2.2 Translating the OSI Model to the physical world

In order make a translation from theoretical models to reality we will introduce physical devices which perform certain tasks for each layer on each model.

Local Area Networks (LANs) are a good place to start and we will define it as a small or large network limited within the same physical site, and that could be a traditional office or a corporate building.

In Figure 1-2 on page 4 you will see a basic network where computers and a printer are interconnected using physical cables and interconnection devices.

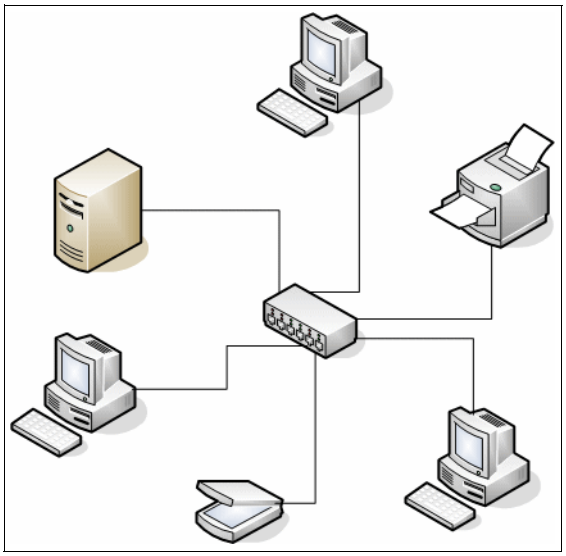

Figure 1-2 Basic network topology

We must keep in mind that any model we choose will define the devices, cables, connectors and interfaces characteristics that we must implement to make it work, as well as supporting the protocols for each model layer.

We will categorize all the network components into five groups:

- End Devices: an end device is a computer system which has a final purpose like desktop computers, printers, storages or servers.
- Network Interface: it is an interface between the media and the end devices which can interact with other network interfaces and understands an interconnection model.
- Connector: this is the physical element at the end of the media which allows connection to the Network Interface.
- Media: this is the physical path used to transmit an electrical or optical signal and it could be wired or wireless, copper or fiber optic cable.
- Network Devices: these are used to interconnect multiple end devices as a single point of interconnection, route communication through different networks, or for providing network security. Examples of network devices are switches, routers, firewalls and directors.

As we mentioned earlier each network component executes a particular role within a network and all of them are required to reach the final goal of making communication possible.

1.3 What do we mean by storage?

To understand what storage is, and because understanding it is a key point for this book, we will start from a basic hard drive progressing through to high performing, fault tolerant, highly available, storage systems. During our explanation we will use instructional examples that may sometimes not reflect reality; however they will make it easier to understand for those that are just beginning to enter the world of storage systems.

1.3.1 Storing data

Data is stored on hard disk drives (HDD) which can be read and written to. Depending on the methodology used to execute those tasks, and the HDD technology on which they were built, the read and write could be faster or slower. The evolution of hard drives has been quite incredible and nowadays we can store hundreds of gigabytes on a single HDD allowing us to keep all the data we could ever imagine. Even though this approach seems to bring us only advantages so far, one question could be what happens if for any reason we are unable to access the HDD?

The first solution could be to have a secondary HDD where we can manually copy our primary HDD to our secondary HDD. Immediately we can see that now our data is safe but how often should we run those manual copies if we expect not to lose data and to keep it as up to date as possible? To keep it as current as possible, every time we change something we must make another copy — but should we copy the entire amount of data from one HDD to the other, or should we copy only what has changed?

Figure 1-3 on page 5 shows a manual copy of data for redundancy.

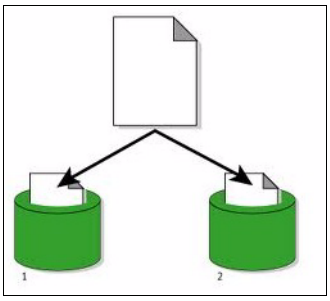

Figure 1-3 Manual copy of data

1.3.2 RAID

Fortunately technology exists that can help us and that is the Redundant Array of Independent Disks (RAID) concept which presents a possible solution to our problem. It is clear that data needs to be copied every time it changes in order to provide us with a reliable fault tolerant system, and it is also clear that it cannot be done in a manual way. A RAID Controller can maintain disks in synchronization and also manage all the writes and reads (Inputs and Outputs I/O) to and from the disks.

So now our RAID system looks like that shown in Figure 1-4. Where A, B and C represent user data such as documents or pictures.

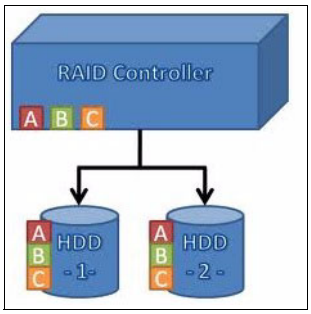

Figure 1-4 Typical RAID scenario

This type of RAID is known as RAID 1 or as mirrored disk.

Advantages to a standalone disk are:

- Provide redundancy to disk failure
- Faster when reading data as it can be taken from either disk

Disadvantages to a standalone disk are:

- Slower when writing as data needs to be written twice
- Only half of the total capacity can be used

This naturally leads us to the next question: is there any other RAID type that can improve things further while conserving the advantages and removing the disadvantages of RAID 1?

The answer is yes, and this is known as RAID 5. It consists of dividing the user data into N-1 parts (where N is the number of disks used to build the RAID) and then calculating a parity part which permits RAID to rebuild user data in case of a disk failure.

RAID5 uses “parity” or redundant information. If a block fails, enough parity information is available to recover the data. The parity information is spread across all the disks. If a disk fails the RAID requires a rebuild and the parity information is used to re-create the data lost. This example is shown in Figure 1-5.

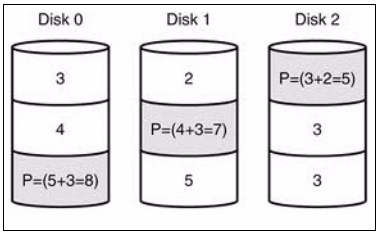

Figure 1-5 Example of RAID 5 with parity.

RAID 5 requires a minimum of 3 disks and in theory there are no limitations to add disks. It combines data safety with efficient use of disk space. Disk failure does not result in a service interruption because data is read from parity blocks. RAID 5 is useful for people who need performance and constant access to their data.

In RAID 5+Spare disk failure does not require immediate attention because the system rebuilds itself using the hot spare, but the failed disk should be replaced as soon as possible. A Spare disk is an empty disk which is used by the RAID controller only when a disk fails. Figure 1-6 on page 7 shows this example.

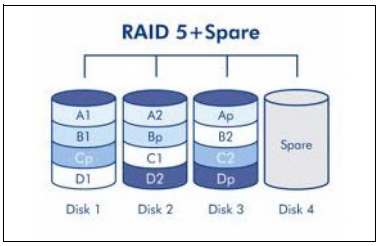

Figure 1-6 RAID 5 with hot spare

RAID 5 has better performance for I/O than RAID 1 and depending on the number of disks used to build the RAID the array disk space utilization is more of 2/3. RAID 5 is also managed by a RAID controller which performs the same role as in RAID 1.

Important: RAID 1 and 5 are the most common RAID levels but there are many others not covered in this book such as RAID 0, 3,4,6 or Nested (hybrid) like RAID 0+1 or RAID 5+1 used in environments where reliability and performance are key points to be covered from the storage perspective.

Figure 1-7 on page 7 shows a brief comparison between the most common RAID levels.

Features	RAID 0	RAID 1	RAID 5
Minimum # Drives	2	2	3
Data Protection	No Protection	Single-drive failure	Single-drive failure
Read Performance	High	High	High
Write Performance	High	Medium	Low
Read Performance (degraded)	N/A	Medium	Low
Write Performance (degraded)	N/A	High	Low
Capacity Utilization	100%	50%	67% - 94%
Typical Applications	High End Workstations, data logging, real-time rendering, very transitory data	Operating System, transaction databases	Data warehousing, web serving, archiving

Figure 1-7 RAID level comparison table

Our disk systems now seem to be ready to support failures, and they are also high performing - but what if our RAID controller fails? Maybe we will not lose data but it will not be accessible, so is there a solution to this?

It is almost the same scenario we initially faced having only one disk as a storage system this type of scenarios is well known as a Single Point of Failure (SPoF) so we must add redundancy by introducing a secondary RAID controller to our storage system.

Now we are sure that no matter what happens data will be available to be used.

Note: The RAID controller role in some cases is performed by the software, and this solution is less expensive than a hardware solution as it does not require controller hardware, however it is a slower solution.

At this point we have a number of physical hard drives managed by two controllers.

Disk Pools

When a logical storage volume needs to be provisioned to servers, firstly the storage RAID needs to be created, and the way of doing this is by selecting available HDDs and grouping them together for a single purpose. The number of grouped HDDs will depend on the RAID type we choose, and the space required for provisioning.

To understand what we mean we will show a basic example using the following assumptions:

- There are 10 hard drives, named A,B,C,D,E,F,G,H,I,J.
- There are two RAID controllers supporting any RAID level, named RC1 and RC2.
- Each RAID controller can manage any hard drive.
- Each RAID controller can act as backup of the other on any time.

Now we can perform the following tasks

- Select hard drives A, B and F and create a RAID 5 array managed by RC1 and we will call it G1
- Select hard drives E, I and J and create a RAID 5 array managed by RC2 and we will call it G2

Figure 1-8 on page 8 shows these steps.

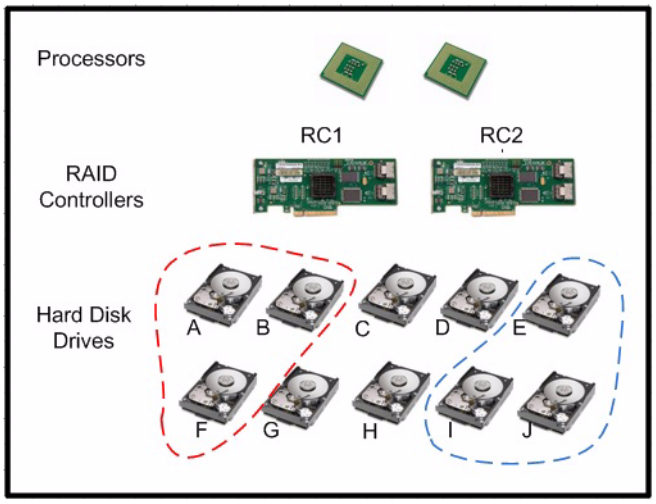

Figure 1-8 Disk pool creation

What we are doing by executing these simple steps is creating disk pools which basically consists of grouping disks together for a single purpose such as creating a RAID level, in our case RAID 5

Earlier we mentioned nested (hybrid) RAIDs such as 5+0. Solutions like these are used when the amount of storage data is significant and is extremely important for business continuity. RAID 50 comprises of RAID 0 striping across lower-level RAID 5 arrays. The benefits of RAID 5 are gained while the spanned RAID 0 allows the incorporation of many more disks into a single logical drive. Up to one drive in each sub-array may fail without loss of data. Also, rebuild times are substantially less then a single large RAID 5 array as shown Figure 1-9 on page 9.

Note: Hybrid or nested RAIDs are a combination of existing RAID levels that create a new RAID to reap the benefits of two different RAID levels.

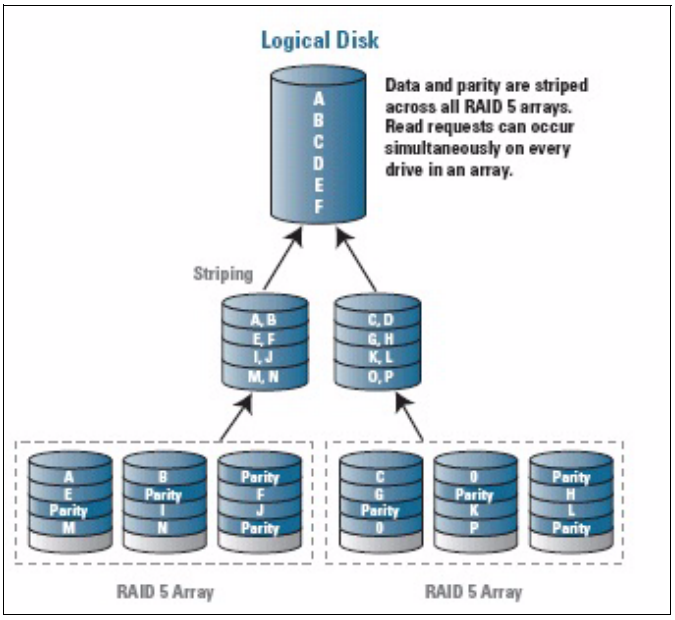

Figure 1-9 Nested (hybrid) RAID 5+0 or RAID 50

This nested RAID 50 can now be managed by RC1 or RC2 so we have full redundancy.

Storage Systems

We are now not so far away from building our basic storage system. However, to answer our previous questions we need to add two new components and an enclosure.

One of those two components will be a CPU which will process all the required instructions in order to allow data to flow. Of course adding one CPU will create a single point of failure (SPoF) so we will add two CPUs.

In addition to this, now that we almost have an independent system and referring back to our networking section, this system must be able to communicate with other systems in a network so it will require a minimum of two network interfaces, once again to avoid a SPoF.

Now there is only one step left and that is to put all these hardware components into an enclosure. Now our Storage System should look like that shown in Figure

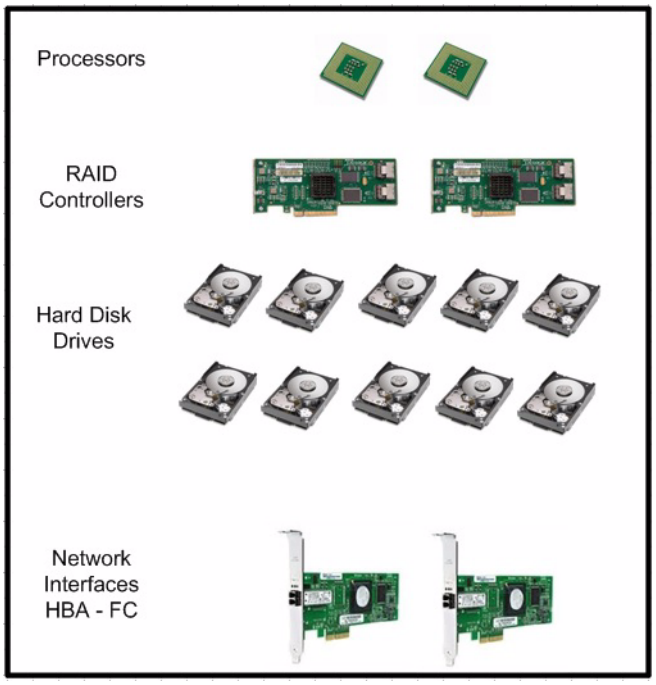

Figure 1-10 Basic Storage System

Note: We have presented a basic storage configuration for understanding of course it could have as many CPUs, RAID Controllers, Network Interfaces and Hard Disk Drives as needed.

1.4 What is a Storage Area Network?

The Storage Network Industry Association (SNIA) defines the SAN as a network whose primary purpose is the transfer of data between computer systems and storage elements. A SAN consists of a communication infrastructure, which provides physical connections; and a management layer, which organizes the connections, storage elements, and computer systems so that data transfer is secure and robust. The term SAN is usually (but not necessarily) identified with block I/O services rather than file access services.

Put in simple terms, a SAN is a specialized, high-speed network attaching servers and storage devices and, for this reason, It is sometimes referred to as “the network behind the servers.” A SAN allows “any-to-any” connection across the network, using interconnect elements such as switches and directors. It eliminates the traditional dedicated connection between a server and storage, and the concept that the server effectively “owns and manages” the storage devices. It also eliminates any restriction to the amount of data that a server can access, currently limited by the number of storage devices attached to the individual server. Instead, a SAN introduces the flexibility of networking to enable one server or many heterogeneous servers to share a common storage utility, which may comprise many storage devices, including disk, tape, and optical storage. Additionally, the storage utility may be located far from the servers that use it.

The SAN can be viewed as an extension to the storage bus concept, which enables storage devices and servers to be interconnected using similar elements as in local area networks (LANs) and wide area networks (WANs)

The diagram in Figure 1-11 shows a tiered overview of a SAN connecting multiple servers to multiple storage systems.

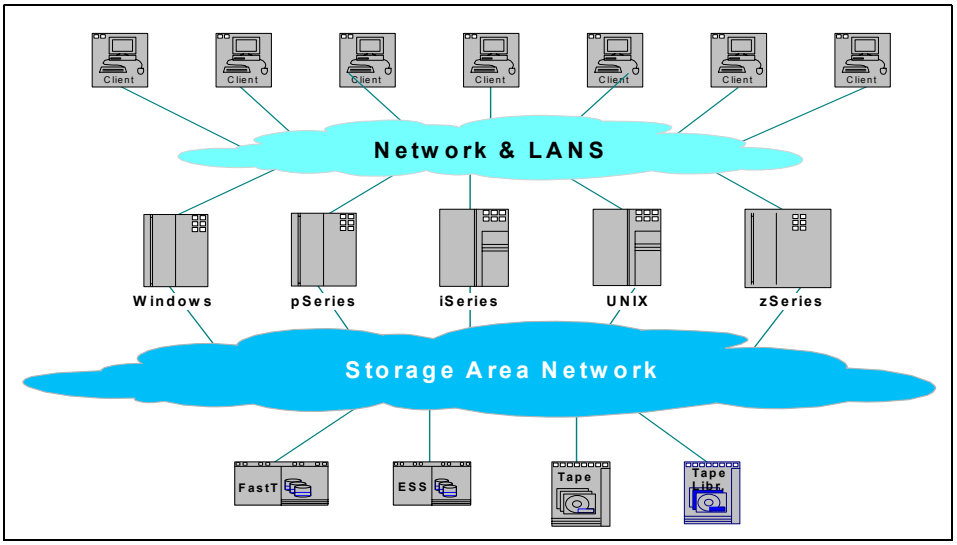

Figure 1-11 A SAN

SANs create new methods of attaching storage to servers. These new methods can enable great improvements in both availability and performance. Today’s SANs are used to connect shared storage arrays and tape libraries to multiple servers, and are used by clustered servers for failover.

A SAN can be used to bypass traditional network bottlenecks. It facilitates direct, high-speed data transfers between servers and storage devices, potentially in any of the following three ways:

- Server to storage: This is the traditional model of interaction with storage devices. The advantage is that the same storage device may be accessed serially or concurrently by multiple servers.
- Server to server: A SAN may be used for high-speed, high-volume communications between servers.
- Storage to storage: This outboard data movement capability enables data to be moved without server intervention, thereby freeing up server processor cycles for other activities like application processing. Examples include a disk device backing up its data to a tape device without server intervention, or remote device mirroring across the SAN.

SANs allow applications that move data to perform better, for example, by having the data sent directly from the source to the target device with minimal server intervention. SANs also enable new network architectures where multiple hosts access multiple storage devices connected to the same network. Using a SAN can potentially offer the following benefits:

- Improvements to application availability: Storage is independent of applications and accessible through multiple data paths for better reliability, availability, and serviceability.
- Higher application performance: Storage processing is off-loaded from servers and moved onto a separate network.
- Centralized and consolidated storage: Simpler management, scalability, flexibility, and availability.
- Data transfer and vaulting to remote sites: Remote copy of data enabled for disaster protection and against malicious attacks.
- Simplified centralized management: Single image of storage media simplifies management.

1.5 SAN components

As stated previously, Fibre Channel is the predominant architecture upon which most SAN implementations are built, with FICON® as the standard protocol for z/OS® systems, and FCP as the standard protocol for open systems. The SAN components described in the following sections are Fibre Channel-based, and are shown in Figure 1-12 on page 13.

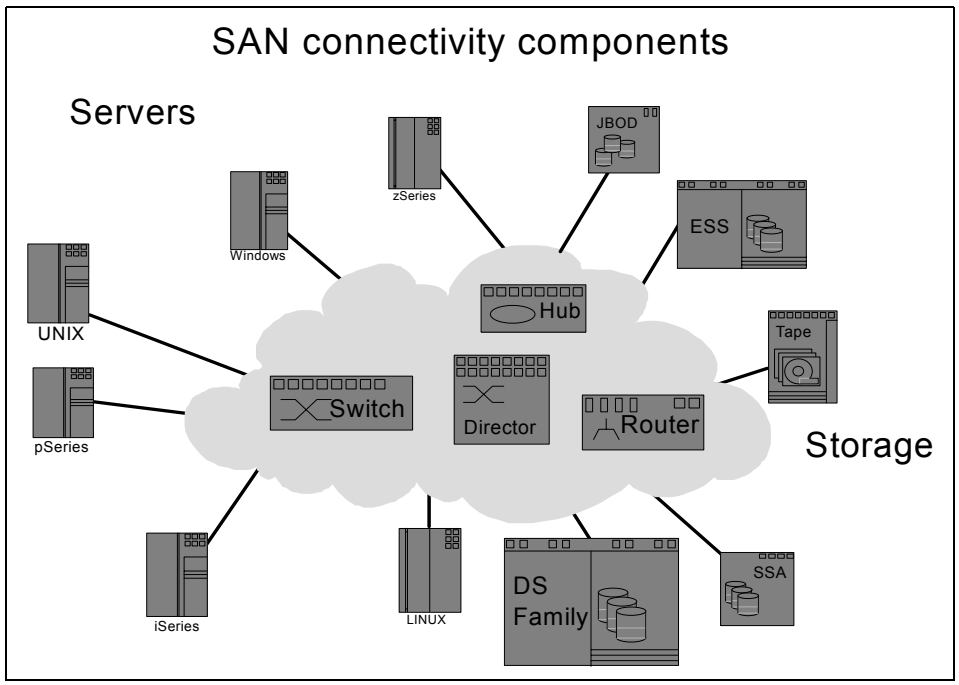

Figure 1-12 SAN components

1.5.1 SAN connectivity

The first element that must be considered in any SAN implementation is the connectivity of storage and server components typically using Fibre Channel. The components listed above have typically been used for LAN and WAN implementations. SANs, like LANs, interconnect the storage interfaces together into many network configurations and across longer distances.

Much of the terminology used for SAN has its origins in IP network terminology. In some cases, the industry and IBM use different terms that mean the same thing, and in some cases, mean different things.

1.5.2 SAN storage

The SAN liberates the storage device so it is not on a particular server bus, and attaches it directly to the network. In other words, storage is externalized and can be functionally distributed across the organization. The SAN also enables the centralization of storage devices and the clustering of servers, which has the potential to make for easier and less expensive centralized administration that lowers the total cost of ownership (TCO).

The storage infrastructure is the foundation on which information relies, and therefore must support a company's business objectives and business model. In this environment simply deploying more and faster storage devices is not enough. A SAN infrastructure provides enhanced network availability, data accessibility, and system manageability, and It is important to remember that a good SAN begins with a good design. This is not only a maxim, but must be a philosophy when we design or implement a SAN.

1.5.3 SAN servers

The server infrastructure is the underlying reason for all SAN solutions. This infrastructure includes a mix of server platforms such as Windows, UNIX (and its various flavors), and z/OS. With initiatives such as server consolidation and e-business, the need for SANs will increase, making the importance of storage in the network greater.

1.6 The importance of standards or models

Why do we care about standards? Standards are the starting point for the potential interoperability of devices and software from different vendors in the SAN marketplace. SNIA, among others, defined and ratified the standards for the SANs of today, and will keep defining the standards for tomorrow. All of the players in the SAN industry are using these standards now, as these are the basis for wide acceptance of SANs. Widely accepted standards potentially allow for heterogeneous, cross-platform, multivendor deployment of SAN solutions.

As all vendors have accepted these SAN standards, there *should* be no problem in connecting the different vendors into the same SAN network. However, nearly every vendor has an interoperability lab where it tests all kind of combinations between their products and those of other vendors. Some of the most important aspects in these tests are the reliability, error recovery, and performance. If a combination has passed the test, that vendor is going to certify or support this combination.

IBM participates in many industry standards organizations that work in the field of SANs. IBM believes that industry standards must be in place, and if necessary, re-defined for SANs to be a major part of the IT business mainstream.

Probably the most important industry standards organization for SANs is the Storage Networking Industry Association (SNIA). IBM is a founding member and board officer in SNIA. SNIA and other standards organizations and IBM is an active participant in many of these organizations.

2

Why, and how can we, use a SAN?

In the previous chapter, we introduced the basics by presenting a network and storage system introduction, we also worked on a standard SAN definition, as well as a brief description of the underlying technologies and concepts that are behind a SAN implementation.

In this chapter, we extend this discussion by presenting real-life SAN alongside well-known technologies and platforms used in SAN implementations. We also discuss some of the trends that are driving SAN evolution, and how they may affect the future of storage technology.

And although the technology is different, a lot of the concepts can also be applied in the Ethernet networking environment too, which we cover later in this book in more depth.

2.1 Why use a SAN?

In this section we describe the main motivators that drive SAN implementations, and present some of the key benefits that this technology might bring to data-dependent business.

2.1.1 The problem

Distributed clients and servers are frequently chosen to meet specific application needs. They may, therefore, run different operating systems (such as Windows Server, UNIX of differing flavors, VMware vSphere, VMS, and so on), and different database software (for example, DB2®, Oracle, Informix®, SQL Server). Consequently, they have different file systems and different data formats.

Managing this multi-platform, multivendor, networked environment has become increasingly complex and costly. Multiple vendor's software tools, and appropriately skilled human resources must be maintained to handle data and storage resource management on the many differing systems in the enterprise. Surveys published by industry analysts consistently show that management costs associated with distributed storage are much greater, up to 10 times more, than the cost of managing consolidated or centralized storage. This includes costs of backup, recovery, space management, performance management, and disaster recovery planning.

Disk storage is often purchased from the processor vendor as an integral feature, and it is difficult to establish if the price you pay per gigabyte (GB) is competitive, compared to the market price of disk storage. Disks and tape drives, directly attached to one client or server, cannot be used by other systems, leading to inefficient use of hardware resources. Organizations often find that they have to purchase more storage capacity, even though free capacity is available in other platforms.

Additionally, it is difficult to scale capacity and performance to meet rapidly changing requirements, such as the explosive growth in server, application and desktop virtualization, and the need to manage information over its entire life cycle, from conception to intentional destruction.

Information stored on one system cannot readily be made available to other users, except by creating duplicate copies, and moving the copy to storage that is attached to another server. Movement of large files of data may result in significant degradation of performance of the LAN/WAN, causing conflicts with mission-critical applications. Multiple copies of the same data may lead to inconsistencies between one copy and another. Data spread on multiple small systems is difficult to coordinate and share for enterprise-wide applications, such as e-business, Enterprise Resource Planning (ERP), Data Warehouse, and Business Intelligence (BI).

Backup and recovery operations across a LAN may also cause serious disruption to normal application traffic. Even using fast Gigabit Ethernet transport, sustained throughput from a single server to tape is about 25 GB per hour. It would take approximately 12 hours to fully back up a relatively moderate departmental database of 300 GBs. This may exceed the available window of time in which this must be completed, and it may not be a practical solution if business operations span multiple time zones. It is increasingly evident to IT managers that these characteristics of client/server computing are too costly, and too inefficient. The islands of information resulting from the distributed model of computing do not match the needs of the enterprise.

New ways must be found to control costs, improve efficiency, and simplify the storage infrastructure to meet the requirements of the modern business world.

2.1.2 The requirements

With this scenario in mind, we can think of a number of requirements that today's storage infrastructures should meet. Some of the most important are:

- Unlimited and just-in-time scalability. Businesses require the capability to flexibly adapt to rapidly changing demands for storage resources without performance degradation.
- System simplification. Businesses require an easy-to-implement infrastructure with the minimum of management and maintenance. The more complex the enterprise environment, the more costs are involved in terms of management. Simplifying the infrastructure can save costs and provide a greater return on investment (ROI).
- Flexible and heterogeneous connectivity. The storage resource must be able to support whatever platforms are within the IT environment. This is essentially an investment protection requirement that allows you to configure a storage resource for one set of systems, and subsequently configure part of the capacity to other systems on an as-needed basis.
- Security. This requirement guarantees that data from one application or system does not become overlaid or corrupted by other applications or systems. Authorization also requires the ability to fence off one system's data from other systems.
- Encryption. When sensitive data is stored it must be read or written only from those authorized systems and if for any reason the storage system is stolen, data must never be available to be read from it.
- Hypervisors. Support of server, application and desktop virtualization hypervisors features for cloud computing.
- Speed. Storage networks and devices must be capable of managing the high number of Gigabytes and intensive I/O required by each business industry.
- Availability. This is a requirement that implies both protection against media failure as well as ease of data migration between devices, without interrupting application processing. This certainly implies improvements to backup and recovery processes: attaching disk and tape devices to the same networked infrastructure allows for fast data movement between devices, which provides enhanced backup and recovery capabilities, such as:
  - Serverless backup. This is the ability to back up your data without using the computing processor of your servers.
  - Synchronous copy. This makes sure your data is at two or more places before your application goes to the next step.
  - Asynchronous copy. This makes sure your data is at two or more places within a short time. It is the disk subsystem that controls the data flow.

In the next section, we discuss the use of SANs as a response to these business requirements.

2.2 How can we use a SAN?

The key benefits that a SAN might bring to a highly data-dependent business infrastructure can be summarized into three rather simple concepts: infrastructure simplification, information lifecycle management and business continuity. They are an effective response to the requirements presented in the previous section, and are strong arguments for the adoption of SANs.

These three concepts are briefly described as follows.

2.2.1 Infrastructure simplification

There are four main methods by which infrastructure simplification can be achieved: ***consolidation, virtualization, automation*** and ***integration***:

- Consolidation

  Concentrating systems and resources into locations with fewer, but more powerful, servers and storage pools can help increase IT efficiency and simplify the infrastructure. Additionally, centralized storage management tools can help improve scalability, availability, and disaster tolerance.

- Virtualization

  Storage virtualization helps in making complexity nearly transparent and at the same time can offer a composite view of storage assets. This may help reduce capital and administrative costs, while giving users better service and availability. Virtualization is designed to help make the IT infrastructure more responsive, scalable, and available.

- Automation

  Choosing storage components with autonomic capabilities can improve availability and responsiveness—and help protect data as storage needs grow. As soon as day-to-day tasks are automated, storage administrators may be able to spend more time on critical, higher-level tasks unique to a company's business mission.

- Integration

  Integrated storage environments simplify system management tasks and improve security. When all servers have secure access to all data, your infrastructure may be better able to respond to your users information needs.

Figure 2-1 illustrates the consolidation movement from the distributed islands of information toward a single, and, most importantly, simplified infrastructure.

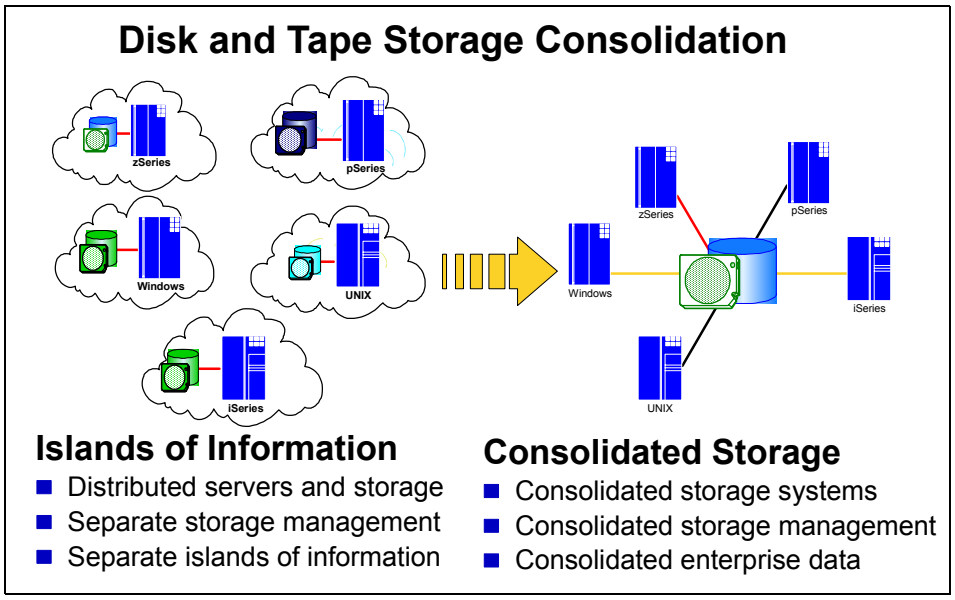

Figure 2-1 Disk and tape storage consolidation

Simplified storage environments have fewer elements to manage, which leads to increased resource utilization, simplifies storage management, and can provide economies of scale for owning disk storage servers. These environments can be more resilient and provide an infrastructure for virtualization and automation.

2.2.2 Information lifecycle management

Information has become an increasingly valuable asset, but as the amount of information grows, it becomes increasingly costly and complex to store and manage it. Information lifecycle management (ILM) is a process for managing information through its life cycle, from conception until intentional disposal, in a manner that optimizes storage, and maintains a high level of access at the lowest cost.

A SAN implementation makes it easier to manage the information lifecycle as it integrates applications and data into a single-view system, in which information resides, and can be managed more efficiently.

IBM Tivoli Productivity Center for Data was specially designed to support ILM.

2.2.3 Business continuity

It goes without saying that the business climate in today's on demand era is highly competitive. Customers, employees, suppliers, and business partners expect to be able to tap into their information at any hour of the day from any corner of the globe. Continuous business operations are no longer optional—they are a business imperative to becoming successful, and maintaining a competitive advantage. Businesses must also be increasingly sensitive to issues of customer privacy and data security, so that vital information assets are not compromised. Factor in those legal and regulatory requirements, and the inherent demands of participating in the global economy, and accountability, and all of a sudden the lot of an IT manager is not a happy one.

Nowadays, with natural disasters seemingly occurring with more frequency a Disaster Recovery (DR) plan is essential and implementing the correct SAN solution can help not only in real-time recovery techniques, but it also can reduce the recovery time objective (RTO) for your current DR plan.

There are many specific vendors solutions in the market which require a SAN running in the background like VMware Site Recovery Manager for business continuity.

It is little wonder that a sound and comprehensive business continuity strategy has become a business imperative, and SANs play a key role in this. By deploying a consistent and safe infrastructure, they make it possible to meet any availability requirements.

2.3 Using the SAN components

The foundation that a SAN is built on is the interconnection of storage devices and servers. This section further discusses storage, interconnection components, and servers, and how different types of servers and storage are used in a typical SAN environment.

2.3.1 Storage

This section briefly describes the main types of storage devices that can be found in the market.

Disk systems

By being contained within a single “box”, a disk system usually has a central control unit that manages all the I/O, simplifying the integration of the system with other devices, such as other disk systems or servers.

We introduced you to what a Storage System consists of in the previous chapter. Depending on the specific functionality offered by a particular storage system, it is possible to make it behave as a small/mid-size or enterprise solution; the decision as to which type of disk system is more suitable for a SAN implementation strongly depends on the performance capacity and availability requirements for the particular SAN. We describe the IBM product portfolio later in this book.

Tape systems

Tape systems, in much the same way as disk systems do, are devices that comprise all the necessary apparatus to manage the use of tapes for storage purposes. In this case, however, the serial nature of a tape makes it impossible for them to be treated in parallel, as RAID devices are leading to a somewhat simpler architecture to manage and use.

There are basically three types of systems: drives, autoloaders and libraries, that are described as follows.

Tape drives

As with disk drives, tape drives are the means by which tapes can be connected to other devices; they provide the physical and logical structure for reading from, and writing to tapes.

Tape autoloaders

Tape autoloaders are autonomous tape drives capable of managing tapes and performing automatic back-up operations. They are usually connected to high-throughput devices that require constant data back-up.

Tape libraries

Tape libraries are devices capable of managing multiple tapes simultaneously and, as such, can be viewed as a set of independent tape drives or autoloaders. They are usually deployed in systems that require massive storage capacity, or that need some kind of data separation that would result in multiple single-tape systems. As a tape is not a random-access media, tape libraries cannot provide parallel access to multiple tapes as a way to improve performance, but they can provide redundancy as a way to improve data availability and fault-tolerance.

Once more, the circumstances under which each of these systems, or even a disk system, should be used, strongly depend on the specific requirements that a particular SAN implementation has. However, we can say that disk systems are usually used for online storage due to their superior performance, whereas tape systems are ideal for offline, high-throughput storage, due to the lower cost of storage per byte.

In the next section we describe the prevalent connectivity interfaces, protocols and services for building a SAN.

2.3.2 SAN connectivity

SAN connectivity comprises of hardware and software components that make possible the interconnection of storage devices and servers. Now we are going to introduce you to the Fibre Channel model for Storage Area Networks.

Standards and models for storage connectivity

As we mentioned previously, networking is governed by adherence to standards and models, and data transfer is also governed by standards. By far the most common is SCSI.

SCSI is an acronym for Small Computer Systems Interface. It is an ANSI standard that has become one of the leading I/O buses in the computer industry.

An industry effort was started to create a stricter standard allowing devices from different vendors to work together. This effort was recognized in the ANSI SCSI-1 standard. The SCSI-1 standard (circa 1985) is rapidly becoming obsolete. The current standard is SCSI-2, with SCSI-3 on the drawing boards.

The SCSI bus is a parallel bus, which comes in a number of variants as shown in Figure 2-2.

Note: If you are not familiar with parallel and serial data transfer refer to Chapter 3, “Fibre Channel internals” on page 31 for more information.

SCSI Standard	Cable Length	Speed (MBps)	Devices Supported
SCSI-1	6	5	8
SCSI-2	6	5 to 10	8 or 16
Fast SCSI-2	3	10 to 20	8
Wide SCSI-2	3	20	16
Fast Wide SCSI-2	3	20	16
Ultra SCSI-3,8-bit	1.5	20	8
Ultra SCSI-3,16-bit	1.5	40	16
Ultra-2 SCSI	12	40	8
Wide Ultra-2 SCSI	12	80	16
Ultra-3 (Ultra160/m)	12	160	16

Figure 2-2 SCSI Standards comparison table

In addition to a physical interconnection standard, SCSI defines a logical (command set) standard to which disk devices must adhere. This standard is called the Common Command Set (CCS) and was developed more or less in parallel with ANSI SCSI-1.

Of course the SCSI bus not only has data lines, but also a number of control signals. A very elaborate protocol is part of the standard to allow multiple devices to share the bus in an efficient manner.

In SCSI-3 even faster bus types are introduced, along with serial SCSI buses that reduce the cabling overhead and allows a higher maximum bus length and it is at this point where the Fibre Channel model comes in.

As always, market demands and needs were pushing for new technologies, especially for faster communications with no distance, or number of connected devices, limitations.

Fibre Channel (FC) is a serial interface (primarily implemented with fiber-optic cable, and is the primary architecture for the vast majority of SANs. To support this there are many vendors in the marketplace producing Fibre Channel adapters, and other FC devices. Fibre Channel brought these advantages by introducing a new protocol stack and by keeping the SCSI-3 CCS on top of it.

Figure 2-3 on page 22 shows the evolution of Fibre Channel speeds, and Fibre Channel will be covered in greater depth later in this book.

NAME	Throughput (Fullduplex) (MBps)	Availability
1GFC	200	1997
2GFC	400	2001
4GFC	800	2005
8GFC	1600	2008
10GFC Serial	2550	2004
16GFC	3200	2011

Figure 2-3 Fibre Channel evolution

Figure 2-4 on page 22 shows an overview of the FC model, and in it we can see that it is divided into 4 lower layers (FC-0, FC-1, FC-2, FC-3) and one upper layer (FC-4) where the upper level protocols are used such as SCSI-3, IP or FICON.

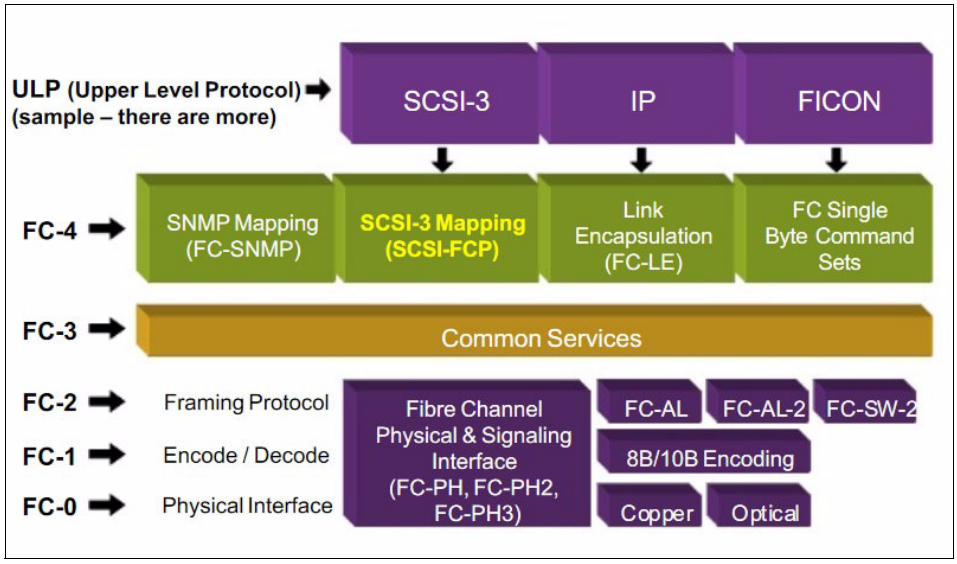

Figure 2-4 Fibre Channel Model Overview

Options for storage connectivity

In this section, we have divided these components into three sections according to the abstraction level to which they belong: lower level layers, middle level layers, and higher level layers. Figure 2-5 gives you an idea of each networking stack.

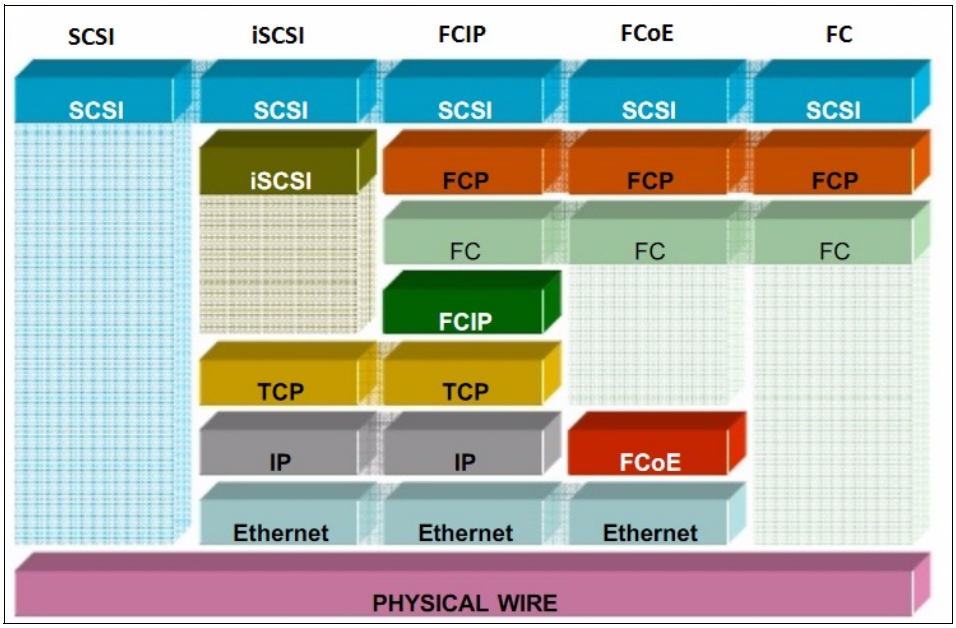

Figure 2-5 Networking stack comparison

Lower level layers

As you can see there are only three stacks that can directly interact with the physical wire, those are Ethernet, SCSI and Fibre Channel. Because of that, these models are considered as lower level layers, all the others are combinations of them such as iSCSI, FCIP and FCoE also called middle level layers.

At this point we are assuming a basic knowledge of Ethernet which is typically used on conventional server-to-server or workstation-to-server network connections. They build up a common bus topology by which every attached device can communicate with each other using this common-bus. Ethernet speed is increasing as it becomes more pervasive in the data center. We will describe key concepts of Ethernet later in this book.

Middle level layers

This section comprises the transport protocol and session layers.

Note: Fibre Channel over Ethernet (FCoE) will be described later in this book as it is a vital model for CNA.

iSCSI

Internet SCSI (iSCSI) is a transport protocol that carries SCSI commands from an initiator to a target. It is a data storage networking protocol that transports standard Small Computer System Interface (SCSI) requests over the standard Transmission Control Protocol/Internet Protocol (TCP/IP) networking technology.

iSCSI enables the implementation of IP-based storage area networks (SANs), enabling customers to use the same networking technologies — for both storage and data networks. As it uses TCP/IP, iSCSI is also well suited to run over almost any physical network. By eliminating the need for a second network technology just for storage, iSCSI has the potential to lower the costs of deploying networked storage.

FCP

The Fibre Channel Protocol (FCP) is the interface protocol of SCSI on Fibre Channel. It is a gigabit speed network technology primarily used for Storage Networking. Fibre Channel is standardized in the T11 Technical Committee of the InterNational Committee for Information Technology Standards (INCITS), an American National Standard Institute (ANSI) accredited standards committee. It started for use primarily in the supercomputer field, but has become the standard connection type for storage area networks in enterprise storage. Despite its name, Fibre Channel signaling can run on both twisted-pair copper wire and fiber optic cables.

FCIP

Fibre Channel over IP (FCIP) is also known as Fibre Channel tunneling or storage tunneling. It is a method to allow the transmission of Fibre Channel information to be tunnelled through the IP network. Because most organizations already have an existing IP infrastructure, the attraction of being able to link geographically dispersed SANs, at a relatively low cost, is enormous.

FCIP encapsulates Fibre Channel block data and subsequently transports it over a TCP socket. TCP/IP services are utilized to establish connectivity between remote SANs. Any congestion control and management, as well as data error and data loss recovery, is handled by TCP/IP services, and does not affect FC fabric services.

The major point with FCIP is that is does not replace FC with IP, it simply allows deployments of FC fabrics using IP tunnelling. The assumption that this might lead to is that the “industry” has decided that FC-based SANs are more than appropriate, and that the only need for the IP connection is to facilitate any distance requirement that is beyond the current scope of an FCP SAN.

FICON

FICON architecture is an enhancement of, rather than a replacement for, the now relatively old ESCON® architecture. As a SAN is Fibre Channel based, FICON is a prerequisite for z/OS systems to fully participate in a heterogeneous SAN, where the SAN switch devices allow the mixture of open systems and mainframe traffic.

FICON is a protocol that uses Fibre Channel as its physical medium. FICON channels are capable of data rates up to 200 MBps full duplex, they extend the channel distance (up to 100 km), increase the number of control unit images per link, increase the number of device addresses per control unit link, and retain the topology and switch management characteristics of ESCON.

Higher level layers

This section comprises of the presentation and application layers.

Server-attached storage

The earliest approach was to tightly couple the storage device with the server. This server-attached storage approach keeps performance overhead to a minimum. Storage is attached directly to the server bus using an adapter card, and the storage device is dedicated to a single server. The server itself controls the I/O to the device, issues the low-level device commands, and monitors device responses.

Initially, disk and tape storage devices had no on-board intelligence. They just executed the server’s I/O requests. Subsequent evolution led to the introduction of control units. Control units are storage off-load servers that contain a limited level of intelligence, and are able to perform functions, such as I/O request caching for performance improvements, or dual copy

of data (RAID 1) for availability. Many advanced storage functions have been developed and implemented inside the control unit.

Network Attached Storage

Network Attached Storage (NAS) is basically a LAN-attached file server that serves files using a network protocol such as Network File System (NFS). NAS is a term used to refer to storage elements that connect to a network and provide file access services to computer systems. A NAS storage element consists of an engine that implements the file services (using access protocols such as NFS or CIFS), and one or more devices, on which data is stored. NAS elements may be attached to any type of network. From a SAN perspective, a SAN-attached NAS engine is treated just like any other server, but a NAS does not provide any of the activities that a server in a server-centric system typically provides, such as e-mail, authentication, or file management.

NAS allows more hard disk storage space to be added to a network that already utilizes servers without shutting them down for maintenance and upgrades. With a NAS device, storage is not an integral part of the server. Instead, in this storage-centric design, the server still handles all of the processing of data, but a NAS device delivers the data to the user. A NAS device does not need to be located within the server but can exist anywhere in the LAN and can be made up of multiple networked NAS devices. These units communicate to a host using Ethernet and file-based protocols. This is in contrast to the disk units discussed earlier, which use Fibre Channel protocol and block-based protocols to communicate.

NAS storage provides acceptable performance and security, and it is often less expensive for servers to implement (for example, ethernet adapters are less expensive than Fibre Channel adapters).

In an effort to bridge the two worlds and to open up new configuration options for customers, some vendors, including IBM, sell NAS units that act as a gateway between IP-based users and SAN-attached storage. This allows for the connection of the storage device and share it between your high-performance database servers (attached directly through Fibre Channel) and your end users (attached through IP) who do not have performance requirements nearly as strict.

NAS is an ideal solution for serving files stored on the SAN to end users in cases where it would be impractical and expensive to equip end users with Fibre Channel adapters. NAS allows those users to access your storage through the IP-based network that they already have.

2.3.3 Servers

Each of the different server platforms (IBM System Z, UNIX, AIX®, HPUX, Sun Solaris, Linux, and others), IBM i, and Windows Server have implemented SAN solutions using various interconnects and storage technologies. The following sections review these solutions and the different implementations on each of the platforms.

Mainframe servers

In simple terms, a mainframe is a single, monolithic and possibly multi-processor high-performance computer system. Apart from the fact that IT evolution has been pointing toward a more distributed and loosely coupled infrastructure, mainframes still play an important role on businesses that depend on massive storage capabilities.

The IBM System Z is a processor(s) and operating system mainframe set. Historically, System Z servers have supported many different operating systems, such as z/OS, OS/390®, VM, VSE, and TPF, which have been enhanced over the years. The processor to storage

device interconnection has also evolved from a bus and tag interface to ESCON channels, and now to FICON channels. Figure 2-6 on page 26 shows the various processor-to-storage interfaces.

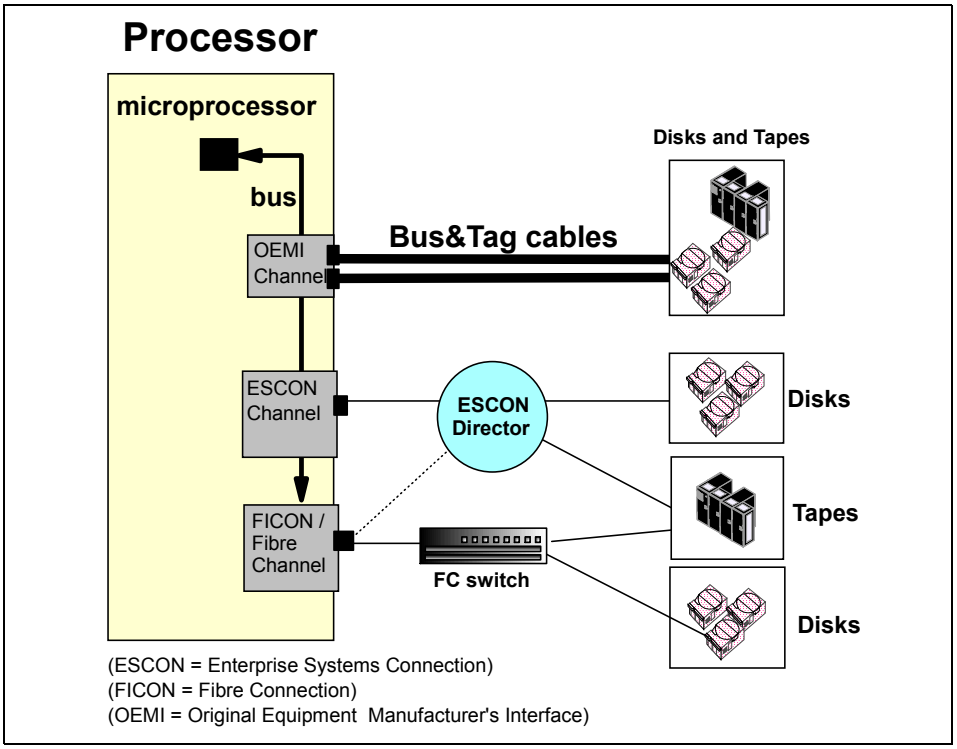

Figure 2-6 Processor-to-storage interface connections

Due to architectural differences, and extremely strict data integrity and management requirements, the implementation of FICON has been somewhat behind that of FCP on open systems. However, at the time of writing, FICON has now caught up with FCP SANs, and they coexist quite amicably.

For the latest news on zSeries® FICON connectivity, refer to:

http://www-03.ibm.com/systems/z/hardware/connectivity/index.html

In addition to FICON for traditional zSeries operating systems, IBM has standard Fibre Channel adapters for use with zSeries servers that can implement Linux.

UNIX-based servers

Originally designed for high-performance computer systems, such as mainframes, the UNIX operating systems is today present on a great variety of hardware platforms, ranging from Linux-based PCs to dedicated large-scale stations. Due to its popularity and maturity, it also plays an important role on both existing and legacy IT infrastructures.

The IBM System p® line of servers, running a UNIX operating system called AIX, offers various processor to storage interfaces, including SCSI, SSA (Serial Storage Architecture), and Fibre Channel. The SSA interconnection has primarily been used for disk storage. Fibre Channel adapters are able to connect to tape and disk. Figure 2-7 shows the various processor-to-storage interconnect options for the System p family.

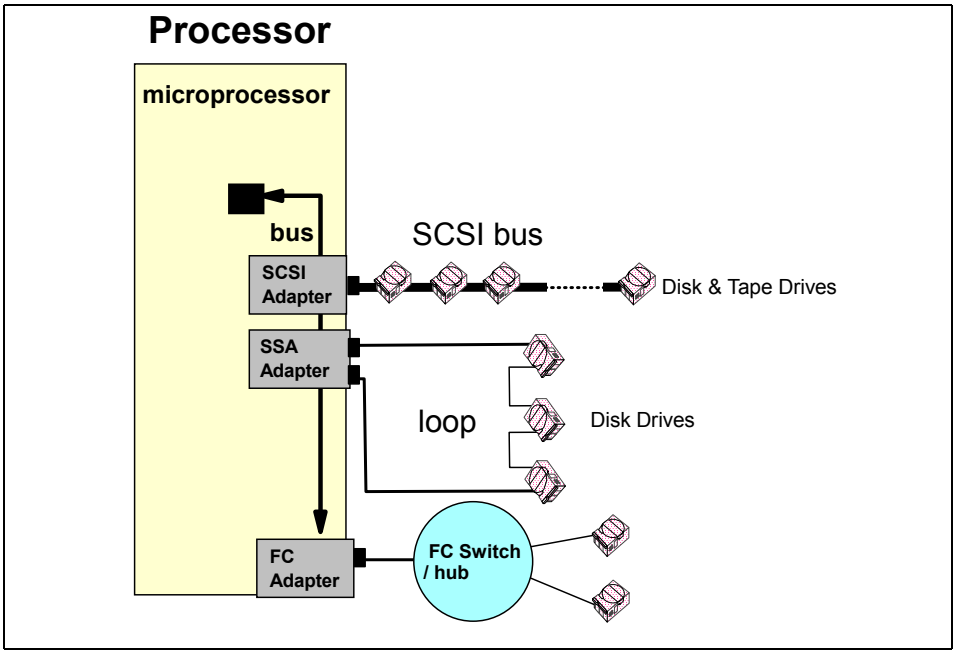

Figure 2-7 System p processor-to-storage interconnections

The various UNIX system vendors in the market deploy different variants of the UNIX operating system, each having some unique enhancements, and often supporting different file systems such as the Journal File System (JFS), Enhanced Journal File System (JFS2), and the Andrew File System (AFS®). The server-to-storage interconnect is similar to System p as shown in Figure 2-7.

For the latest System p Power Systems™ products, refer to:

http://www.ibm.com/systems/storage/product/power.html

Windows-based servers

Based on the reports of various analysts regarding growth in the Windows server market (both in the number and size of Windows servers), Windows will become the largest market for SAN solution deployment. More and more Windows servers will host mission-critical applications that will benefit from SAN solutions, such as disk and tape pooling, tape sharing, multipathing, and remote copy.

The processor-to-storage interfaces on System x® servers (IBM Intel-based processors that support the Microsoft Windows Server operating system) are similar to those supported on UNIX servers, including SCSI and Fibre Channel.

For more information, see the System x SAN Web site at:

http://www.ibm.com/systems/storage/product/x.html

Single-level storage

Single-level storage (SLS) is probably the most significant differentiator in a SAN solution implementation on a System i® server, as compared to other systems such as z/OS, UNIX, and Windows. In IBM i, both the main storage (memory) and the secondary storage (disks) are treated as a very large virtual address space known as SLS.

Figure 2-8 compares the IBM i SLS addressing with the way Windows or UNIX systems work, using the processor local storage. With 32-bit addressing, each process (job) has 4 GB of

addressable memory. With 64-bit SLS addressing, over 18 million terabytes (18 exabytes) of addressable storage is possible. Because a single page table maps all virtual addresses to physical addresses, task switching is very efficient. SLS further eliminates the need for address translation, thus speeding up data access.

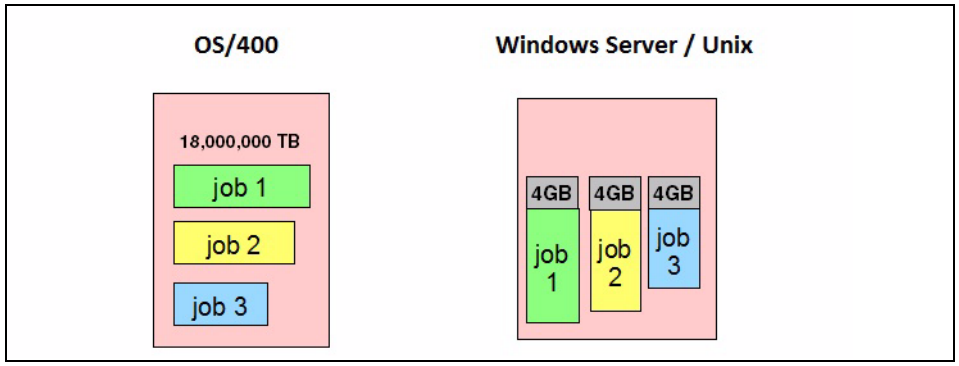

Figure 2-8 IBM i versus Windows Server 32 bits or UNIX storage addressing

System i SAN support has rapidly expanded, and System i servers now support attachment to switched fabrics, and to most of IBM SAN-attached storage products.

For more information, see the System i SAN Web site:

http://www.ibm.com/systems/i/hardware/storage/

2.3.4 Putting the components together

After going through this myriad of technologies and platforms, we can easily understand why it is a challenge to implement true heterogeneous storage and data environments across different hardware and operating systems platforms; for example, disk and tape sharing across z/OS, IBM i, UNIX, and Windows Server.

One of the SAN principles, which is infrastructure simplification, cannot be easily achieved; each platform, along with its operating system, treats data differently at various levels in the system architecture, thus creating some of these many challenges:

- Different attachment interfaces and protocols, such as SCSI, ESCON and FICON.
- Different data formats, such as Extended Count Key Data (ECKD™), blocks, clusters, and sectors.
- Different file systems, such as Virtual Storage Access Method (VSAM), Journal File System (JFS), Enhanced Journal File System (JFS2), Andrew File System (AFS), and Windows Server File System (NTFS).
- IBM i, with the concept of single-level storage.
- Different file system structures, such as catalogs and directories.
- Different file naming conventions, such as AAA.BBB.CCC and DIR/Xxx/Yyy.
- Different data encoding techniques, such as EBCDIC, ASCII, floating point, and little or big endian.

In Figure 2-9 on page 29 is a brief summary of these differences for several different systems.

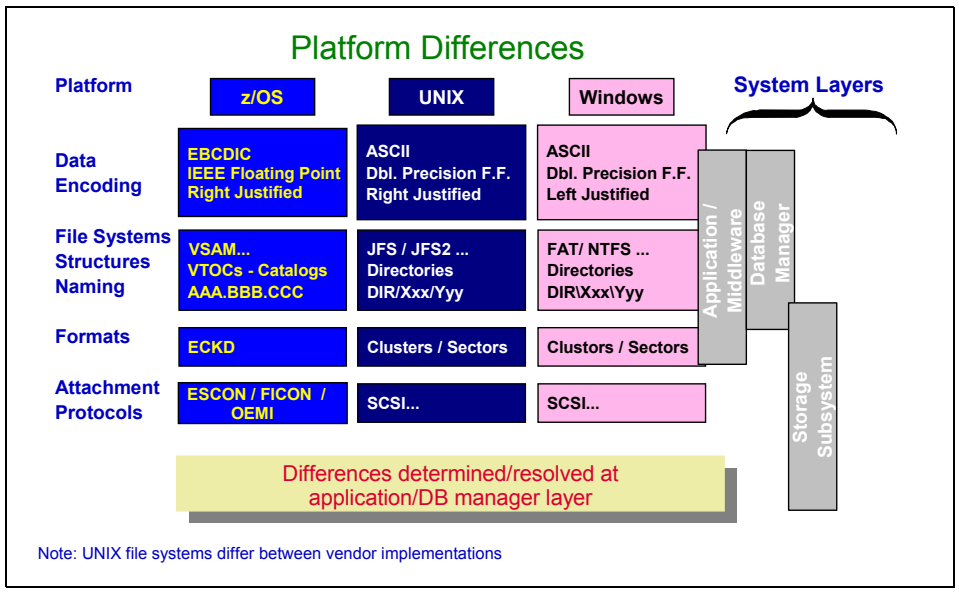

Figure 2-9 Hardware and operating systems differences

3

Fibre Channel internals

Fibre Channel is the predominant architecture upon which SAN implementations are built. Fibre Channel is a technology standard that allows data to be transferred at extremely high speeds. Current implementations support data transfers at up to 16 Gbps or even more. The Fibre Channel standard is accredited by many standards bodies, technical associations, vendors, and industry-wide consortiums. There are many products on the market that take advantage of FC's high-speed, high-availability characteristics.

Fibre Channel was completely developed through industry cooperation, unlike SCSI, which was developed by a vendor, and submitted for standardization afterwards.

Fibre or Fiber?: Fibre Channel was originally designed to support fiber optic cabling only. When copper support was added, the committee decided to keep the name in principle, but to use the UK English spelling (Fibre) when referring to the standard. We retain the US English spelling when referring generically to fiber optics and cabling.

Some people refer to Fibre Channel architecture as the Fibre version of SCSI. Fibre Channel is an architecture used to carry IPI traffic, IP traffic, FICON traffic, FCP (SCSI) traffic, and possibly traffic using other protocols, all on the standard FC transport. An analogy could be Ethernet, where IP, NetBIOS, and SNA are all used simultaneously over a single Ethernet adapter, since these are all protocols with mappings to Ethernet. Similarly, there are many protocols mapped onto FC.

FICON is the standard protocol for z/OS, and will replace all ESCON environments over time. FCP is the standard protocol for open systems, both using Fibre Channel architecture to carry the traffic.

3.1 Firstly, why the Fibre Channel architecture?

Before we delve into the internals of Fibre Channel we will describe why Fibre Channel became the predominant SAN architecture.

3.1.1 The SCSI legacy

The Small Computer Systems Interface (SCSI) is the conventional, server centric method of connecting peripheral devices (disks, tapes and printers) in the open client/server environment. As its name indicates, it was designed for the PC and small computer environment. It is a bus architecture, with dedicated, parallel cabling between the host and storage devices, such as disk arrays. This is similar in implementation to the Original Equipment Manufacturer's Information (OEMI) bus and tag interface commonly used by mainframe computers until the early 1990's.

In addition to being a physical transport, SCSI is also a protocol, which specifies commands and controls for sending blocks of data between the host and the attached devices. SCSI commands are issued by the host operating system, in response to user requests for data. Some operating systems, for example, Windows NT, treat all attached peripherals as SCSI devices, and issue SCSI commands to deal with all read and write operations.SCSI was used in Direct attached Storage with internal and external devices connected VIA SCSI channel in Daisy chain fashion.

Typical SCSI device connectivity is shown in Figure 3-1 on page 32.

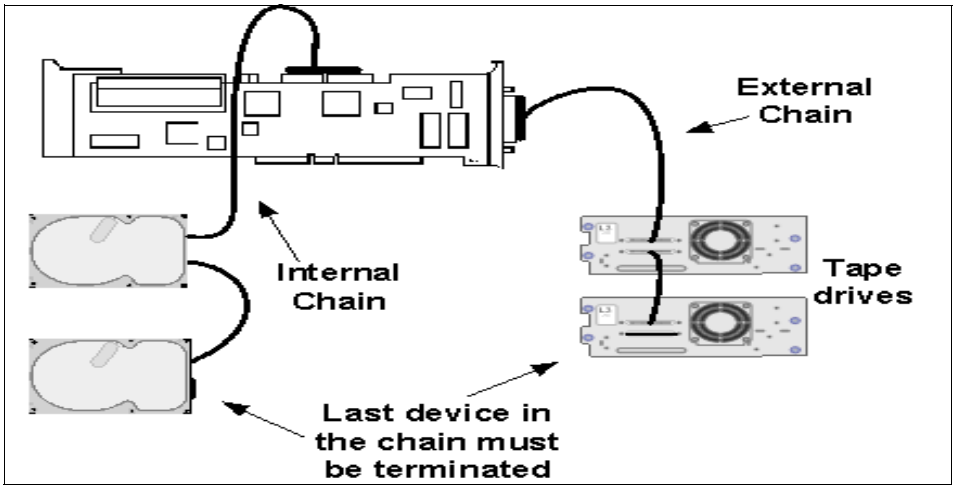

Figure 3-1 SCSI Device connectivity

3.1.2 Limitations of SCSI

We list some of the limitations of SCSI in the topics that follow.

Scalability limitations

The amount of data available to the server is determined by the number of devices which can attach to the bus, and by the number of buses attached to the server. Up to 15 devices can be attached to a server on a single SCSI bus. In practice, because of performance limitations due to arbitration, it is common for no more than four or five devices to be attached in this way, thus limiting scalability in terms of the number of devices able to be connected to the server.

Reliability and availability limitations

SCSI shares aspects with bus and tag in that cables and connectors are bulky, relatively expensive, and are prone to failure. Access to data is lost in the event of a failure of any of the SCSI connections to the disks. This also applies in the event of reconfiguration or servicing of a disk device attached to the SCSI bus, because all the devices in the string must be taken

offline. In today's environment, when many applications need to be available continuously, this downtime is unacceptable.

Speed and latency limitations

The data rate of the SCSI bus is determined by the number of bits transferred, and the bus cycle time (measured in megahertz (MHz)). Decreasing the cycle time increases the transfer rate, but, due to limitations inherent in the bus architecture, it may also reduce the distance over which the data can be successfully transferred. The physical transport was originally a parallel cable comprising eight data lines, to transmit eight bits in parallel, plus control lines. Later implementations widened the parallel data transfers to 16 bit paths (SCSI Wide), to achieve higher bandwidths.

Propagation delays in sending data in parallel along multiple lines lead to a well known phenomenon known as skew, meaning that all bits may not arrive at the target device at the same time. This is shown in Figure 3-2.

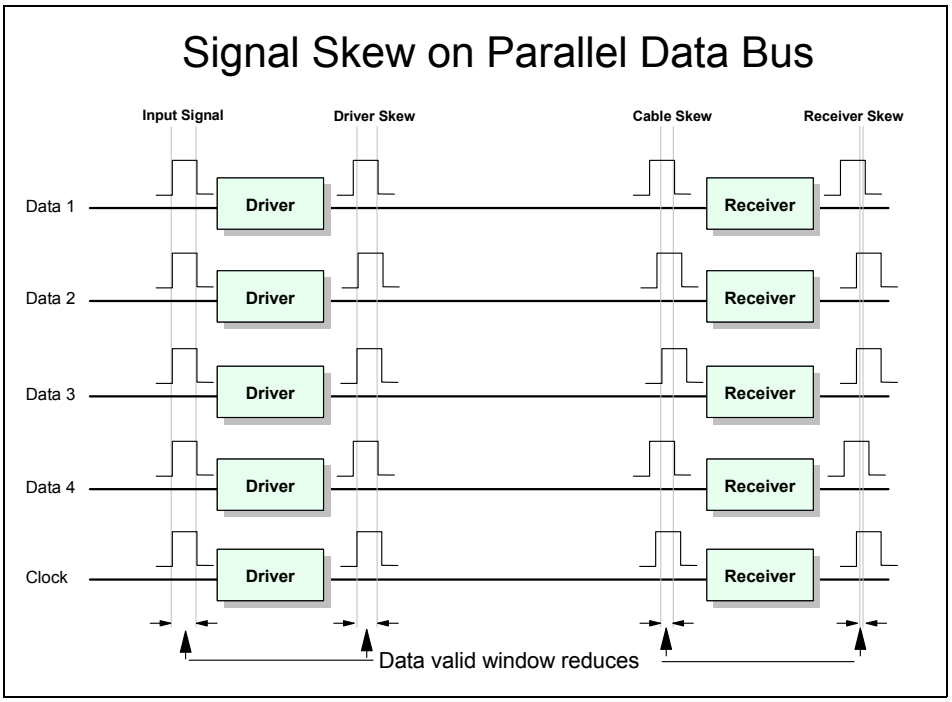

Figure 3-2 SCSI Propagation delay results in skew

Arrival occurs during a small window of time, depending on the transmission speed, and the physical length of the SCSI bus. The need to minimize the skew limits the distance that devices can be positioned away from the initiating server to between 2 to 25 meters, depending on the cycle time. Faster speed means shorter distance.

Distance limitations

The distances refer to the maximum length of the SCSI bus, including all attached devices. The SCSI distance limitations are shown in Figure 3-3 on page 34. These distance limitations may severely restrict the total GB capacity of the disk storage which can be attached to an individual server.

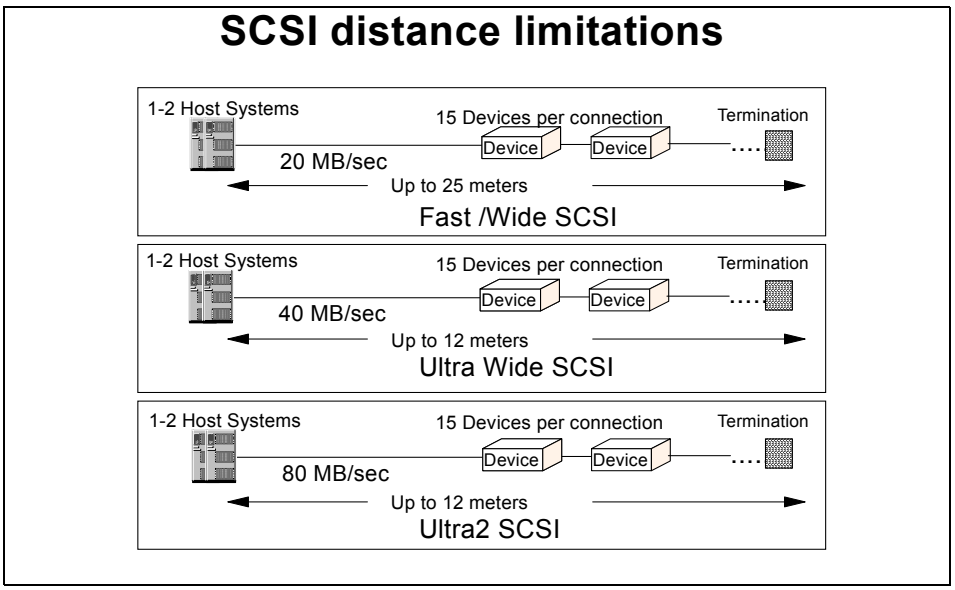

Figure 3-3 SCSI bus distance limitations

Device sharing

Many applications require the system to access several devices, or for several systems to share a single device. SCSI can enable this by attaching multiple servers or devices to the same bus. This is known as a multi-drop configuration. A multi-drop configuration is shown in Figure 3-4.

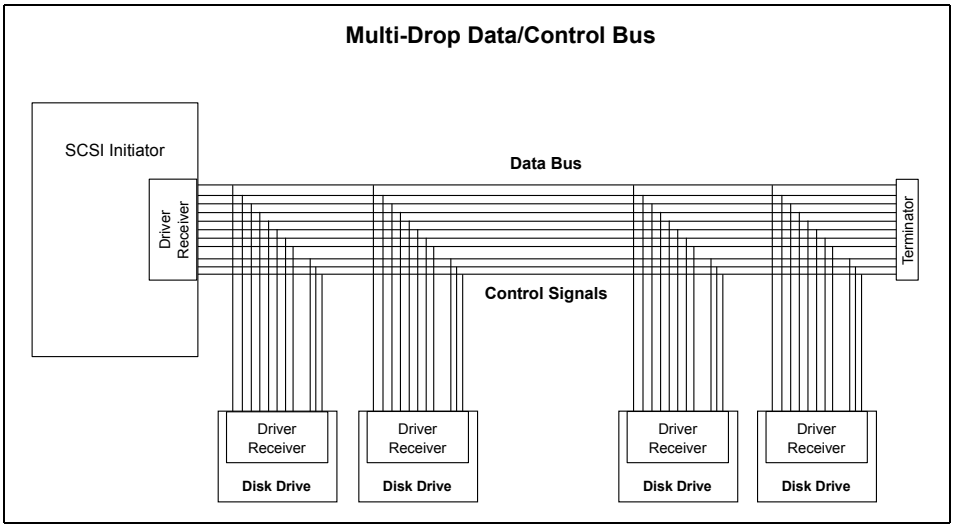

Figure 3-4 Multi-drop bus structure

To avoid signal interference, and therefore possible data corruption, all unused ports on a parallel SCSI bus must be properly terminated. Incorrect termination can result in transaction errors or failures.

Normally, only a single server can access data on a specific disk by means of a SCSI bus. In a shared bus environment, it is clear that all devices cannot transfer data at the same time. SCSI uses an arbitration protocol to determine which device can gain access to the bus. Arbitration occurs before and after every data transfer on the bus. While arbitration takes place, no data movement can occur. This represents an additional overhead which reduces

bandwidth utilization, substantially reducing the effective data rate achievable on the bus. Actual rates are typically less than 50% of the rated speed of the SCSI bus.

It is clear that the physical parallel SCSI bus architecture has a number of significant speed, distance, and availability limitations, which make it increasingly less suitable for many applications in today's networked IT infrastructure. However, since the SCSI protocol is deeply embedded in the way that commonly encountered operating systems handle user requests for data, it would be a major inhibitor to progress if we were obliged to move to new protocols.

3.1.3 Why Fibre Channel?

Fibre Channel is an open, technical standard for networking which incorporates the "channel transport" characteristics of an I/O bus, with the flexible connectivity and distance characteristics of a traditional network.

Because of its channel-like qualities, hosts and applications see storage devices attached to the SAN as though they are locally attached storage. Because of its network characteristics it can support multiple protocols and a broad range of devices, and it can be managed as a network. Fibre Channel can use either optical fiber (for distance) or copper cable links (for short distance at low cost).

Fibre Channel is a multi-layered network, based on a series of American National Standards Institute (ANSI) standards which define characteristics and functions for moving data across the network. These include definitions of physical interfaces, such as cabling, distances and signaling; data encoding and link controls; data delivery in terms of frames, flow control and classes of service; common services; and protocol interfaces.

Like other networks, information is sent in structured packets or frames, and data is serialized before transmission. But, unlike other networks, the Fibre Channel architecture includes a significant amount of hardware processing to deliver high performance.

Fibre Channel uses a serial data transport scheme, similar to other computer networks, streaming packets, (frames) of bits one behind the other in a single data line to achieve high data rates.

Serial transfer by its very nature, of course, does not suffer from the problem of skew, so speed and distance is not restricted as with parallel data transfers as we show in Figure 3-5.

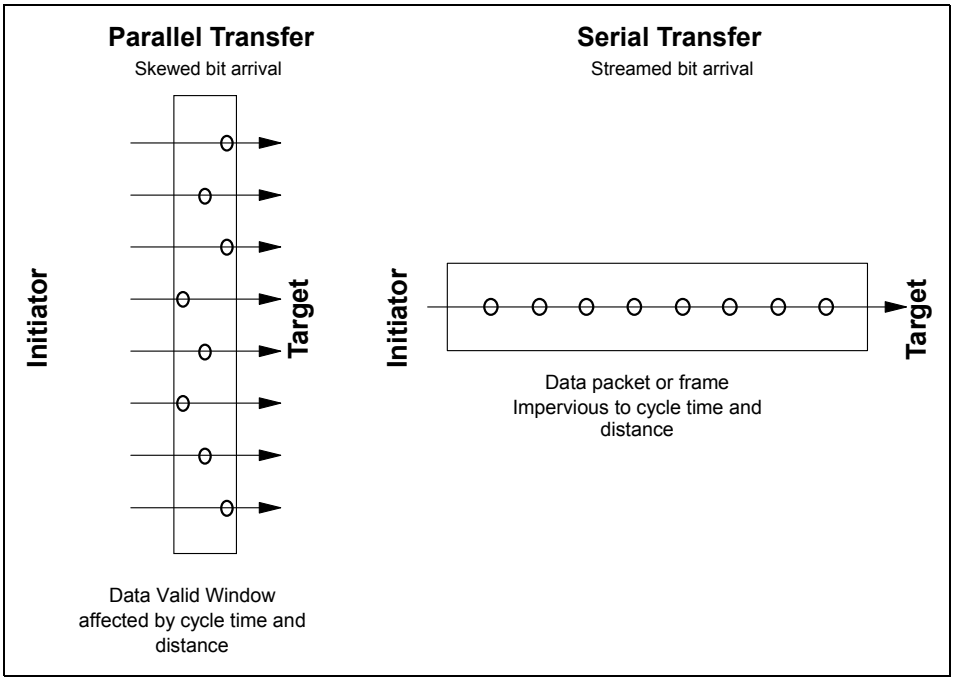

Figure 3-5 Parallel data transfers versus serial data transfers

Serial transfer enables simpler cabling and connectors, and also routing of information through switched networks. Fibre Channel can operate over longer distances, both natively and by implementing cascading, and longer with the introduction of repeaters. Just as LANs can be interlinked in WANs by using high speed gateways, so can campus SANs be interlinked to build enterprise wide SANs.

Whatever the topology, information is sent between two nodes, which are the source (transmitter or initiator) and destination (receiver or target). A node is a device, such as a server (personal computer, workstation, or mainframe), or peripheral device, such as disk or tape drive, or video camera. Frames of information are passed between nodes, and the structure of the frame is defined by a protocol. Logically, a source and target node must utilize the same protocol, but each node may support several different protocols or data types.

Therefore, Fibre Channel architecture is extremely flexible in its potential application. Fibre Channel transport layers are protocol independent, enabling the transmission of multiple protocols.

Using a credit based flow control methodology, Fibre Channel is able to deliver data as fast as the destination device buffer is able to receive it. And low transmission overheads enable high sustained utilization rates without loss of data.

Therefore, Fibre Channel combines the best characteristics of traditional I/O channels with those of computer networks:

- High performance for large data transfers by using simple transport protocols and extensive hardware assists
- Serial data transmission
- A physical interface with a low error rate definition
- Reliable transmission of data with the ability to guarantee or confirm error free delivery of the data
- Packaging data in packets (frames in Fibre Channel terminology)

- Flexibility in terms of the types of information which can be transported in frames (such as data, video and audio)
- Use of existing device oriented command sets, such as SCSI and FCP
- A vast expansion in the number of devices which can be addressed when compared to I/O interfaces — a theoretical maximum of more than 15 million ports

It is this high degree of flexibility, availability and scalability; the combination of multiple protocols at high speeds over long distances; and the broad acceptance of the Fibre Channel standards by vendors throughout the IT industry, which makes the Fibre Channel architecture ideal for the development of enterprise SANs.

In the topics that follow we describe some of the key concepts that we have touched upon in the previous pages and that are behind Fibre Channel SAN implementations. We also introduce some more Fibre Channel SAN terminology and jargon that the reader can expect to encounter.

3.2 Layers

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

The five layers can be categorized into these two:

- Physical and signaling layer
- Upper layer

Fibre Channel is a layered protocol. as shown in Figure 3-6.

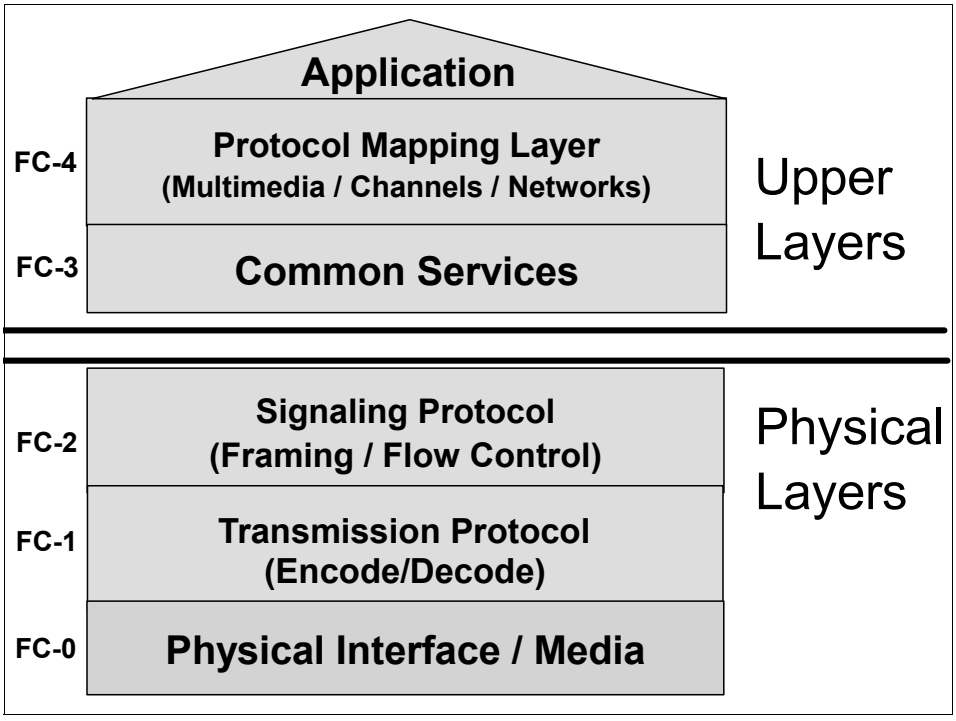

Figure 3-6 Upper and physical layers

The layers can be briefly described as follows:

Physical and signaling layers

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

Physical interface and media: FC-0

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

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

Laser safety

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

Transmission protocol: FC-1

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

Framing and signaling protocol: FC-2

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

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

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

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

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

Upper layers

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

Common services: FC-3

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

- Hunt Groups
  - A *Hunt Group* is a set of associated N_Ports attached to a single node. This set is assigned an alias identifier that allows any frames containing the alias to be routed to any available N_Port within the set. This decreases latency in waiting for an N_Port to become available.
- Striping
  - *Striping* is used to multiply bandwidth, using multiple N_Ports in parallel to transmit a single information unit across multiple links.
- Multicast
  - *Multicast* delivers a single transmission to multiple destination ports. This includes the ability to broadcast to all nodes or a subset of nodes.

Upper layer protocol mapping (ULP): FC-4

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

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

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

3.3 Optical cables

An optical fiber is a very thin strand of silica glass and in its geometry quite like a human hair. In reality it is a very narrow, very long glass cylinder with special characteristics. When light enters one end of the fiber it travels (confined within the fiber) until it leaves the fiber at the other end. Two critical factors stand out:

- Very little light is lost in its journey along the fiber.
- Fiber can bend around corners and the light will stay within it and be guided around the corners.

An optical fiber consists of two parts: the core and the cladding. See Figure 3-7 on page 42. The core is a narrow cylindrical strand of glass and the cladding is a tubular jacket surrounding it. The core has a (slightly) higher refractive index than the cladding. This means that the boundary (interface) between the core and the cladding acts as a perfect mirror. Light travelling along the core is confined by the mirror to stay within it — even when the fiber bends around a corner.

When light is transmitted on a fiber, the most important consideration is “what kind of light?” The electromagnetic radiation that we call light exists at many wavelengths. These wavelengths go from invisible infrared through all the colors of the visible spectrum to invisible ultraviolet. Because of the attenuation characteristics of fiber, we are only interested in infrared “light” for communication applications. This light is usually invisible, since the wavelengths used are usually longer than the visible limit of around 750 nanometers (nm).

If a short pulse of light from a source such as a laser or an LED is sent down a narrow fiber, it will be changed (degraded) by its passage down the fiber. It will emerge (depending on the

distance) much weaker, lengthened in time (“smeared out”), and distorted in other ways. The reasons for this are described in the topics that folllow.

3.3.1 Attenuation

The pulse will be weaker because all glass absorbs light. More accurately, impurities in the glass can absorb light but the glass itself does not absorb light at the wavelengths of interest. In addition, variations in the uniformity of the glass cause scattering of the light. Both the rate of light absorption and the amount of scattering are dependent on the wavelength of the light and the characteristics of the particular glass. Most light loss in a modern fiber is caused by scattering.

3.3.2 Maximum power

There is a practical limit to the amount of power that can be sent on a fiber. This is about half a watt (in standard single-mode fiber) and is due to a number of non-linear effects that are caused by the intense electromagnetic field in the core when high power is present.

Polarization

Conventional communication optical fiber is cylindrically symmetric but contains imperfections. Light travelling down such a fiber is changed in polarization. (In current optical communication systems this does not matter but in future systems it may become a critical issue.)

Dispersion

Dispersion occurs when a pulse of light is spread out during transmission on the fiber. A short pulse becomes longer and ultimately joins with the pulse behind, making recovery of a reliable bit stream impossible. (In most communications systems bits of information are sent as pulses of light. 1 = light, 0 = dark. But even in analogue transmission systems where information is sent as a continuous series of changes in the signal, dispersion causes distortion.) There are many kinds of dispersion, each of which works in a different way, but the most important three are discussed below:

Material dispersion (chromatic dispersion)

Both lasers and LEDs produce a range of optical wavelengths (a band of light) rather than a single narrow wavelength. The fiber has different refractive index characteristics at different wavelengths and therefore each wavelength will travel at a different speed in the fiber. Thus, some wavelengths arrive before others and a signal pulse disperses (or smears out).

Modal dispersion

When using multimode fiber, the light is able to take many different paths or “modes” as it travels within the fiber. The distance traveled by light in each mode is different from the distance travelled in other modes. When a pulse is sent, parts of that pulse (rays or quanta) take many different modes (usually all available modes). Therefore, some components of the pulse will arrive before others. The difference between the arrival time of light taking the fastest mode versus the slowest obviously gets greater as the distance gets greater.

Waveguide dispersion

Waveguide dispersion is a very complex effect and is caused by the shape and index profile of the fiber core. However, this can be controlled by careful design and, in fact, waveguide dispersion can be used to counteract material dispersion.

Noise

One of the great benefits of fiber optical communications is that the fiber doesn't pick up noise from outside the system. However, there are various kinds of noise that can come from components within the system itself. Mode partition noise can be a problem in single-mode fiber and modal noise is a phenomenon in multimode fibre.

It is not our intention to delve any deeper into optical than this in the book.

3.3.3 Fiber in the SAN

Fibre Channel can be run over optical or copper media, but fiber-optic cables enjoy a major advantage in noise immunity as we mentioned previously. It is for this reason that fiber-optic cabling is preferred. However, copper is also used, and it is likely that in the short term a mixed environment will need to be tolerated and supported although this is less likely to be the case as SANs mature.

In addition to the noise immunity, fiber-optic cabling provides a number of distinct advantages over copper transmission lines that make it a very attractive medium for many applications. At the forefront of the advantages are:

- Greater distance capability than is generally possible with copper
- Insensitive to induced electro-magnetic interference (EMI)
- No emitted electro-magnetic radiation (RFI)
- No electrical connection between two ports
- Not susceptible to crosstalk
- Compact and lightweight cables and connectors

However, fiber-optic and optical links do have some drawbacks. Some of the considerations are:

- Optical links tend to be more expensive than copper links over short distances.
- Optical connections do not lend themselves to backplane printed circuit wiring.
- Optical connections may be affected by dirt and other contamination.

Overall, optical fibers have provided a very high-performance transmission medium, which has been refined and proven over many years.

Mixing fiber-optical and copper components in the same environment is supported, although not all products provide that flexibility, and this should be taken into consideration when planning a SAN. Copper cables tend to be used for short distances, up to 30 meters, and can be identified by their DB-9, 9 pin, connector.

Normally, fiber-optic cabling is referred to by mode or the frequencies of light waves that are carried by a particular cable type. Fiber cables come in two distinct types, as shown in Figure 3-7.

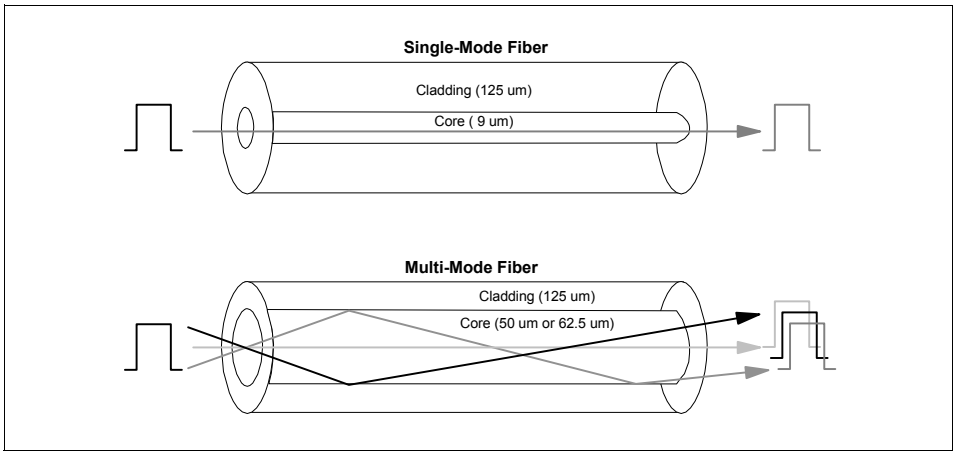

Figure 3-7 Cable types

- Multi-mode fiber (MMF) for shorter distances

  Multi-mode cabling is used with shortwave laser light and has either a 50 micron or a 62.5 micron core with a cladding of 125 micron. The 50 micron or 62.5 micron diameter is sufficiently large for injected light waves to be reflected off the core interior.

  Multi-mode fiber allows more than one mode of light. Common MM core sizes are 50 micron and 62.5 micron. 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 fibre is more economical because it can be used with inexpensive connectors and laser devices, thereby reducing the total system cost.

- Single-mode fiber (SMF) for longer distances

  Single-mode (SM) fibre allows only one pathway, or mode, of light to travel within the fibre. The core size is typically 8.3 micron. Single-mode fibres 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.

Fibre Channel architecture supports both short wave and long wave optical transmitter technologies, as follows:

- Short wave laser

  This technology uses a wavelength of 780 nanometers and is only compatible with multi-mode fiber.

- Long wave laser

  This technology uses a wavelength of 1300 nanometers. It is compatible with both single-mode and multi-mode fiber.

Different cable types along with their speed and distance are listed in Table 3-1.

Table 3-1 Fibre Channel modes, speeds, and distances

Fiber mode	Speed (MBps)	Transmitter	Medium	Distance
Single-mode fiber	1600	1310 nm longwave light	1600-SM-LC-L	0.5 m - 10 km
		1490 nm longwave light	1600-SM-LZ-I	0.5 m - 2 km
	800	1310 nm longwave light	800-SM-LC-L	2 m - 10 km
			800-SM-LC-I	2 m - 1.4 km
	400	1310 nm longwave light	400-SM-LC-L	2 m - 10 km
			400-SM-LC-M	2 m - 4 km
			400-SM-LL-I	2 m - 2 km
	200	1550 nm longwave light	200-SM-LL-V	2 m - 50 km
		1310 nm longwave light	200-SM-LC-L	2 m - 10 km
			200-SM-LL-I	2 m - 2 km
	100	1550 nm longwave light	100-SM-LL-V	2 m - 50 km
		1310 nm longwave light	100-SM-LL-L	2 m - 10 km
			100-SM-LC-L	2 m - 10 km
			100-SM-LL-I	2 m - 2 km

Fiber mode	Speed (MBps)	Transmitter	Medium	Distance
Multimode Fiber[a]	1600	850 nm shortwave light	1600-M5F-SN-I	0.5 m - 125 m
			1600-M5E-SN-I	0.5 - 100 m
			1600-M5-SN-S	0.5 - 35 m
			1600-M6-SN-S	0.5 - 15 m
	800		800-M5F-SN-I	0.5 - 190 m
			800-M5E-SN-I	0.5 - 150 m
			800-M5-SN-S	0.5 - 50 m
			800-M6-SN-S	0.5 - 21 m
	400		400-M5F-SN-I	0.5 - 400 m
			400-M5E-SN-I	0.5- 380 m
			400-M5-SN-I	0.5 - 150 m
			400-M6-SN-I	0.5- 70 m
	200		200-M5E-SN-I	0.5 - 500 m
			200-M5-SN-I	0.5 - 300 m
			200-M6-SN-I	0.5 - 150 m
	100		100-M5E-SN-I	0.5 - 860 m
			100-M5-SN-I	0.5 - 500 m
			100-M6-SN-I	0.5 - 300 m
			100-M5-SL-I	2 - 500 m
			100-M6-SL-I	2 - 175 m

a. See Table 3-2 for multimode fiber details

Table 3-2 shows multimode fiber designations, optical multimode (OM) numbering, fiber-optic cable diameters, and FC media designation.

Table 3-2 Optical multimode designations

Multimode fiber	Fiber Diameter (microns)	FC media designation
OM1	62.5 µm	M6
OM2	50 µm	M5
OM3	50 µm	M5E
OM4	50 µm	M5F

3.3.4 Dark fiber

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

If the two units which need to be connected are in different cities, then the problem is much larger. Larger, in this case, is typically associated with more expensive. As most businesses are not in the cable laying business they will lease fiber optic cables to meet their needs. When a company does this, the fiber optic cable that they lease is known as *dark fiber*.

Dark fiber generically refers to a long, dedicated fiber optic link that can be used without the need for any additional equipment. It can be as long as the particular technology supports.

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

But these cables are simply cables. They are not used in anyway by the company who owns them. They remain dark until the user puts their own light down the fiber. Hence, the term *dark fiber*.

3.4 Classes of service

Applications may require different levels of service and guarantees with respect to delivery, connectivity, and bandwidth. Some applications will need to have bandwidth dedicated to them for the duration of the data exchange. An example of this would be a tape backup application. Other applications may be “bursty” in nature and not require a dedicated connection but they may insist that an acknowledgement is sent for each successful transfer. The Fibre Channel standards provide different classes of service to accommodate different applications needs. Table 3-3 provides brief details of the different class of service.

Table 3-3 Fibre Channel class of service

Class	Description	Requires acknowledge
1	Dedicated connection with full bandwidth	Yes
2	Connectionless switch to switch communication for frame transfer / delivery.	Yes
3	Connectionless switch to switch communication for frame transfer / delivery.	No
4	Dedicated Connection with fraction of bandwidth between ports using Virtual circuits	Yes

Class	Description	Requires acknowledge
6	Dedicated connection for multicast.	Yes
F	Switch to switch communication.	Yes

3.4.1 Class 1

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

3.4.2 Class 2

Class 2 is a connectionless, acknowledged service. Class 2 makes better use of available bandwidth since it allows the fabric to multiplex several messages on a frame-by-frame basis. As frames travel through the fabric they can take different routes, so class 2 service does not guarantee in-order delivery. Class 2 relies on upper layer protocols to take care of frame sequence. The use of acknowledgments reduces available bandwidth, which needs to be considered in large-scale busy networks.

3.4.3 Class 3

There is no dedicated connection in class 3 and the received frames are not acknowledged. Class 3 is also called *datagram connectionless* service. It optimizes the use of fabric resources, but it is now up to the upper layer protocol to ensure that all frames are received in the proper order, and to request to the source device the retransmission of missing frames. Class 3 is a commonly used class of service in Fibre Channel networks.

3.4.4 Class 4

Class 4 is a connection-oriented service like class 1, but the main difference is that it allocates only a fraction of available bandwidth of path through the fabric that connects two N_Ports. Virtual Circuits (VCs) are established between two N_Ports with guaranteed Quality of Service (QoS), including bandwidth and latency. Like class 1, class 4 guarantees in-order delivery of frames and provides acknowledgment of delivered frames, but now the fabric is responsible for multiplexing frames of different VCs. Class 4 service is mainly intended for multimedia applications such as video and for applications that allocate an established bandwidth by department within the enterprise. Class 4 was added in the FC-PH-2 standard.

3.4.5 Class 5

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

3.4.6 Class 6

Class 6 is a variant of class 1, known as multicast class of service. It provides dedicated connections for a reliable multicast. An N_Port may request a class 6 connection for one or more destinations. A multicast server in the fabric will establish the connections and get 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. Class 6 was designed for applications like audio and video requiring multicast functionality. It appears in the FC-PH-3 standard.

3.4.7 Class F

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

3.5 Fibre Channel data movement

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

The reflective differences in fiber-optic cabling mean that intermodal, or modal, dispersion (signal degradation) may occur.

This may result in frames arriving at different times. This bit error rate (BER) is referred to as the jitter budget. No products are entirely jitter free, and this is an important consideration when selecting the components of a SAN.

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

Embedded clocking, though, can be accomplished by different means. Fibre Channel uses a byte-encoding scheme (which is covered in more detail in 3.5.1, “Byte encoding schemes” on page 47) and clock and data recovery (CDR) logic to recover the clock. From this, it determines the data bits that comprise bytes and words.

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

3.5.1 Byte encoding schemes

In order to transfer data over a high-speed serial interface, the data is encoded prior to transmission and decoded upon reception. The encoding process ensures that sufficient

clock information is present in the serial data stream to allow the receiver to synchronize to the embedded clock information and successfully recover the data at the required error rate. This 8b/10b encoding will find errors that a parity check cannot. A parity check will not find even numbers of bit errors, only odd numbers. The 8b/10b encoding logic will find almost all errors.

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

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

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

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

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

We illustrate an encoding example in Figure 3-8.

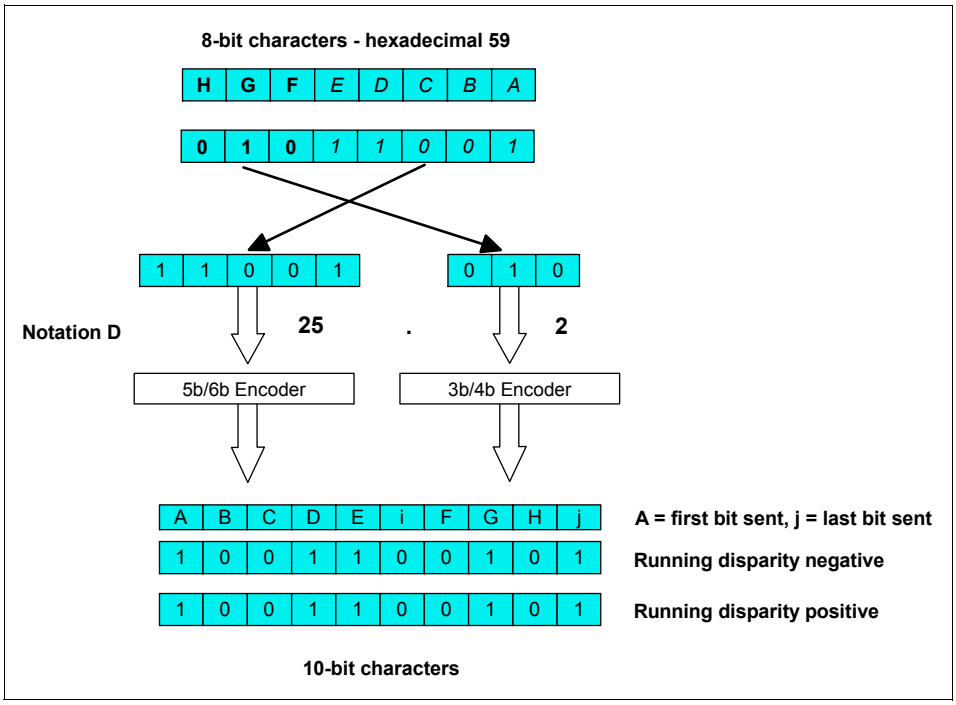

Figure 3-8 8b/10b encoding logic

In the encoding example the following occurs:

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

Running disparity

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

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

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

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

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

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

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

K28.5

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

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

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

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

The two 10-bit encoding of K28.5 are shown in Table 3-4.

Table 3-4 10-bit encoding of K28.5

Name of character	Encoding for current running disparity of	
	Negative	Positive
K28.5	001111 1010	110000 0101

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

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

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

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

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

64b/66b encoding

10 and 16 Gbps communications use 64/66b encoding. 64 bits of data are transmitted as a 66-bit entity.The 66 bit entity is made by prefixing one of two possible two-bit 'preambles' to the 64 bits to be transmitted. If the preamble is '01', the 64 bits are entirely data.

If the preamble is '10', an eight-bit type field follows, plus 56 bits of control information and/or data.The preambles '00' and '11' are not used, and generate an error if seen.

The use of the '01' and '10' preambles guarantees a bit transition every 66 bits, which means that a continuous stream of 0s or 1s cannot be valid data. It also allows easier clock/timer synchronization, as a transition must be seen every 66 bits.

The overhead of the 64B/66B encoding is considerably less than the more common 8b/10b encoding scheme.

3.6 Data transport

In order for Fibre Channel devices to be able to communicate with each other, there needs to be some strict definitions regarding the way that data is sent and received. To this end, some data structures have been defined. It is fundamental to understanding Fibre Channel that you have some knowledge of the way that data is moved around and the mechanisms that are used to accomplish this.

3.6.1 Ordered set

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

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

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

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

- Offline state (OLS)

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

- Not operational (NOS)

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

- Link reset (LR)

  The link reset primitive sequence is used to initiate a link reset.

- Link reset response (LRR)

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

Data transfer

To send data over Fibre Channel, though, we need more than just the control mechanisms. Data is sent in frames. One or more related frames make up a sequence. One or more related sequences make up an exchange.

3.6.2 Frames

Fibre Channel places a restriction on the length of the data field of a frame at 528 transmission words, which is 2112 bytes. (See Table 3-5 on page 52.) Larger amounts of data must be transmitted in several frames. This larger unit that consists of multiple frames is called a *sequence*. An entire transaction between two ports is made up of sequences administered by an even larger unit called an *exchange*.

Note: Some classes of Fibre Channel communication guarantee that the frames will arrive at the destination in the same order in which they were transmitted; other classes do not. If the frames do arrive in the same order in which they were sent, then we are said to have *in order* delivery of frames.

A frame consists of the following elements:

- SOF delimiter

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

Figure 3-9 on page 52 shows the layout of a Fibre Channel frame.

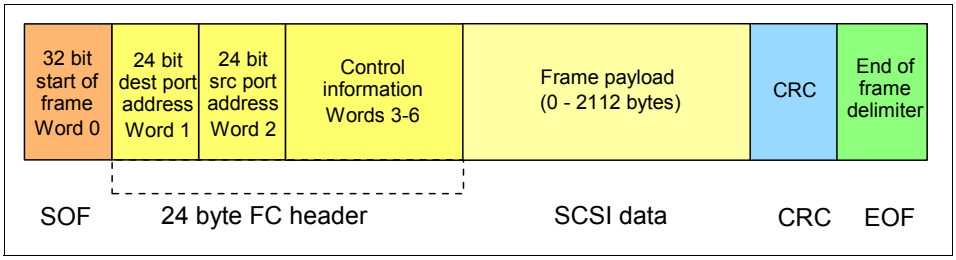

Figure 3-9 Fibre Channel frame structure

Framing rules

The following rules apply to the framing protocol:

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

Transmission word

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

Frames are the building blocks of Fibre Channel. A *frame* is a string of transmission words prefixed by a Start Of Frame (SOF) delimiter and followed by an End Of Frame (EOF) delimiter. The way that transmission words make up a frame is shown in Table 3-5.

Table 3-5 Transmission words in a frame

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

Frame header

Each frame includes a header that identifies the source and destination of the frame as well as control information that manages the frame as well as sequences and exchanges associated with that frame. The structure of the Frame header is shown in Table 3-6 on page 52. The abbreviations are explained below the table.

Table 3-6 The frame header

	Byte 0	Byte 1	Byte 2	Byte 3
Word 0	R_CTL	Destination_ID (D_ID)		
Word 1	Reserved	Source_ID (S_ID)		
Word 2	Type	Frame Control (F_CTL)		
Word 3	SEQ_ID	DF_CTL	SequenceCount (SEQ_CNT)	

	Byte 0	Byte 1	Byte 2	Byte 3
Word 4	Originator X_ID (OX_ID)		Responder X_ID (RX_ID)	
Word 5	Parameter			

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

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

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

Type
Type identifies the protocol of the frame content for data frames, such as SCSI, or a reason code for control frames.

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

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

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

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

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

RX_ID
This field identifies the exchange ID to the responder.

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

3.6.3 Sequences

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

3.6.4 Exchanges

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

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

- Command IU
- Transfer ready IU
- Data IU
- Response IU

Each IU is one sequence of the exchange. Only one participant sends a sequence at a time. Below Figure 3-10 on page 54 indicates the flow of the exchange, sequence and frames.

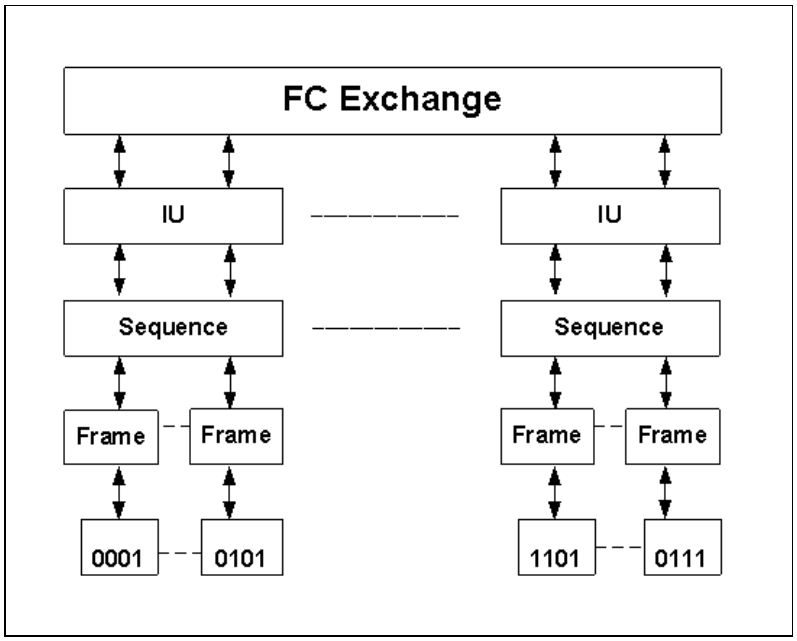

Figure 3-10 FC Exchange, Sequence & frame flow

3.6.5 In order and out of order

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

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

3.6.6 Latency

The term *latency* means the delay between an action requested and an action taking place.

Latency occurs almost everywhere. A simple fact is that it takes time and energy to perform an action. The areas where we particularly need to be aware of latency in a SAN are:

- Ports
- Switches, directors
- Inter Chassis links in a DCX director
- Long distance links
- Inter Switch links
- ASICs

3.6.7 Open Fiber Control

When dealing with lasers there is potential danger to the eyes. Generally, the lasers in use in Fibre Channel are low-powered devices designed for quality of light and signaling rather than for maximum power. However, they can still be dangerous.

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

To add a degree of safety, the concept of Open Fiber Control (OFC) was developed. The idea is as follows:

1. A device is turned on and it sends out low powered light.
2. If it does not receive light back, then it assumes that there is no fiber connected. This is a fail-safe option.
3. When it receives light, it assumes that there is a fiber connected and switches the laser to full power.
4. If one of the devices stops receiving light, then it will revert to the low power mode.

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

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

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

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

3.7 Flow control

Now that we know data is sent in frames, we also need to understand that devices need to temporarily store the frames as they arrive, and until they are assembled in sequence, and then delivered to the upper layer protocol. The reason for this is that due to the high bandwidth that Fibre Channel is capable of, it would be possible to inundate and overwhelm a

target device with frames. There needs to be a mechanism to stop this happening. The ability of a device to accept a frame is called its credit. This credit is usually referred to as the number of buffers (its buffer credit) that a node maintains for accepting incoming data.

3.7.1 Buffer to buffer

Buffer to buffer credits is the maximum frame transfers a port can support. During login, N_Ports and F_Ports at both ends of a link establish its buffer to buffer credit (BB_Credit). Each port states the maximum BB_Credit that they can offer and the lower of the two is used.

3.7.2 End to end

At 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.

3.7.3 Controlling the flow

Two counters are used to accomplish successful flow control: 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 one. When it receives R_RDY from the adjacent port it decrements BB_Credit_CNT by one, and when it receives ACK from the destination port it decrements EE_Credit_CNT by one. If at any time BB_Credit_CNT becomes 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.

3.7.4 Performance

Here we can see the importance that the number of buffers has in overall performance. We need enough buffers to make sure the transmitting port can continue sending frames without stopping in order to use the full bandwidth. This is particularly true with distance. At 1 Gbps a frame occupies between about 75m and 4km of fiber depending on the size of the data payload. In a 100km link we could send many frames before the first one reaches its destination. We need an acknowledgement (ACK) back to start replenishing EE_Credit or a receiver ready (R_RDY) indication to replenish BB_Credit.

For a moment, let us consider frames with 2 KB of data. These occupy approximately 4 km of fiber. We will be able to send about 25 frames before the first arrives at the far end of our 100 km link. We will be able to send another 25 before the first R_RDY or ACK is received, so we would need at least 50 buffers to allow for non-stop transmission at 100 km distance with frames of this size. If the frame size is reduced, more buffers would be required to allow non-stop transmission. In brief the buffer credit management is critical in case of long distance communication, hence the appropriate buffer credit allocation is important to obtain optimal performance. Inappropriate allocation of buffer credit may result in a delay of transmission over the FC link. As a best practice always refer to the default buffer and maximum buffer credit values for each model of switch from each vendor.

4

Ethernet and system networking concepts

In this chapter we will introduce you to Ethernet and system networking concepts, and we will also show you the SAN IP networking options and how we arrive at converged networks.

4.1 Ethernet

Earlier we gave a brief introduction to what a network is and the importance of the models. The Ethernet Standard fits into layer 2 of the OSI Model and refers to the media access layer that devices are connected to (the cable) and compete for access using a Carrier Sense Multiple Access with Collision Detection (CSMA/CD) protocol.

Ethernet is a standard communications protocol embedded in software and hardware devices, intended for building a local area network (LAN). Ethernet was designed by Bob Metcalfe in 1973, and through the efforts of Digital, Intel and Xerox (for whom Metcalfe worked), 'DIX' Ethernet became the standard model for LANs worldwide.

The formal designation for standardization of the Ethernet protocol is sometimes referred to as IEEE 802.3. The IEEE (Institute of Electrical and Electronics Engineers) proposed a working group in February 1980 to standardize network protocols. The third subcommittee worked on a flavor essentially identical to Ethernet, though there are insignificant variances. Consequently, generic use of the term 'Ethernet' might refer to IEEE 802.3 or DIX Ethernet.

Ethernet was originally based on the idea of computers communicating over a shared coaxial cable acting as a broadcast transmission medium. The methods used were similar to those used in radio systems, with the common cable providing the communication channel likened to the luminiferous aether (light-bearing aether) in 19th century physics, and it was from this reference that the name Ethernet was derived.

4.1.1 Shared media

Since all communications happen on the same wire, any information sent by one computer is received by all, even if that information is intended for just one destination.The network interface card (NIC) interrupts the CPU only when applicable packets are received: The card ignores information not addressed to it. Use of a single cable also means that the bandwidth is shared, so that network traffic can be very slow when many stations are simultaneously active.

Collisions reduce throughput by their very nature. In the worst case, when there are lots of hosts with long cables that attempt to transmit many short frames, excessive collisions can reduce throughput dramatically.

Ethernet networks are composed of broadcast domains and there is no clock signal on the wire, as serial connections often have. Instead, Ethernet systems must determine if the wire is in use, and if not, send enough data to enable the remote station to allow it to synchronize properly. This synchronization mechanism combined with the ability to detect other computers attempting to access the wire is a formalized protocol called Carrier Sense Multiple Access-Collision Detect (CSMA/CD).

4.1.2 Ethernet frame

In Figure 4-1 we show an Ethernet frame.

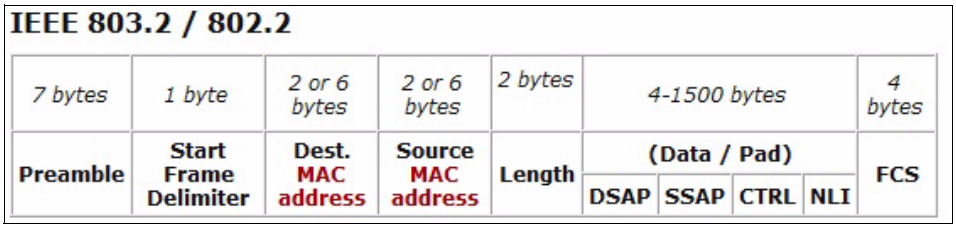

Figure 4-1 Ethernet frame

The breakdown is as follows:

Preamble

This is a stream of bits used to allow the transmitter and receiver to synchronize their communication. The preamble is an alternating pattern of 56 binary ones and zeroes. The preamble is immediately followed by the Start Frame Delimiter.

Start Frame Delimiter

This is always 10101011 and is used to indicate the beginning of the frame information.

Destination MAC

This is the media access control (MAC) address of the machine receiving data. When a network interface card (NIC) is listening to the wire is checking this field for its own MAC address.

Source MAC

This is the MAC address of the machine transmitting data.

Length

This is the length of the entire Ethernet frame in bytes. Although this field can hold any value between 0 and 65,534, it is rarely larger than 1500 as that is usually the maximum transmission frame size for most serial connections. Ethernet networks tend to use serial devices to access the Internet.

Data/Pad (a.k.a. payload)

The data is inserted here. This is where the IP header and data is placed if you are running IP over Ethernet. This field contains IPX information if you are running IPX/SPX (Novell). Contained within the data/padding section of an IEEE 803.2 frame are four specific fields:

- DSAP - Destination Service Access Point
- SSAP - Source Service Access Point
- CTRL - Control bits for Ethernet communication
- NLI - Network Layer Interface
- FCS - Frame Check Sequence

This field contains the Frame Check Sequence (FCS) which is calculated using a Cyclic Redundancy Check (CRC). The FCS allows Ethernet to detect errors in the Ethernet frame and reject the frame if it appears damaged.

4.1.3 How Ethernet works

When a device connected to an Ethernet network wants to send data it first checks to make sure it has a carrier on which to send its data (usually a piece of copper cable connected to a hub or another machine). This is known as *Carrier Sense*.

All machines on the network are free to use the network whenever they like so long as no-one else is transmitting. This is known as *Multiple Access*.

There also needs to be a means of ensuring that when two machines start to transmit data simultaneously, that the resultant corrupted data is discarded, and re-transmissions are generated at differing time intervals. This is known as *Collision Detection*.

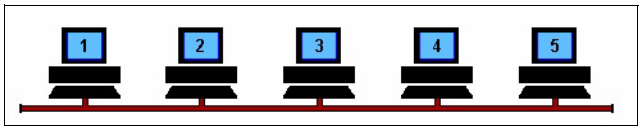

Figure 4-2 Bus Ethernet network

Referring to Figure 4-2 we will assume machine 2 wants to send a message to machine 4, but first it 'listens' to make sure no one else is using the network.

If it is all clear it starts to transmit its data on to the network, each packet of data contains the destination address, the senders address, and of course the data to be transmitted.

The signal moves down the cable and is received by every machine on the network — but because it is only addressed to machine 4, the other machines ignore it. Machine 4 then sends a message back to machine 2 acknowledging receipt of the data.

But what happens when two machines try to transmit at the same time? A collision occurs, and each machine has to 'back off' for a random period of time before re-trying. Figure 4-3 illustrates what happens when two machines are transmitting simultaneously.

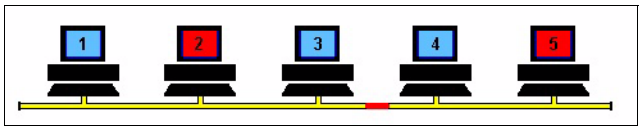

Figure 4-3 Machine 2 and machine 5 both trying to transmit simultaneously.

The resulting collision destroys both signals and each machine knows this has happened because they do not 'hear' their own transmission within a given period of time (this time period is the propagation delay and is equivalent to the time it takes for a signal to travel to the furthest part of the network and back again).

Both machines then wait for a random period of time before re-trying. On small networks this all happens so quickly that it is virtually unnoticeable, however, as more and more machines are added to a network the number of collisions rises dramatically and eventually results in slow network response. The exact number of machines that a single Ethernet segment can handle depends upon the applications being used, but it is generally considered that between 40 and 70 users are the limit before network speed is compromised.

Figure 4-4 on page 61 shows two different scenarios: hub and switch. The hub is where all the machines are interconnected so that only one machine at a time can use the media. In the switch network, more than one machine can be using the media at the time.

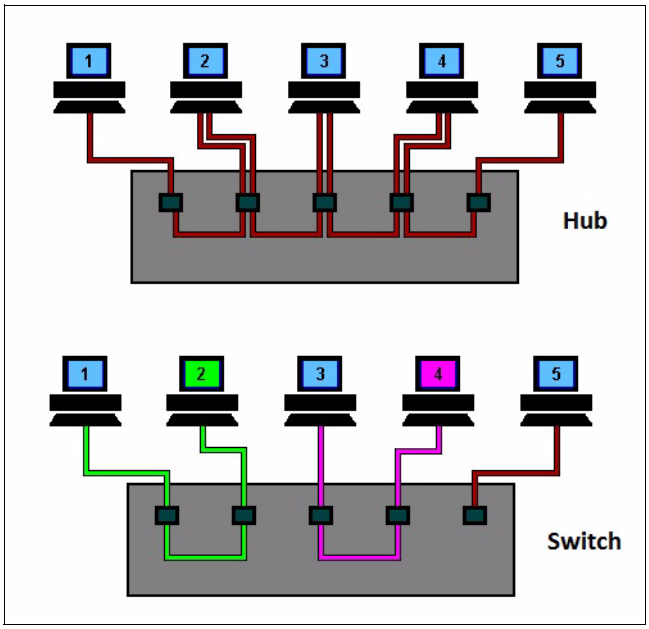

Figure 4-4 Hub and switch scenarios

An Ethernet hub changes the topology from a 'bus' to a 'star wired bus', and as an example let's say again that machine 1 is transmitting data to machine 4 — but this time the signal travels in and out of the hub to each of the other machines.

As you can see, it is still possible for collisions to occur but hubs have the advantage of centralised wiring, and they can automatically bypass any ports that are disconnected or have a cabling fault. This makes the network much more fault tolerant than a coax based system where disconnecting a single connection will bring the whole network down.

With a switch, machines can transmit simultaneously. Each switch reads the destination addresses and 'switches' the signals directly to the recipients without broadcasting to all of the machines on the network.

This 'point to point' switching alleviates the problems associated with collisions and considerably improves network speed

4.1.4 Speed and bandwidth

By convention, network data rates are denoted either in bit/s (bits per second/bps) or byte/s (bytes per second/Bps). In general, parallel interfaces are quoted in byte/s and serial in bit/s.

The numbers below are simplex data rates, which may conflict with the duplex rates vendors sometimes use in promotional materials. Where two values are listed, the first value is the downstream rate and the second value is the upstream rate.

All quoted figures are in metric decimal units, where:

1 Byte = 8 bits
1 Kbps = 1,000 bits per second
1 Mbps = 1,000,000 bits per second
1 Gbps = 1,000,000,000 bits per second
1 KBps = 1,000 bytes per second
1 MBps = 1,000,000 bytes per second

1 GBps = 1,000,000,000 bytes per second
1 TBps = 1,000,000,000,000 bytes per second

Note that this goes against the traditional use of binary prefixes for memory size. These decimal prefixes have long been established in data communications.

Table 4-1 shows the technology rates and the medium.

Table 4-1 Technology rates and medium

Technology	Rate (Bit/s)	Rate (Byte/s)	Media
Fast Ethernet (100BASE-X)	100 Mb/s	12.5 MB/s	UTP Cat 5
Gigabit Ethernet (1000BASE-X)	1000 Mb/s	125 MB/s	UTP Cat 5e / 6
10 Gigabit Ethernet (10GBASE-X)	10000 Mb/s	1250 MB/s	UTP Cat 7 - Fiber

4.1.5 10GbE

From its origin more than 25 years ago, Ethernet has evolved to meet the increasing demands of packet-based networks. Due to its proven low implementation cost, reliability, and relative simplicity of installation and maintenance, Ethernet's popularity has grown to the point that nearly all traffic on the Internet originates or terminates with an Ethernet connection. Further, as the demand for ever-faster network speeds has increased, Ethernet has been adapted to handle these higher speeds, as well as the surges in volume demand that accompany them.

The IEEE 802.3ae* 2002 (10 Gigabit Ethernet standard) is different in some respects from earlier Ethernet standards in that it will only operates in full duplex mode (collision-detection protocols are unnecessary).

Ethernet can now progress to 10 gigabits per second while retaining its critical Ethernet properties, such as the packet format, and the current capabilities are easily transferable to the new standard.

10 Gigabit Ethernet continues the evolution of Ethernet in speed and distance, while retaining the same Ethernet architecture used in other Ethernet specifications, except for one key ingredient. Since 10 Gigabit Ethernet is a full-duplex only technology, it does not need the carrier-sensing multiple-access with collision detection (CSMA/CD) protocol used in other Ethernet technologies. In every other respect, 10 Gigabit Ethernet matches the original Ethernet model.

4.1.6 10GbE copper versus fiber

Once the decision is made to implement 10GbE functionality, organizations must consider the data carrying techniques that facilitate such bandwidth. Copper and fiber cabling are the preeminent technologies for data transmission and provide their own unique benefits and drawbacks.

Copper is the default standard for transmitting data between devices due to its low cost, easy installation and flexibility. It also possesses distinct shortcomings. Copper is best when utilized in short lengths, typically 100 meters or less. When employed over long distances, electromagnetic signal characteristics hinder performance. In addition, bundling copper

cabling can cause interference, making it difficult to employ as a comprehensive backbone. For these reasons, copper cabling has become the principal data carrying technique for communication among PCs and LANs, but not campus or long-distance transmission.

On the other hand, fiber cabling is typically used for remote campus connectivity, crowded wiring closets, long-distance communications and environments that need protection from interference, such as manufacturing areas. Since it is very reliable and less susceptible to attenuation, it is optimum for sending data beyond 100 meters.

However, fiber is also more costly than copper and its use is typically limited to those applications that demand it.

As a result, most organizations utilize a combination of copper and fiber cabling. As these companies transition to 10GbE functionality, they must have a solid understanding of the various cabling technologies and a sound migration strategy to ensure their cabling infrastructure will support their network infrastructure both today and tomorrow.

The IEEE 802.3 Higher Speed Study Group was formed in 1998, and development of 10GigE began the following year. By 2002, the 10GigE standard was first published as IEEE Std 802.3ae-2002. This standard defines a normal data rate of 10 Gigabits, making it ten times faster than the Gigabit Ethernet.

Subsequent standard updates ensued in relation to the first 10GigE version published in 2002. The IEEE 802.3ae-2002 fiber, followed by 802.3ak-2004 in 2004, which were later consolidated into IEEE 802.3-2005 in the year 2005. In 2006, 802.3an-2006, which is a 10 Gigabit Base-T copper twisted pair and an enhanced version with fiber- LRM PMD followed known as 802.3aq-2006. Finally, in 2007, the 802.3ap-2007 with copper backplane involved.

As a result of those standards there are two main types of 10 Gigabit Ethernet;

For fiber:

- 10GBASE-LX4: It supports ranges of between 240 metres (790 ft) and 300 metres (980 ft) over legacy multi-mode cabling. This is achieved through the use of four separate laser sources operating at 3.125 Gbit/s in the range of 1300 nm on unique wavelengths. 10GBASE-LX4 also supports 10 kilometres (6.2 mi) over SMF.
- 10GBASE-SR: Over obsolete 62.5 micron multi-mode fiber cabling (OM1), it has a maximum range of 26–82 metres (85–269 ft), depending on cable type. Over standard 50 µm 2000 MHz·km OM3 multi-mode fiber (MMF), it has a maximum range of 300 metres (980 ft).
- 10GBASE-LR: Has a specified reach of 10 kilometres (6.2 mi), but 10GBASE-LR optical modules can often manage distances of up to 25 kilometres (16 mi) with no data loss.
- 10GBASE-LRM: supports distances up to 220 metres (720 ft) on FDDI-grade 62.5 µm multi-mode fibre originally installed in the early 1990s for FDDI and 100BaseFX networks and 260 metres (850 ft) on OM3. 10GBASE-LRM reach is not quite as far as the older 10GBASE-LX4 standard
- 10GBASE-ER: Extended range has a reach of 40 kilometres (25 mi)
- 10GBASE-ZR: Several manufacturers have introduced 80 km (50 mi) range ER pluggable interfaces under the name 10GBASE-ZR. This 80 km PHY is not specified within the IEEE 802.3ae standard and manufacturers have created their own specifications based upon the 80 km PHY described in the OC-192/STM-64 SDH/SONET specifications.

For copper:

10G Ethernet can also run over twin-ax cabling and twisted pair cabling.

- 10GBASE-CX4: was the first 10G copper standard published by 802.3 (as 802.3ak-2004). It is specified to work up to a distance of 15 m (49 ft). Each lane carries 3.125 G baud of signaling bandwidth.
- 10GBASE-T, or IEEE 802.3an-2006, is a standard released in 2006 to provide 10 Gbit/s connections over un-shielded or shielded twisted pair cables, over distances up to 100 metres (330 ft).

Note: Category 6A or better balanced twisted pair cables specified in ISO 11801 amendment 2 or ANSI/TIA-568-C.2 are needed to carry 10GBASE-T up to distances of 100m. Category 6 cables can carry 10GBASE-T for shorter distances when qualified according to the guidelines in ISO TR 24750 or TIA-155-A.

For backplane:

- 10GBASE-X
- 10GBASE-KX4
- 10GBASE-KR

4.1.7 Virtual local area network

A VLAN is a networking concept in which a network is logically divided into smaller virtual LANs. The Layer 2 traffic in one VLAN is logically isolated from other VLANs as illustrated in Figure 4-5.

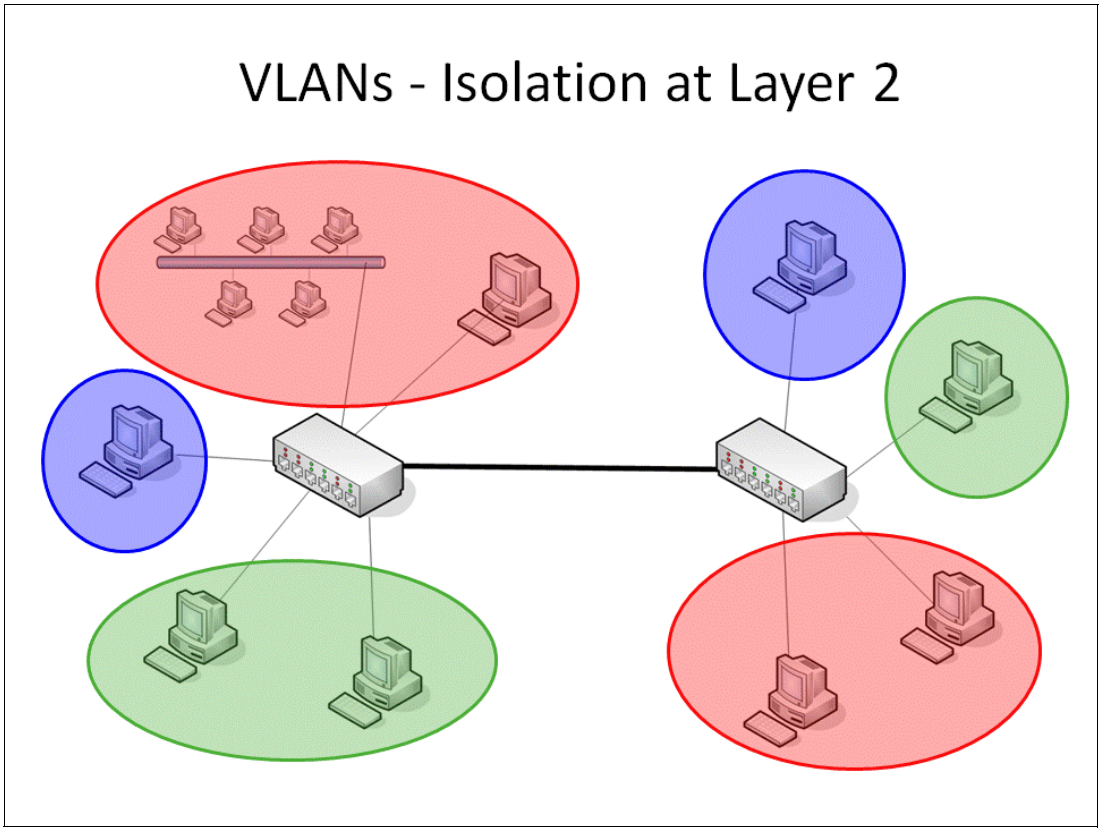

Figure 4-5 Isolation at Layer 2

Figure 4-6 shows two methods for maintaining isolation of VLAN traffic between switches.

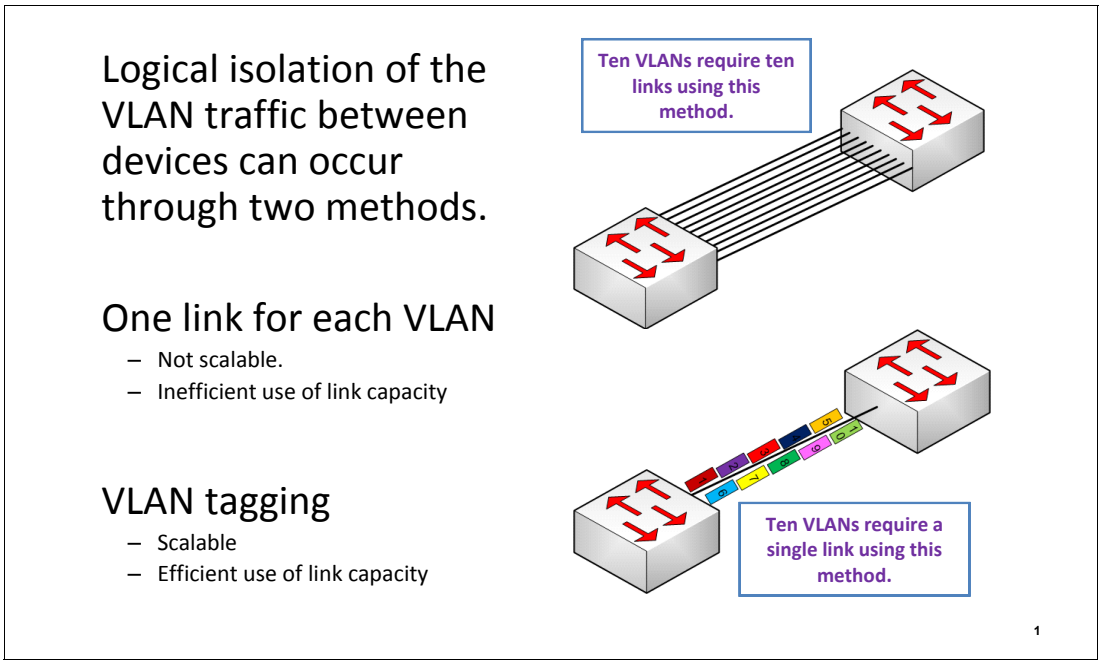

Figure 4-6 VLAN tagging

The first method uses a single link for each VLAN. This method does not scale well because it uses many ports in networks with multiple VLANs and multiple switches. Also, this method does not use link capacity efficiently when traffic in the VLANs is not uniform.

The second method is VLAN tagging over a single link in which each frame in tagged with its VLAN ID. This method is highly scalable because only a single link is required to provide connectivity to many VLANs, which provides for better utilization of the link capacity when VLAN traffic is not uniform.

The protocol for VLAN tagging of frames in a LAN environment is defined by the IEEE 802.1 P/Q standard.

Inter-Switch Link (ISL): ISL is another protocol for providing the VLAN tagging function in a network. This protocol is not compatible with the IEEE 802.1P/Q standard.

Tagged frames

The IEEE 802.1P/Q standard provides a methodology for added information such as VLAN membership and priority to the frame as shown in Figure 4-7.

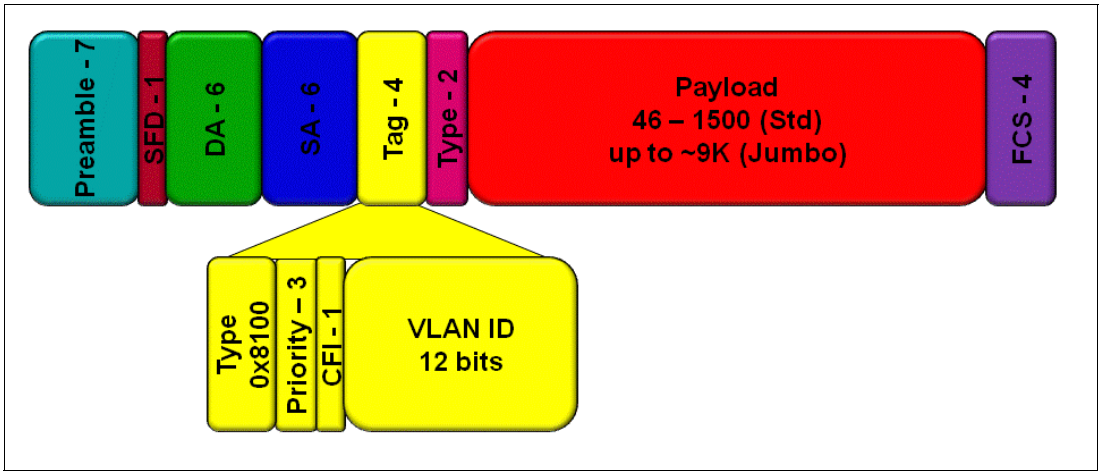

Figure 4-7 IEEE 802.1 P/Q tagged Ethernet frame

The standard provides an additional 4 bytes of information to be added to each Ethernet frame. A frame including this extra information is known as a *tagged frame*.

The 4 byte tag has four component fields:

- A type field that is 2-bytes long with the hexadecimal value of x8100 to identify the frame as an 802.1P/Q tagged frame
- A priority field of 3-bits long to allow a priority value of eight different values to be included in the tag; has the “P” portion of the 802.1P/Q standard
- A Canonical Format Indicator field that is 1-bit long to identify when the contents of the payload field are in canonical format
- A VLAN ID field that is 12-bits long to identify which VLAN the frame is a member of, with 4096 different VLANs possible

4.1.8 Interface VLAN operation modes

Interfaces on a switch can operate in two VLAN modes: single VLAN mode or multiple VLAN mode.

Single VLAN mode

Single VLAN mode operation is also referred to as *access mode*. A port operating in this mode is associated with a single VLAN. Incoming traffic does not have any VLAN identification. While the untagged frames enter the port, the VLAN identification for the VLAN configured for the port is added to the inbound frames.

Switch ports: Some vendors use terms other than access mode for ports operating in single VLAN mode. The switch ports of those vendors might be configured to operate in single VLAN mode by configuring a Port VLAN ID (PVID) and adding the port as a member of the VLAN.

Multiple VLAN mode

Multiple VLAN mode operation is also referred to as *trunk mode*. A port operating in this mode can receive frames that have VLAN tags. The port is also configured with VLANs to which the port is allowed to send and receive frames.

With the IEEE 802.1Q specification, untagged traffic on a multi-VLAN port can be associated with a single VLAN, which is referred to as the “native” VLAN for the port (Figure 4-8). By using this provision, traffic with no VLAN tag can be received and associated with the VLAN that is configured as the PVID or native VLAN. Outbound traffic for this VLAN on a port configured in this manner is transmitted with no tag so that the receiving device can receive the frame in an untagged format.

This method provides compatibility with existing devices or devices that are configured in single VLAN mode and attached to a port configured as a multi-VLAN port.

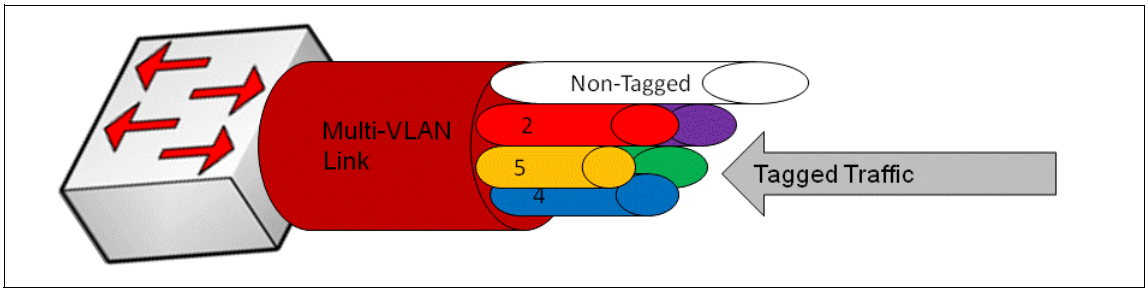

Figure 4-8 Multiple VLAN mode link

Variations in the meaning of trunk: The term *trunk* is used to express different ideas in the networking industry. When using this term, keep in mind that others might use the term in a different manner. The term trunk can mean a port operating in multiple VLAN mode or it can mean a link aggregated port.

4.1.9 Link aggregation

Link aggregation combines multiple physical links to operate as a single larger logical link. The member links no longer function as independent physical connections but as members of the larger logical link (Figure 4-9).

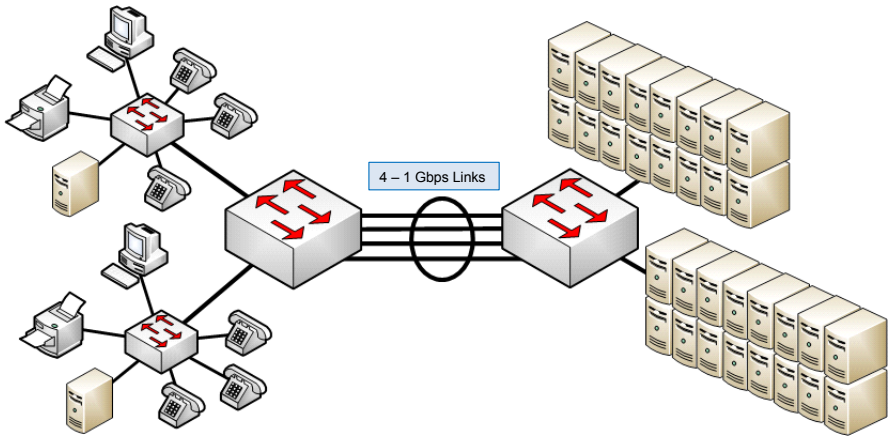

Figure 4-9 Link aggregation

Link aggregation provides greater bandwidth between the devices at each end of the aggregated link. Another advantage of link aggregation is increased availability, because the aggregated link is composed of multiple member links. If one member link fails, the aggregated link continues to carry traffic over the remaining member links.

Each of devices interconnected by the aggregated link uses a hashing algorithm to determine on which of the member links frames will be transmitted. The hashing algorithm might use varying information in the frame to make the decision. This algorithm might include a source MAC, destination MAC, source IP, destination IP and more. It might also include a combination of these values.

4.1.10 Spanning Tree Protocol

Spanning Tree Protocol (STP) provides Layer 2 loop prevention and is commonly in different forms, such as existing STP, Rapid STP (RSTP), Multiple STP (MSTP), and VLAN STP (VSTP). RSTP is a common default STP. It provides faster convergence times than STP. However, some existing networks require the slower convergence times of basic STP.

The operation of STP

STP uses Bridge Protocol Data Unit (BPDU) packets to exchange information with other switches. BPDUs send out hello packets at regular intervals to exchange information across bridges and detect loops in a network topology.

Two types of BPDUs are available:

- Configuration BPDUs

  These BPDUs contain configuration information about the transmitting switch and its ports, including switch and port MAC addresses, switch priority, port priority, and port cost.

- Topology Change Notification (TCN) BPDUs

  When a bridge must signal a topology change, it starts to send TCNs on its root port. The designated bridge receives the TCN, acknowledges it, and generates another one for its own root port. The process continues until the TCN reaches the root bridge.

STP uses the information provided by the BPDUs to elect a root bridge, identify root ports for each switch, identify designated ports for each physical LAN segment, and prune specific redundant links to create a loop-free tree topology. All leaf devices calculate the best path to the root device and place their ports in blocking or forwarding states based on the best path to the root. The resulting tree topology provides a single active Layer 2 data path between any two end stations.

Rapid Spanning Tree Protocol

RSTP provides better re-convergence time than the original STP. RSTP identifies certain links as point to point. When a point-to-point link fails, the alternate link can make the transition to the forwarding state.

An RSTP domain has the following components:

Root port	The “best path” to the root device.
Designated port	Indicates that the switch is the designated bridge for the other switch connecting to this port.
Alternate port	Provides an alternate root port.
Backup port	Provides an alternate designated port.

RSTP was originally defined in the IEEE 802.1w draft specification and later incorporated into the IEEE 802.1D-2004 specification.

Multiple Spanning Tree Protocol

Although RSTP provides faster convergence time than STP, it still does not solve a problem inherent in STP, that all VLANs within a LAN must share the same spanning tree. To solve this problem, we use MSTP to create a loop-free topology in networks with multiple spanning-tree regions.

In an MSTP region, a group of bridges can be modeled as a single bridge. An MSTP region contains multiple spanning tree instances (MSTI). MSTIs provide different paths for different VLANs. This functionality facilitates better load sharing across redundant links.

An MSTP region can support up to 64 MSTIs, and each instance can support anywhere from 1 through 4094 VLANs.

MSTP was originally defined in the IEEE 802.1s draft specification and later incorporated into the IEEE 802.1Q-2003 specification.

VLAN Spanning Tree Protocol

With VSTP, switches can run one or more STP or RSTP instances for each VLAN on which VSTP is enabled. For networks with multiple VLANs, VSTP enables more intelligent tree spanning. This level of tree spanning is possible because each VLAN can have interfaces enabled or disabled depending on the paths that are available to that specific VLAN.

By default, VSTP runs RSTP, but you cannot have both stand-alone RSTP and VSTP running simultaneously on a switch. VSTP can be enabled for up to 253 VLANs.

BPDU protection

BPDU protection can help prevent STP misconfigurations that can lead to network outages. Receipt of BPDUs on certain interfaces in an STP, RSTP, VSTP, or MSTP topology, can lead to network outages.

BPDU protection is enabled on switch interfaces connected to user devices or on interfaces on which no BPDUs are expected, such as edge ports. If BPDUs are received on a protected interface, the interface is disabled and stops forwarding frames.

Loop protection

Loop protection increases the efficiency of STP, RSTP, VSTP, and MSTP by preventing ports from moving into a forwarding state that might result in a loop opening in the network.

A blocking interface can transition to forwarding state in error if the interface stops receiving BPDUs from its designated port on the segment. Such a transition error can occur when there is a hardware error on the switch or software configuration error between the switch and its neighbor.

When loop protection is enabled, the spanning tree topology detects root ports and blocked ports and ensures that both keep receiving BPDUs. If a loop-protection-enabled interface stops receiving BPDUs from its designated port, it reacts as it might react to a problem with the physical connection on this interface. It does not transition the interface to a forwarding state, but instead transitions it to a loop-inconsistent state. The interface recovers and then transitions back to the spanning-tree blocking state as soon as it receives a BPDU.

You must enable loop protection on all switch interfaces that have a chance of becoming root or designated ports. Loop protection is most effective when enabled in the entire switched

network. When you enable loop protection, you must configure at least one action (**alarm**, **block**, or both).

An interface can be configured for either loop protection or root protection, but not for both.

Root protection

Root protection increases the stability and security of STP, RSTP, VSTP, and MSTP by limiting the ports that can be elected as root ports. A root port elected through the regular process has the possibility of being wrongly elected. A user bridge application running on a PC can also generate BPDUs and interfere with root port election. With root protection, network administrators can manually enforce the root bridge placement in the network.

Root protection is enabled on interfaces that should not receive superior BPDUs from the root bridge and should not be elected as the root port. These interfaces become designated ports and are typically on an administrative boundary. If the bridge receives superior STP BPDUs on a port that has root protection enabled, that port transitions to a root-prevented STP state (inconsistency state), and the interface is blocked. This blocking prevents a bridge that should not be the root bridge from being elected the root bridge. After the bridge stops receiving superior STP BPDUs on the interface with root protection, the interface returns to a listening state, followed by a learning state, and ultimately back to a forwarding state. Recovery back to the forwarding state is automatic.

When root protection is enabled on an interface, it is enabled for all the STP instances on that interface. The interface is blocked only for instances for which it receives superior BPDUs. Otherwise, it participates in the spanning tree topology. An interface can be configured for either root protection or loop protection, but not for both.

4.1.11 Link Layer Discovery Protocol

Link Layer Discovery Protocol (LLDP) is a vendor independent protocol for network devices to advertise information about their identity and capabilities. It is referred to as *Station and Media Access Control Connectivity Discovery*, which is specified in the 802.1ab standard. With LLDP and Link Layer Discovery Protocol–Media Endpoint Discovery (LLDP-MED), network devices can learn and distribute device information on network links. With this information, the switch can quickly identify various devices, resulting in a LAN that interoperates smoothly and efficiently.

LLDP-capable devices transmit information in Type Length Value (TLV) messages to neighbor devices. Device information can include specifics, such as chassis and port identification and system name and system capabilities.

LLDP-MED goes one step further, exchanging IP-telephony messages between the switch and the IP telephone. These TLV messages provide detailed information about the PoE policy. With the PoE Management TLVs, the switch ports can advertise the power level and power priority needed. For example, the switch can compare the power needed by an IP telephone running on a PoE interface with available resources. If the switch cannot meet the resources required by the IP telephone, the switch can negotiate with the telephone until a compromise on power is reached.

The switch also uses these protocols to ensure that voice traffic gets tagged and prioritized with the correct values at the source itself. For example, 802.1p class-of-service (COS) and 802.1Q tag information can be sent to the IP telephone.

4.1.12 LLDP TLVs

Basic TLVs include the following information:

Chassis identifier The MAC address associated with the local system.

Port identifier The port identification for the specified port in the local system.

Port description The user-configured port description. The port description can be a maximum of 256 characters.

System name The user-configured name of the local system. The system name can be a maximum of 256 characters.

System description The system description containing information about the software and current image running on the system. This information is not configurable, but taken from the software.

System capabilities The primary function performed by the system. The capabilities that system supports, for example, bridge or router. This information is not configurable, but is based on the model of the product.

Management address
The IP management address of the local system.

Additional 802.3 TLVs include the following details:

Power via MDI A TLV that advertises MDI power support, Power Sourcing Equipment (PSE) power pair, and power class information.

MAC/PHY configuration status
A TLV that advertises information about the physical interface, such as auto-negotiation status and support and MAU type. The information is not configurable, but based on the physical interface structure.

Link aggregation A TLV that advertises if the port is aggregated and its aggregated port ID.

Maximum frame size A TLV that advertises the maximum transmission unit (MTU) of the interface sending LLDP frames.

Port VLAN A TLV that advertises the VLAN name configured on the interface.

LLDP-MED provides the following TLVs:

LLDP MED capabilities
A TLV that advertises the primary function of the port. The capabilities values range 0 through 15:
0Capabilities
1Network policy
2Location identification
3Extended power via MDI-PSE
4Inventory
5 through 15 Reserved

LLDP-MED device class values
0Class not defined
1Class 1 device
2Class 2 device
3Class 3 device
4Network connectivity device
5 through 255 Reserved

Network policy A TLV that advertises the port VLAN configuration and associated Layer 2 and Layer 3 attributes. Attributes include the policy identifier, application types, such as voice or streaming video, 802.1Q VLAN tagging, and 802.1p priority bits and Diffserv code points.

Endpoint location A TLV that advertises the physical location of the endpoint.

Extended power via MDI
A TLV that advertises the power type, power source, power priority, and power value of the port. It is the responsibility of the PSE device (network connectivity device) to advertise the power priority on a port.

4.2 SAN IP networking

Now that we have introduced the protocols at a high level, what are the strategic differences between them all? Do I need them all, any, or none? What are some of the benefits that these technologies will bring me? Some of the benefits that can be realized are:

- Departmental isolation and resource sharing alleviation
- Technology migration and integration
- Remote replication of disk systems
- Remote access to disk and tape systems
- Low-cost connection to SANs
- Inter fabric routing
- Overcoming distance limitations

People do not want to make any large financial investment without any guarantees that there will be some form of return. However, the beauty of these protocols is that they immediately bring benefits. As these are standards based protocols they allow the leveraging of both the existing TCP/IP and FCP infrastructure, they support existing FC devices, and enable simplification of the infrastructure by removing any SAN islands.

4.2.1 The multiprotocol environment

As with any technology it comes with its unique jargon and terminology. Typically it is borrowed from the networking world, but may have a different meaning. It is not our intent to cover each and every unique description, but we will make some distinctions that we feel are important for a basic introduction to routing in an IP SAN.

4.2.2 Fibre Channel switching

A Fibre Channel switch filters and forwards packets between Fibre Channel connections on the *same* fabric, but it cannot transmit packets between fabrics. As soon as you join two switches together, you merge the two fabrics into a single fabric with one set of fabric services.

4.2.3 Fibre Channel routing

A router forwards data packets *between* two or more fabrics. Routers use headers and forwarding tables to determine the best path for forwarding the packets.

Separate fabrics each have their own addressing schemes. When they are joined by a router, there must be a way to translate the addresses between the two fabrics. This mechanism is called *network address translation* (NAT) and is inherent in all the IBM System Storage multiprotocol switch/router products. It is sometimes referred to as FC-NAT to differentiate it from a similar mechanism which exists in IP routers.

4.2.4 Tunneling

Tunneling is a technique that allows one network to send its data via another network's connections. Tunneling works by encapsulating a network protocol within packets carried by the second network. For example, in a Fibre Channel over Internet Protocol (FCIP) solution, Fibre Channel packets can be encapsulated inside IP packets. Tunneling raises issues of packet size, compression, out-of-order packet delivery, and congestion control.

4.2.5 Routers and gateways

When a Fibre Channel router needs to provide protocol conversion or tunneling services, it is a *gateway* rather than a router. However, it has become common usage to broaden the term *router* to include these functions. FCIP is an example of tunneling, while Small Computer System Interface over IP (iSCSI) and Internet Fibre Channel Protocol (iFCP) are examples of protocol conversion.

4.2.6 Internet Storage Name Service

The Internet Storage Name Service (iSNS) protocol facilitates automated discovery, management, and configuration of iSCSI and Fibre Channel devices that exist on a TCP/IP network. iSNS provides storage discovery and management services comparable to those that are found in Fibre Channel networks. What this means is that the IP network appears to operate in a similar capacity as a SAN. Coupling this with its ability to emulate Fibre Channel fabric services, iSNS allows for a transparent integration of IP and Fibre Channel networks as it can manage both iSCSI and Fibre Channel devices.

4.3 Delving deeper into the protocols

We have introduced all the protocols at a high level. Now we will show how they do what they do with the Fibre Channel traffic in greater depth.

4.3.1 FCIP

FCIP is a method for tunneling Fibre Channel packets through an IP network. FCIP encapsulates Fibre Channel block data and transports it over a TCP socket, or tunnel. TCP/IP services are used to establish connectivity between remote devices. The Fibre Channel packets are not altered in any way. They are simply encapsulated in IP and transmitted.

Figure 4-10 shows FCIP tunneling, assuming that the Fibre Channel packet is small enough to fit inside a single IP packet.

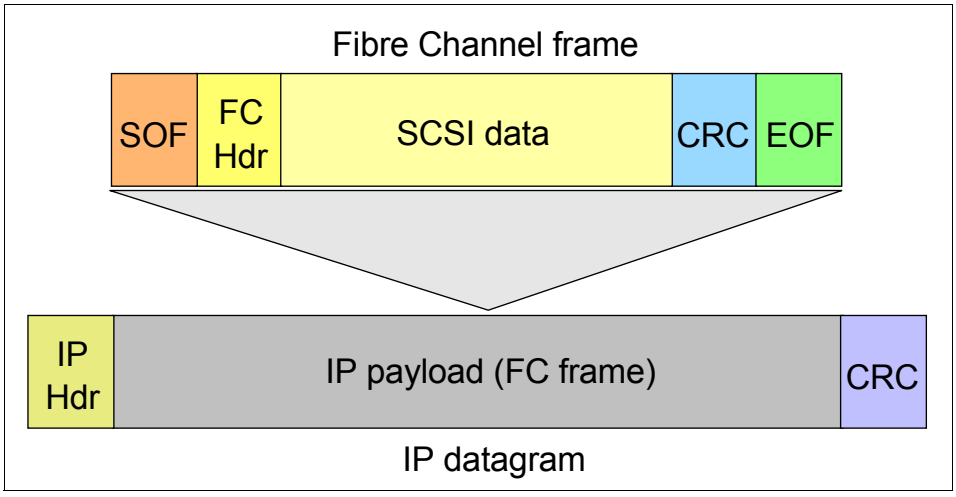

Figure 4-10 FCIP encapsulates the Fibre Channel frame into IP packets

The main advantage is that FCIP overcomes the distance limitations of native Fibre Channel. It also enables geographically distributed devices to be linked using the existing IP infrastructure, while keeping fabric services intact.

The architecture of FCIP is outlined in the Internet Engineering Task Force (IETF) Request for Comment (RFC) 3821, “Fibre Channel over TCP/IP (FCIP)”, available on the Web at:

http://ietf.org/rfc/rfc3821.txt

Because FCIP simply tunnels Fibre Channel, creating an FCIP link is like creating an inter-switch link (ISL), and the two fabrics at either end are merged into a single fabric. This creates issues in situations where you do not want to merge the two fabrics for business reasons, or where the link connection is prone to occasional fluctuations.

Many corporate IP links are robust, but it can be difficult to be sure because traditional IP-based applications tend to be retry-tolerant. Fibre Channel fabric services are not as retry-tolerant. Each time the link disappears or reappears, the switches re-negotiate and the fabric is reconfigured.

By combining FCIP with FC-FC routing, the two fabrics can be left “un-merged”, each with its own separate Fibre Channel services.

4.3.2 iFCP

iFCP is a gateway-to-gateway protocol. It provides Fibre Channel fabric services to Fibre Channel devices over a TCP/IP network. iFCP uses TCP to provide congestion control, error detection, and recovery. iFCP’s primary purpose allows interconnection and networking of existing Fibre Channel devices at wire speeds over a IP network.

Under iFCP, IP components and technology replace the Fibre Channel switching and routing infrastructure. iFCP was originally developed by Nishan Systems who were acquired by McDATA in September 2003 and then McDATA was acquired by Brocade.

To learn more about the architecture and specification of iFCP, refer to the document at the following IETF Web site:

http://www.ietf.org/internet-drafts/draft-ietf-ips-ifcp-14.txt

There is a popular myth that iFCP does not use encapsulation. In fact, iFCP encapsulates the Fibre Channel packet in much the same way that FCIP does. In addition, it maps the Fibre Channel header to the IP header and a TCP session, as shown in Figure 4-11.

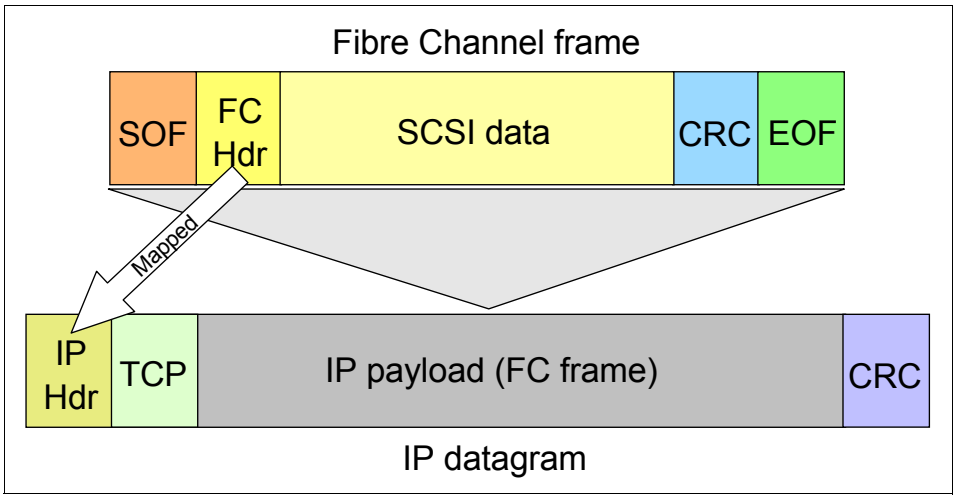

Figure 4-11 iFCP encapsulation and header mapping

iFCP uses the same Internet Storage Name Server (iSNS) mechanism that is used by iSCSI.

iFCP also allows data to fall across IP packets and share IP packets. Some FCIP implementations can achieve a similar result when running software compression, but not otherwise. FCIP typically break each large Fibre Channel packet into two dedicated IP packets. iFCP compression is payload compression only. Headers are not compressed to simplify diagnostics.

iFCP uses one TCP connection per fabric login (FLOGI), while FCIP typically uses one connection per router link (although more are possible). A FLOGI is the process by which an N_PORT registers its presence on the fabric, obtains fabric parameters such as classes of service supported, and receives its N_PORT address. Because under iFCP there is a separate TCP connection per N_PORT to N_PORT couple, each connection can be managed to have its own Quality of Service (QoS) identity. A single incidence of congestion does not need to drop the sending rate for all connections on the link.

While all iFCP traffic between a given remote and local N_PORT pair must use the same iFCP session, that iFCP session can be shared across multiple gateways or routers.

4.3.3 iSCSI

The Small Computer Systems Interface (SCSI) protocol has a client/server architecture. Clients (called *initiators*) issue SCSI commands to request services from logical units on a server known as a *target*. A SCSI *transport* maps the protocol to a specific interconnect.

The SCSI protocol has been mapped over various transports, including Parallel SCSI, Intelligent Peripheral Interface (IPI), IEEE-1394 (firewire), and Fibre Channel. All of these transports are ways to pass SCSI commands. Each transport is I/O specific and has limited distance capabilities.

The iSCSI protocol is a means of transporting SCSI packets over TCP/IP to take advantage of the existing Internet infrastructure.

A session between a iSCSI initiator and an iSCSI target is defined by a session ID that is a combination of an initiator part (ISID) and a target part (Target Portal Group Tag).

The iSCSI transfer direction is defined with respect to the initiator. Outbound or outgoing transfers are transfers from an initiator to a target. Inbound or incoming transfers are transfers from a target to an initiator.

For performance reasons, iSCSI allows a “phase-collapse”. A command and its associated data may be shipped together from initiator to target, and data and responses may be shipped together from targets.

An iSCSI name specifies a logical initiator or target. It is not tied to a port or hardware adapter. When multiple network interface cards (NICs) are used, they should generally all present the same iSCSI initiator name to the targets, because they are simply paths to the same SCSI layer. In most operating systems, the named entity is the operating system image.

The architecture of iSCSI is outlined in IETF RFC 3720, “Internet Small Computer Systems Interface (iSCSI)”, which you can find on the Web at:

http://www.ietf.org/rfc/rfc3720.txt

Figure 4-12 shows the format of the iSCSI packet.

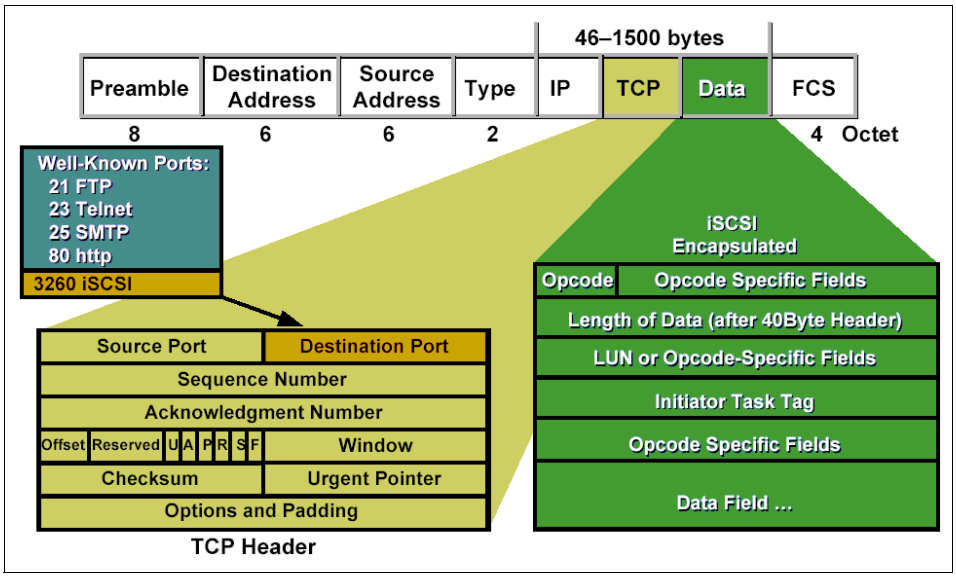

Figure 4-12 iSCSI packet format

Testing on iSCSI latency has shown a difference of up to 1 ms of additional latency for each disk I/O as compared to Fibre Channel. This does not include such factors as trying to do iSCSI I/O over a shared, congested or long-distance IP network, all of which may be tempting for some customers. iSCSI generally uses a shared 1 Gbps network.

iSCSI naming and discovery

Although we do not propose to go for an iSCSI deep dive in this redbook, there are three ways for an iSCSI initiator to understand which devices are in the network:

- In small networks, you can use the **`sendtargets`** command.
- In larger networks, you can use the Service Location Protocol (SLP, multicast discovery).
- In large networks, we recommend that you use Internet Storage Name Service (iSNS).

Note: At time of writing, not all vendors’ have delivered iSNS.

You can find a range of drafts that cover iSCSI naming, discovery, and booting on the following Web site:

http://www.ietf.org/proceedings/02mar/220.htm

4.3.4 Routing considerations

As you would expect with any technology there are going to be a unique set of characteristics that need to be given consideration. The topics that follow briefly describe some of the issues, or items, that are considerations in a multiprotocol Fibre Channel environment.

4.3.5 Packet size

The standard size of a Fibre Channel packet is 2148 bytes, and the standard IP packet size is 1500 bytes (with a 1460 byte payload). It does not take an Einstein to work out that one is larger than the other and will need to be accommodated somehow.

When transporting Fibre Channel over IP, you can use jumbo IP packets to accommodate larger Fibre Channel packets. Keep in mind that jumbo IP packets must be turned on for the whole data path. In addition, a jumbo IP packet is not compatible with any devices in the network that do not have jumbo IP packets enabled.

Alternatively, you can introduce a variety of schemes to split Fibre Channel packets across two IP packets. Some compression algorithms can allow multiple small Fibre Channel packets or packet segments to share a single IP packet.

Each technology and each vendor may implement this differently. But the key point is this: they all try to avoid sending small inefficient packets.

4.3.6 TCP congestion control

Sometimes standard TCP congestion mechanisms may not be suitable for tunneling storage. Standard TCP congestion control is designed to react quickly and severely to network congestion, but recover slowly. This is well suited to traditional IP networks being somewhat variable and unreliable. But for storage applications, this approach is not always appropriate and may cause disruption to latency-sensitive applications.

When three duplicate unanswered packets are sent on a traditional TCP network, the sending rate backs-off by 50%. When packets are successfully sent, it does a slow-start linear ramp-up again.

Some vendors tweak the back-off and recovery algorithms. For example, the tweak causes the send rate to drop by 12.5% each time congestion is encountered, and then to recover rapidly to the full sending rate by doubling each time until full rate is regained.

Other vendors take a simpler approach to achieve much the same end.

If you are sharing your IP link between storage and other IP applications, then using either of these storage friendly congestion controls may impact your other applications.

You can find the specification for TCP congestion control on the Web at:

http://www.ietf.org/rfc/rfc2581.txt

4.3.7 Round-trip delay

Round-trip link latency is the time it takes for a packet to make a round-trip across the link. The term *propagation delay* is also sometime used. Round-trip delay generally includes both inherent latency and delays due to congestion.

Fibre Channel cable has an inherent latency of approximately five microseconds per kilometer each way. Typical Fibre Channel devices, like switches and routers, have inherent latencies of around five microseconds each way. IP routers might vary between five and one hundred microseconds in theory, but when tested with filters applied, the results are more likely to be measured in milliseconds.

This is the essential problem with tunneling Fibre Channel over IP. Fibre Channel applications are generally designed for networks that have round-trip delays measured in microseconds. IP networks generally deliver round-trip delays measured in milliseconds or tens of milliseconds. Internet connections often have round-trip delays measured in hundreds of milliseconds.

Any round-trip delay caused by additional routers and firewalls along the network connection also needs to be added to the total delay. The total round-trip delay varies considerably depending on the models of routers or firewalls used, and the traffic congestion on the link.

So how does this affect you? If you are purchasing the routers or firewalls yourself, we recommend that you include the latency of any particular product in the criteria that you use to choose the products. If you are provisioning the link from a service provider, we recommend that you include at least the maximum total round-trip latency of the link in the service-level agreement (SLA).

Time of frame in transit

The time of frame in transit is the actual time that it takes for a given frame to pass through the slowest point of the link. Therefore it depends on both the frame size and link speed.

The maximum size of the payload in a Fibre Channel frame is 2112 bytes. The Fibre Channel headers add 36 bytes to this, for a total Fibre Channel frame size of 2148 bytes. When transferring data, Fibre Channel frames at or near the full size are usually used.

If we assume that we are using jumbo frames in the Ethernet, the complete Fibre Channel frame can be sent within one Ethernet packet. The TCP and IP headers and the Ethernet medium access control (MAC) add a minimum of 54 bytes to the size of the frame, giving a total Ethernet packet size of 2202 bytes, or 17616 bits.

For smaller frames, such as the Fibre Channel acknowledgement frames, the time in transit is much shorter. The minimum possible Fibre Channel frame is one with no payload. With FCIP encapsulation, the minimum size of a packet with only the headers is 90 bytes, or 720 bits.

Table 4-2 details the transmission times of this FCIP packet over some common wide area network (WAN) link speeds.

Table 4-2 FCIP packet transmission times over different WAN links

Link type	Link speed	Large packet	Small packet
Gigabit Ethernet	1250 Mbps	14 µs	0.6 µs
OC-12	622.08 Mbps	28 µs	1.2 µs
OC-3	155.52 Mbps	113 µs	4.7 µs

Link type	Link speed	Large packet	Small packet
T3	44.736 Mbps	394 µs	16.5 µs
E1	2.048 Mbps	8600 µs	359 µs
T1	1.544 Mbps	11 400 µs	477 µs

If we cannot use jumbo frames, each large Fibre Channel frame needs to be divided into two Ethernet packets. This doubles the amount of TCP, IP, and Ethernet MAC overhead for the data transfer.

Normally each Fibre Channel operation transfers data in only one direction. The frames going in the other direction are close to the minimum size.

4.4 Multiprotocol solution briefs

The solution briefs in the following sections show how you can use multiprotocol routers.

4.4.1 Dividing a fabric into sub-fabrics

Let us suppose that you have eight switches in your data center, and they are grouped into two fabrics of four switches each. Two of the switches are used to connect the development/test environment, two are used to connect a joint-venture subsidiary company, and four are used to connect the main production environment.

The development/test environment does not follow the same change control disciplines as the production environment. Also systems and switches can be upgraded, downgraded, or rebooted on occasions, usually unscheduled and without any form of warning.

The joint-venture subsidiary company is up for sale. The mandate is to provide as much separation and security as possible between it and the main company, and the subsidiary. The backup/restore environment is shared between the three environments.

In summary, we have a requirement to provide a degree of isolation, and a degree of sharing. In the past this would have been accommodated through zoning. Some fabric vendors may still recommend that approach as the simplest and most cost-effective. However as the complexity of the environment grows, zoning can become complex. Any mistakes in setup can disrupt the entire fabric. Adding FC-FC routing to the network allows each of the three environments to run separate fabric services and provides the capability to share the tape backup environment.

In larger fabrics with many switches and separate business units, for example in a shared services hosting environment, separation and routing are valuable in creating a larger number of simple fabrics, rather than fewer more complex fabrics.

4.4.2 Connecting a remote site over IP

Suppose you want to replicate your disk system to a remote site, perhaps 50 km away synchronously, or 500 km away asynchronously. Using FCIP tunneling or iFCP conversion, you can transmit your data to the remote disk system over a standard IP network. The router includes Fibre Channel ports to connect back-end devices or switches and IP ports to connect to a standard IP wide area network router. Standard IP networks are generally much

lower in cost to provision than traditional high quality dedicated dense wavelength division multiplexing (DWDM) networks. They also often have the advantage of being well understood by internal operational staff.

Similarly you might want to provision storage volumes from your disk system to a remote site by using FCIP or iFCP.

Note: FCIP and iFCP can provide a low cost way to connect remote sites using familiar IP network disciplines.

4.4.3 Connecting hosts using iSCSI

Many hosts do not require high bandwidth low latency access to storage. For such hosts, iSCSI may be a more cost-effective connection method. iSCSI can be thought of as an IP SAN. There is no requirement to provide a Fibre Channel switch port for every server, nor to purchase Fibre Channel host bus adapters (HBAs), nor to lay Fibre Channel cable between storage and servers.

The iSCSI router has both Fibre Channel ports and Ethernet ports to connect to servers located either locally on the Ethernet or remotely over a standard IP wide area network connection.

The iSCSI connection delivers block I/O access to the server so it is application independent. That is, an application cannot really tell the difference between direct SCSI, iSCSI, or Fibre Channel, since all three are delivery SCSI block I/Os.

Different router vendors quote different limits on the number of iSCSI connections that are supported on a single IP port.

iSCSI places a significant packetizing and depacketizing workload on the server CPU. This can be mitigated by using TCP/IP offload engine (TOE) Ethernet cards. However since these cards can be expensive, they somewhat undermine the low-cost advantage of iSCSI.

Note: iSCSI can be used to provide low-cost connections to the SAN for servers that are not performance critical.

5

Topologies and other fabric services

In this chapter we will introduce Fibre Channel topologies, and other fabric services commonly encountered in a SAN.

We will also provide an insight into the emerging converged topology and the option to merge FC to FCOE.

5.1 Fibre Channel topologies

Fibre Channel based networks support three types of base topologies, which include point-to-point, arbitrated loop, and switched fabric. A switched fabric is the most commonly encountered topology today and it has sub-classifications of topology. Figure 5-1 depicts the various classifications of SAN topology.

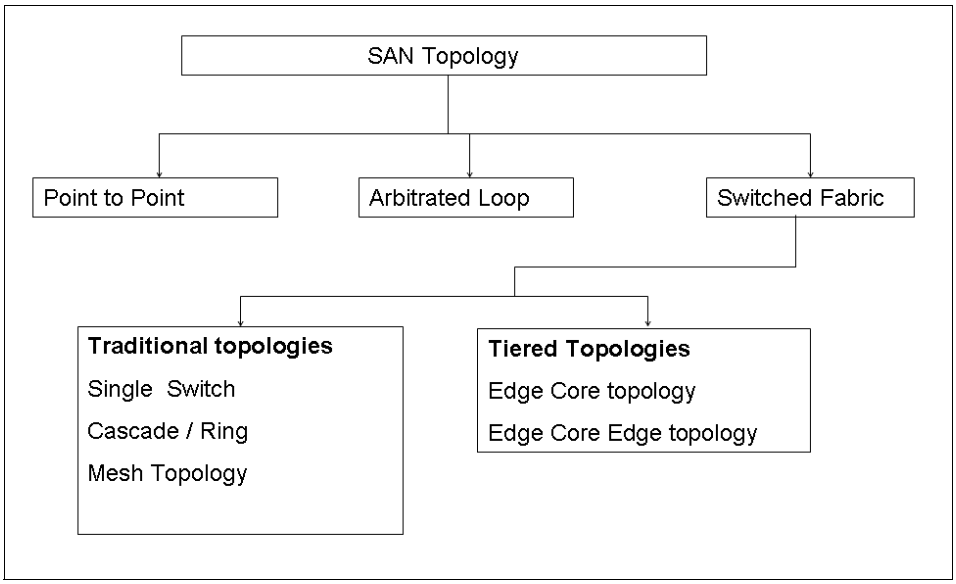

Figure 5-1 SAN topologies

5.1.1 Point-to-point

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

Fibre Channel is a full duplex protocol, which means both paths transmit data simultaneously. As an example, Fibre Channel connections based on the 1 Gbps standard are able to transmit at 100 MBps and receive at 100 MBps simultaneously. Again, as an example, for Fibre Channel connections based on the 2 Gbps standard, they can transmit at 200 MBps and receive at 200 MBps simultaneously. This will extend to 4 Gbps,8 Gbps and 16 GBPS technologies as well.

Illustrated in Figure 5-2 on page 85 is a simple point-to-point connection.

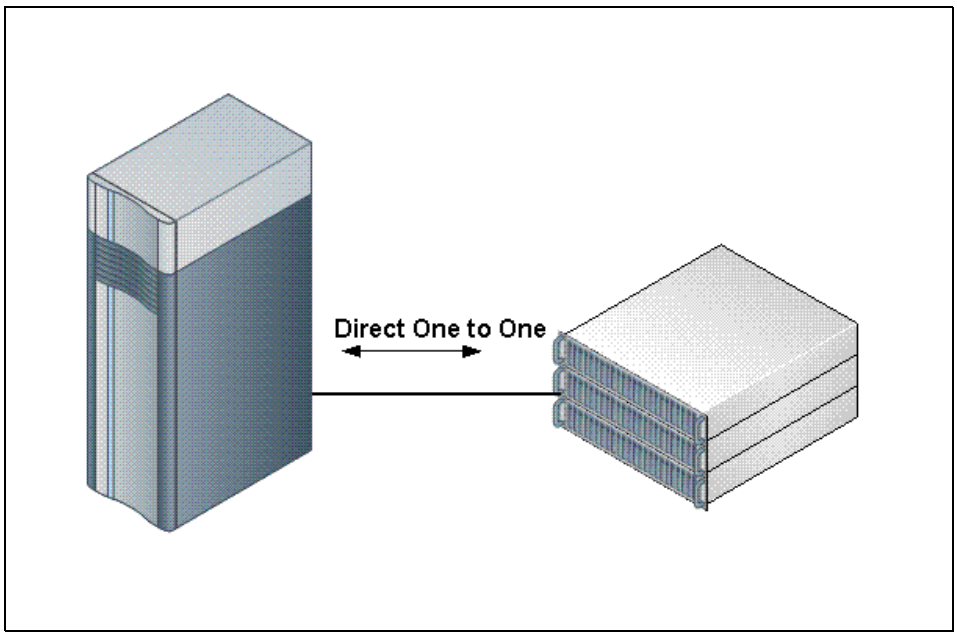

Figure 5-2 Point-to-point

5.1.2 Arbitrated loop

Note: Although this topology is rarely encountered these days, and is considered as a legacy topology, we include it for historical reasons only.

Our second topology is Fibre Channel Arbitrated Loop (FC-AL). FC-AL is more useful for storage applications. It is a loop of up to 126 nodes (NL_Ports) that is managed as a shared bus. Traffic flows in one direction, carrying data frames and primitives around the loop with a total bandwidth of 400 MBps (or 200 MBps for a loop based on 2 Gbps technology).

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

A simple loop, configured using a hub, is shown in Figure 5-3 on page 86.

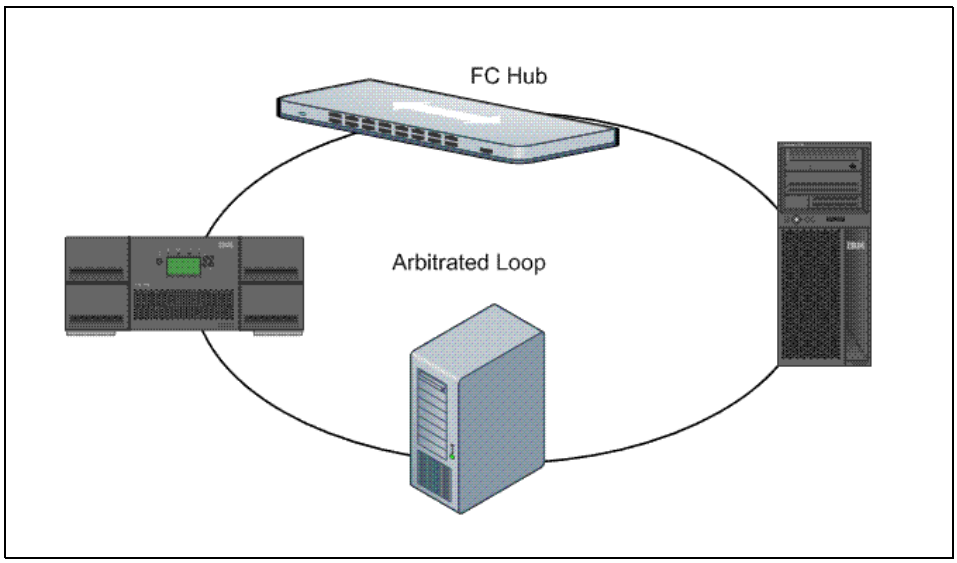

Figure 5-3 Arbitrated loop

We discuss FC-AL in more depth in 5.4, “Fibre Channel Arbitrated Loop protocols” on page 106.

5.1.3 Switched fabric

Our third, and the most useful topology used in SAN implementations, is Fibre Channel Switched Fabric (FC-SW). It applies to switches and directors that support the FC-SW standard, that is, it is not limited to switches as its name suggests. A Fibre Channel fabric is one or more fabric switches in a single, sometimes extended, configuration. Switched fabrics provide full bandwidth per port compared to the shared bandwidth per port in arbitrated loop implementations.

One of the key differentiators is that if you add a new device into the arbitrated loop, you further divide the shared bandwidth. However, in a switched fabric, adding a new device or a new connection between existing ones actually increases the bandwidth. For example, an 8-port switch (for example let’s assume it is based on 2 Gbps technology) with three initiators and three targets can support three concurrent 200 MBps conversations or a total of 600 MBps throughput (1,200 MBps if full-duplex applications were available).

A switched fabric configuration is shown in Figure 5-4.

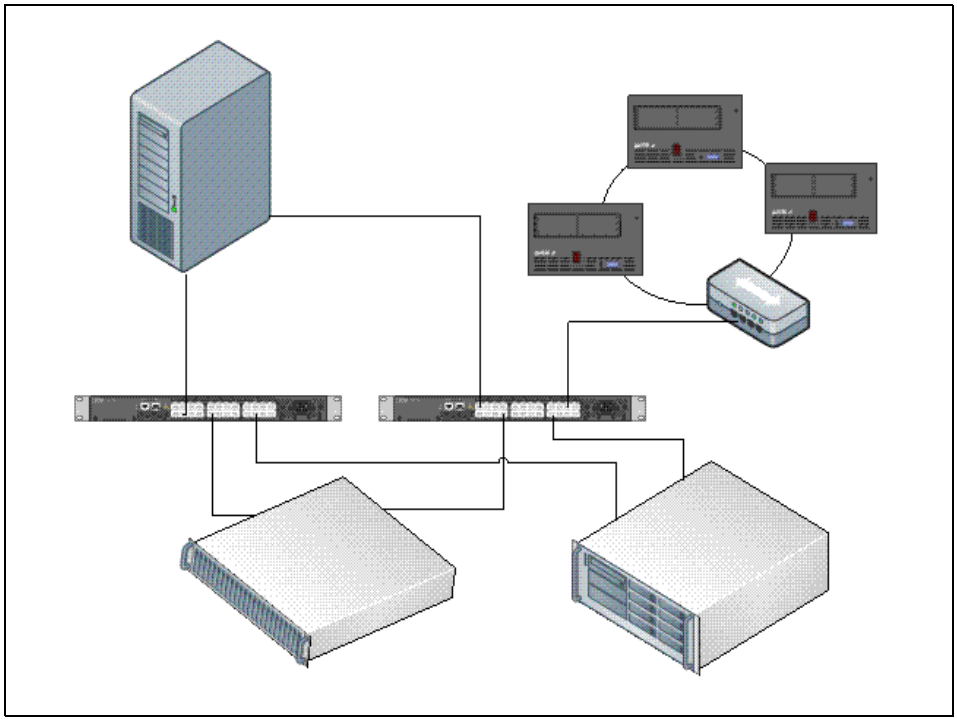

Figure 5-4 Sample switched fabric topology

This is one of the major reasons why arbitrated loop is considered a legacy SAN topology. A switched fabric is usually referred to as a *fabric.*

In terms of switch interconnections the switched SAN topologies can be classified as:

- Single switch topology
- Cascaded and ring topology
- Mesh topology

5.1.4 Single switch topology.

This topology has only one switch and has no Inter switch links (ISL). It is the simplest design for infrastructures which do not need any redundancy. Due to the issues of it introducing a single point of failure, this topology is very rarely used. Figure 5-5 indicates a single switch topology with all devices connected to same switch.

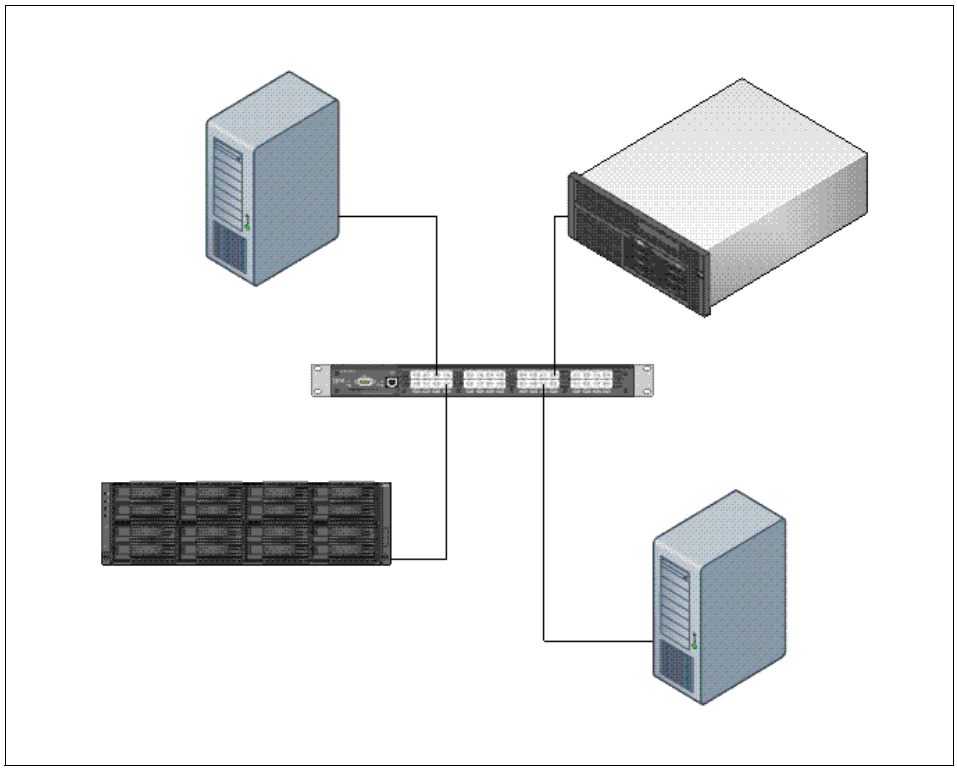

Figure 5-5 Single switch Topology

5.1.5 Cascaded and ring topology

In a cascaded topology switches are connected in a ‘queue’ fashion as shown in Figure 5-6, Even in a ring topology the switches are connected in a queue fashion but it forms a closed ring with an additional Inter Switch link as shown in Figure 5-7.

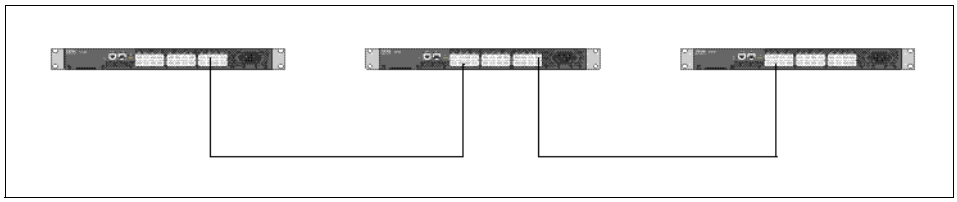

Figure 5-6 Cascade topology

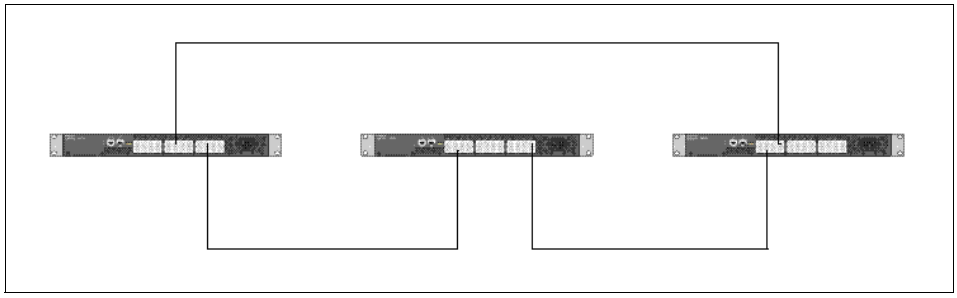

Figure 5-7 Ring Topology

5.1.6 Mesh topology

In a full mesh topology each switch is connected to every other switch in the fabric as shown in Figure 5-8 on page 89.

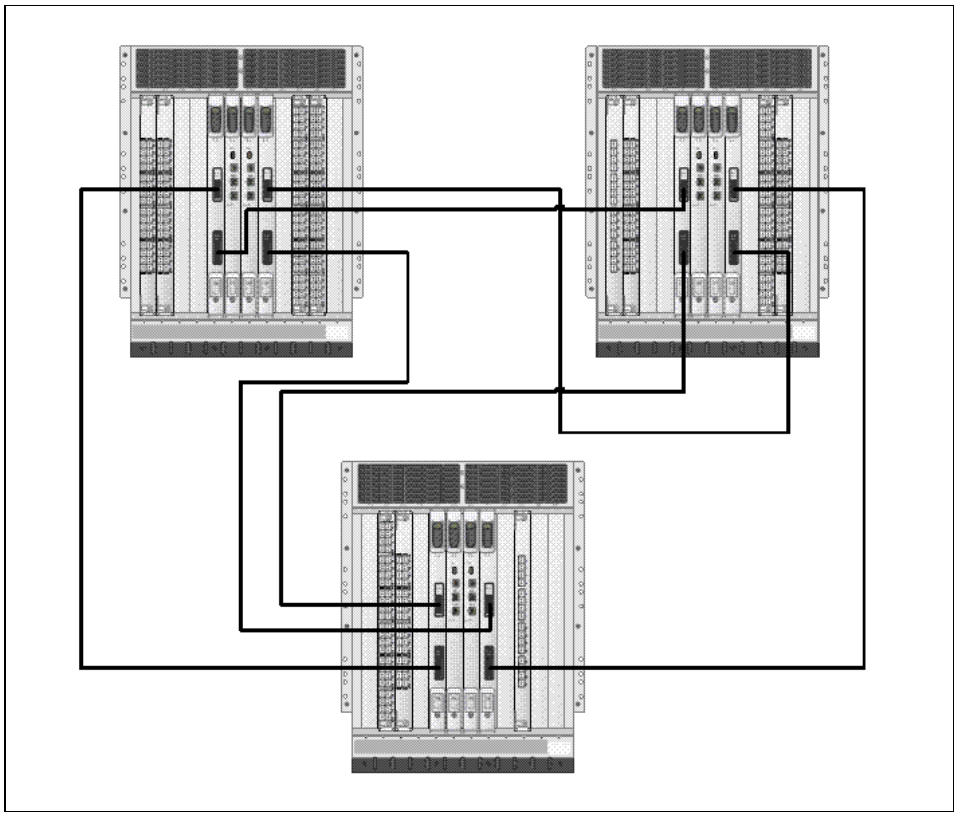

Figure 5-8 IBM SAN768B connected to form a mesh topology

In terms of a tiered approach the switched fabric can be further classified as:

- Core Edge Topology
- Edge Core Edge topology

5.1.7 Core Edge Topology

In this topology the servers are connected to the edge fabric and the storage is connected to core switches. Figure 5-9 on page 90 depicts the core edge topology.

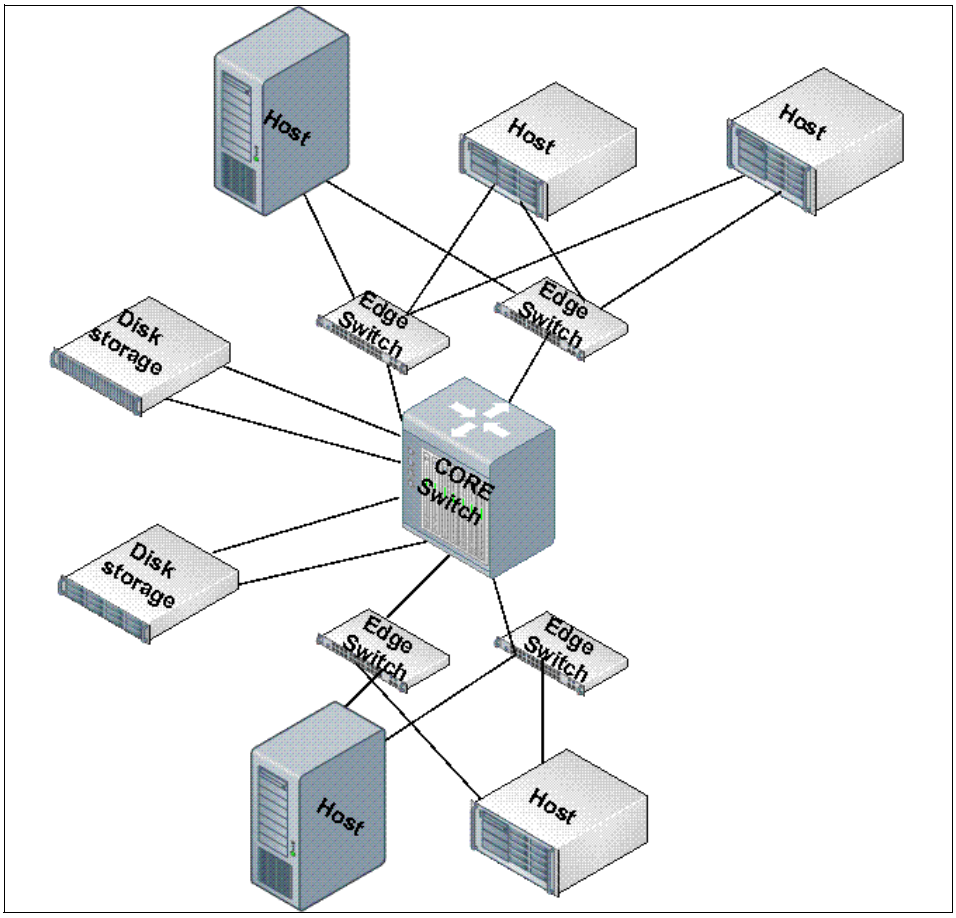

Figure 5-9 Core Edge topology

5.1.8 Edge Core Edge topology

In this topology the server and storage are connected to the edge fabric and the core switch connectivity is used only for scalability in terms of connecting to edge switches, expanding the SAN traffic flow to long distance via DWDM, connecting to virtualisation appliances, and encryption switches. Also the servers may be isolated to one edge and storage can be at the other edge which helps with management. Figure 5-10 on page 91 shows the edge core edge topology.

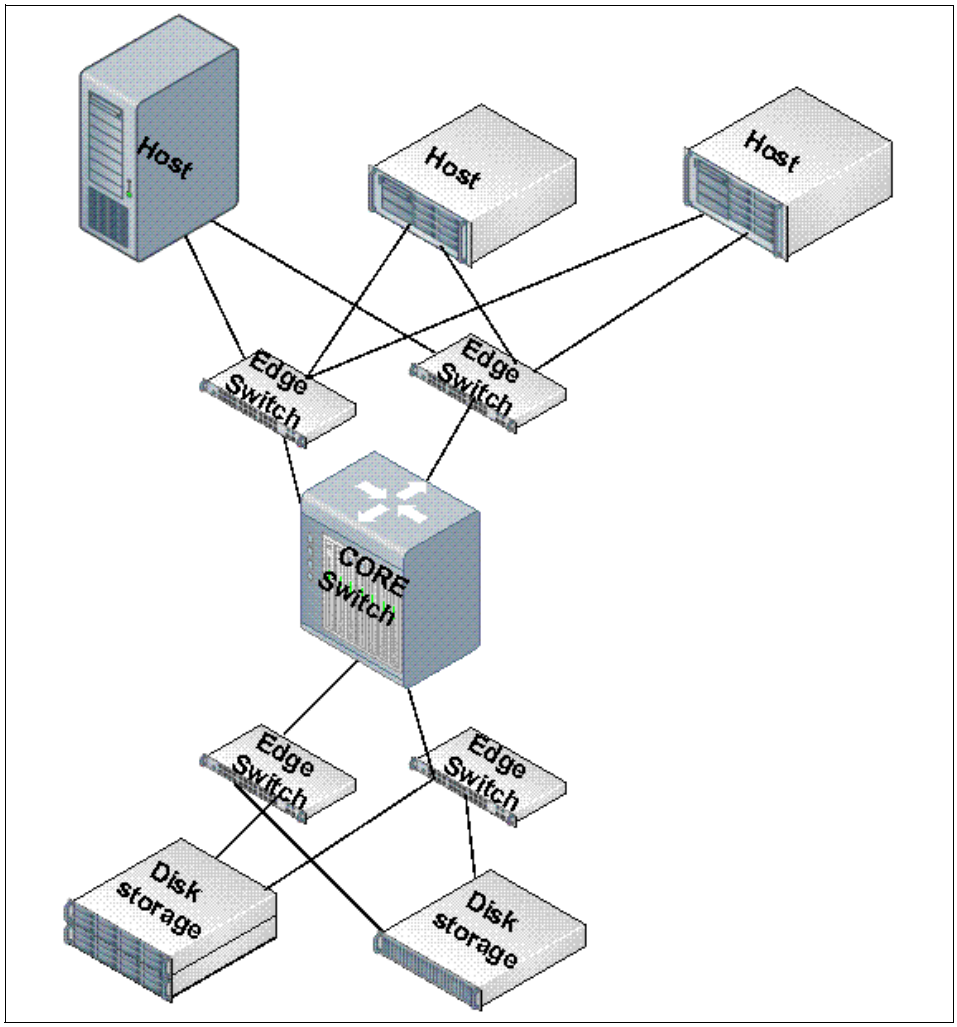

Figure 5-10 Edge Core Edge topology

5.2 Port types

The basic building block of the Fibre Channel is the port. The following lists the various types of Fibre Channel port types and their purposes in switches, servers, and storage.

5.2.1 Common Port types

- F_Port: This is a fabric port that is connected to an N_Port point-point to a switch.
- FL_Port: This is a fabric port that is counted to a loop device. It is used to connect an NL_Port to the switch in a public loop configuration.
- TL_port: Cisco specific port type - It is a Translative loop port connected with non fabric aware, private loop devices.
- G_Port: This is a generic port that can operate as either an E_Port or an F_Port. A port is defined as a G_Port after it is connected but has not received a response to *loop* initialization or has not yet completed the link initialization procedure with the adjacent Fibre Channel device.

- L_Port: This is a loop-capable node or switch port.
- U_Port: This is a universal port—a more generic switch port than a G_Port. It can operate as either an E_Port, F_Port, or FL_Port. A port is defined as a U_Port when it is not connected or has not yet assumed a specific function in the fabric.
- N_Port: This is a node port that is not loop capable. It is host end port used to connect to the fabric switch.
- NL_Port: This is a node port that is loop capable. It is used to connect an equipment port to the fabric in a loop configuration through an L_Port or FL_Port. Figure 5-11 on page 92 depicts different common port types of switch and nodes.

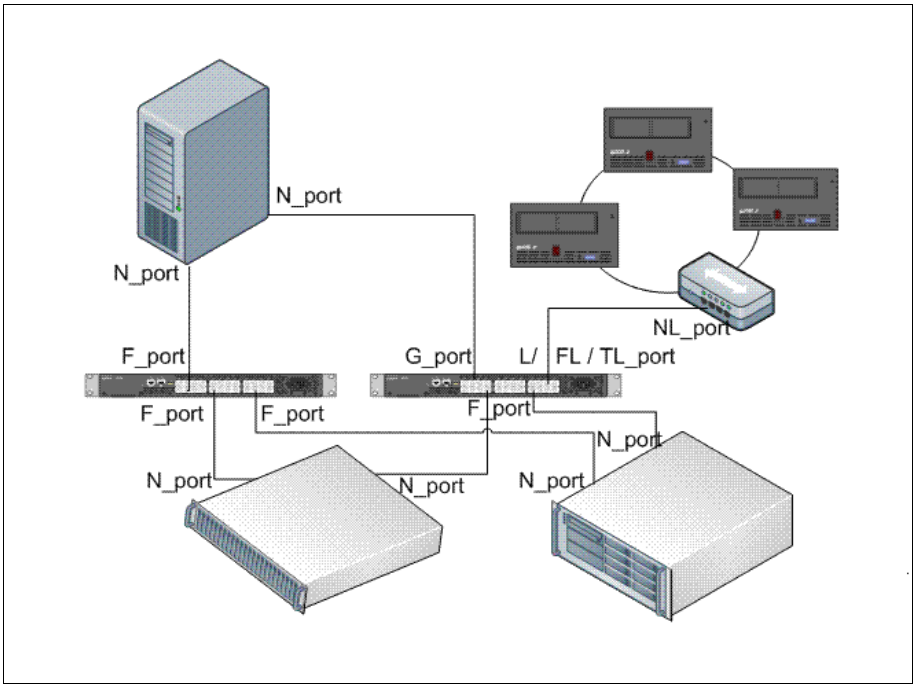

Figure 5-11 Common Port types

5.2.2 Expansion Port types

These ports are found in a multi-switch fabric where switches are interconnected via a FC link.

- E_Port: This is an expansion port. A port is designated an E_Port when it is used as an inter-switch expansion port (ISL) to connect to the E_Port of another switch, to enlarge the switch fabric.
- Ex_port: The type of E_Port used to connect a Multiprotocol Router to an edge fabric. An EX_Port follows standard E_Port protocols, and supports FC-NAT, but does not allow fabric merging across EX_Ports.
- VE_port: A virtual E port is a port that emulates an E_Port over an FCIP link. VE port connectivity is supported over point-to-point links.
- VEX_port: VEX_Ports are routed VE_Ports, just as Ex_Ports are routed E_Ports. VE_Ports and VEX_Ports have the same behavior and functionality.
- TE_port: The TE_port provides not only standard E_port functions but allows for routing of multiple VSANs (Virtual SANs). This is accomplished by modifying the standard Fibre Channel frame (vsan tagging) upon ingress/egress of the VSAN environment. It is also known as a Trunking E_port.

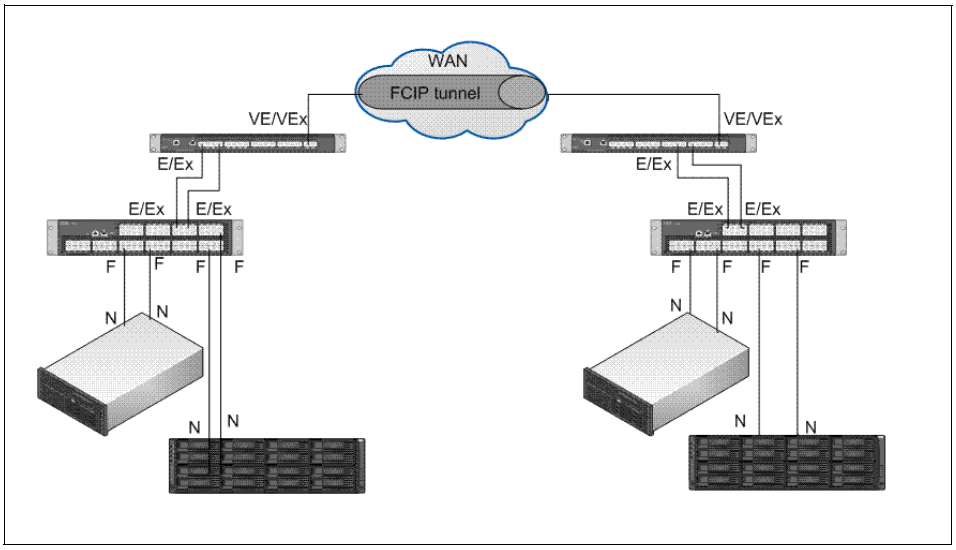

Figure 5-12 Fabric with expansion ports

5.2.3 Diagnostic Port types

- D_port: This is a diagnostic port type which can be enabled only on the 16 Gbps b-type switches with Fabric Operating System 7.0. This uses Spinfab test and performs electrical loop back, optical loop back, measures link distance, and also stress tests with a link saturation test. Figure 5-13 on page 94 describe the different test options, also long distance cable checks also can be done with D_Port diagnostic capabilities.

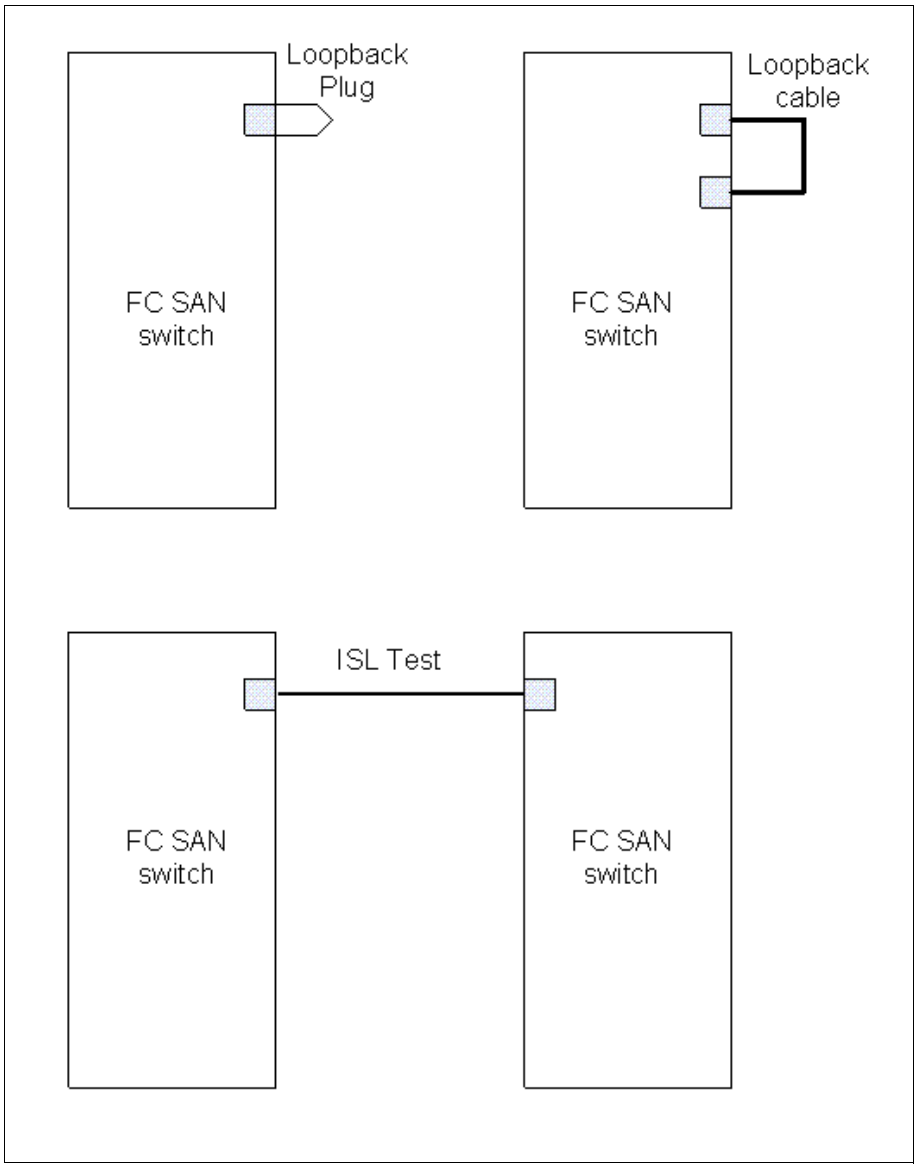

Figure 5-13 D_Port type diagnostics

- MTx_Port: CNT port used as a mirror for viewing the transmit stream of the port to be diagnosed.
- MRx_Port: CNT port used as a mirror for viewing the receive stream of the port to be diagnosed.
- SD_Port: Cisco SPAN diagnostic port used for diagnostic capture with a connection to SPAN- switch port analyzer.
- ST_port: Cisco's port type for Remote SPAN monitoring in a source switch. This will be an undedicated port used for RSPAN analysis, and will not be connected to any other device.

Figure 5-14 on page 95 represents the Cisco specific Fibre Channel port types.

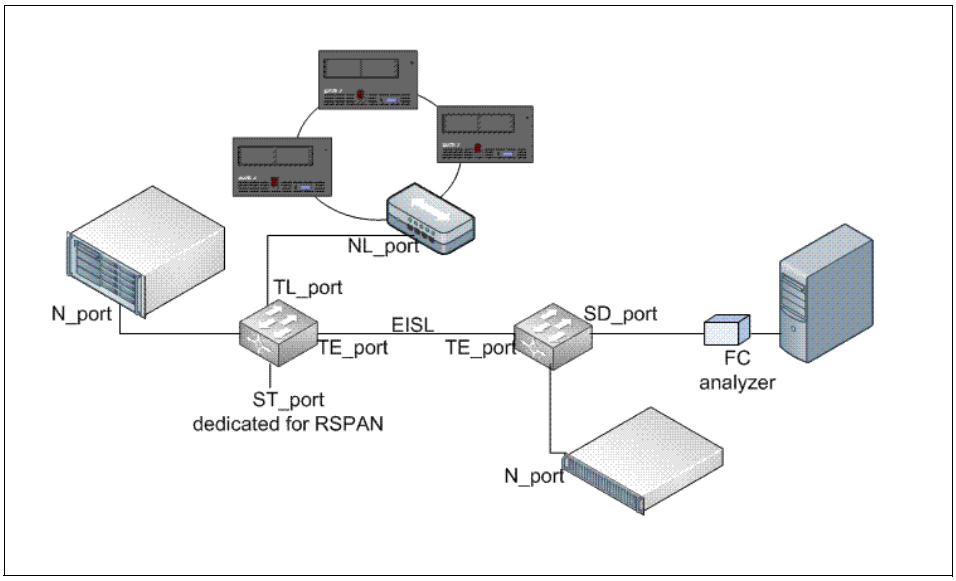

Figure 5-14 CISCO specific Fibre Channel ports

5.3 Addressing

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

5.3.1 World Wide Name

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

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

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

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

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

Both the old and new WWN formats are shown in Figure 5-15 on page 96.

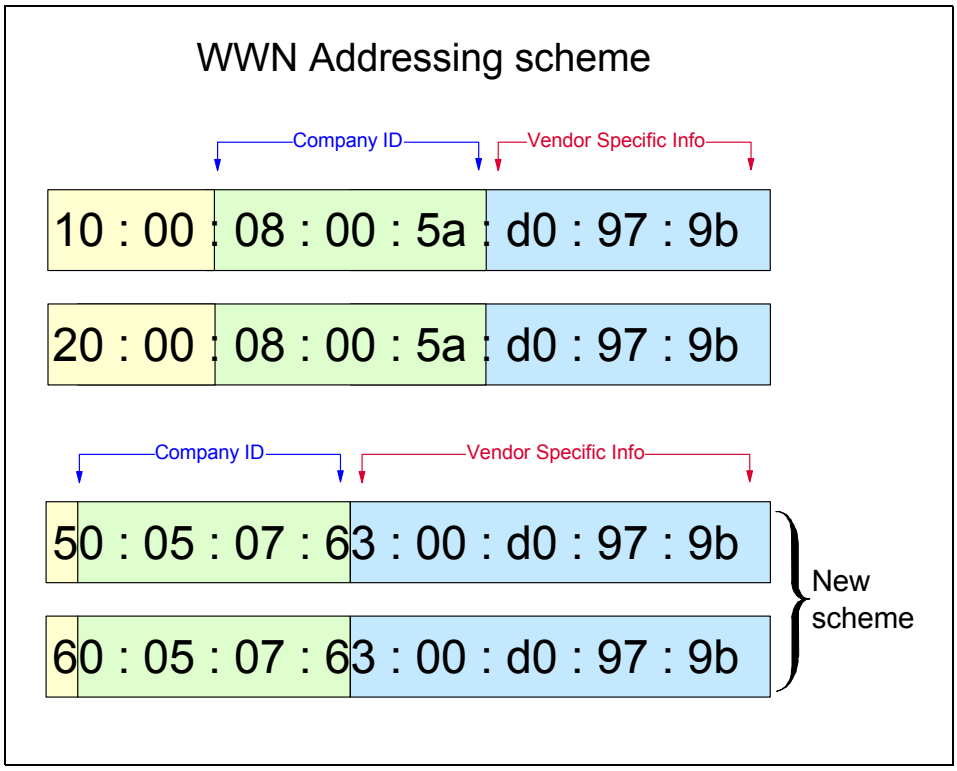

Figure 5-15 World Wide Name addressing scheme

The new addressing scheme starts with a hex 5 or 6 in the first half-byte followed by the vendor identifier in the next 3 bytes. The vendor-specific information is then contained in the following fields. Both these formats are currently in use and depends on the hardware manufacturer standards to follow either of the formats. However the Vendor ID / company ID are assigned uninque by the IEEE standards and each vendor and their identifier can be found in this text file:

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

A worldwide node name (WWNN) is a globally unique 64-bit identifier assigned to each Fibre Channel *node or device*. For servers/ hosts, the WWNN is unique for each HBA, and in a case of a server with two HBAs they have two WWNN. For a SAN switch the WWNN will be common for the chassis, and for storage the WWNN is common for each controller unit of midrange storage, and in a case of high-end enterprise storage the WWNN is unique for the entire array.

A worldwide port number WWPN is a unique identifier for each FC port of any Fibre Channel device. For server we have a WWPN for each port of the HBA, for a switch the WWPN is available for each port in the chassis, and for storage each host port has an individual WWPN.

Server WWNN and WWPN

Figure 5-16 on page 97 indicates the WWNN is for every HBA and every port in the HBA will have an individual WWPN.

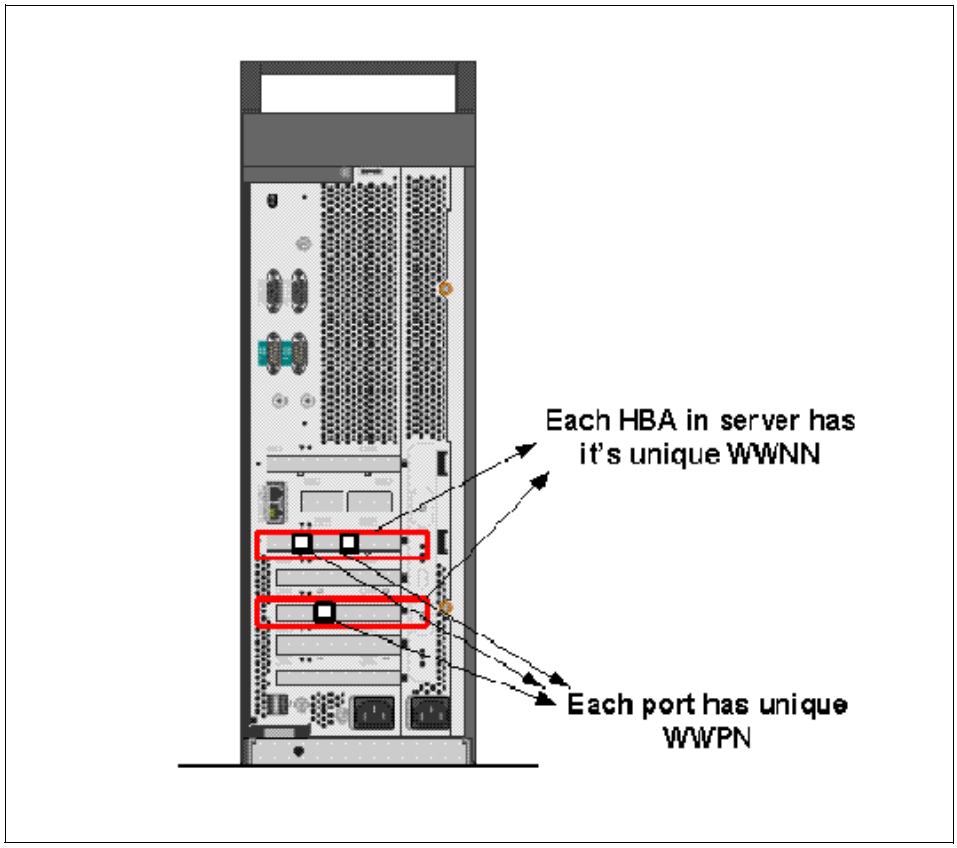

Figure 5-16 Server WWNN and WWPN

SAN WWNN and WWPN

Figure 5-17 on page 98 indicates the WWNN is for the entire SAN switch chassis and the WWPN is for each FC port in the SAN switch chassis.

Note: The new 16 Gbps b-type switches with FOS 7.0 could also have a virtual WWPN defined by switches called Fabric Assigned PWWWNs - (FAPWWN).These FAPWWN can be used for pre-configuring zoning before physical servers are connected. This feature helps to simplify and accelerate server deployment and improve operational efficiency by avoiding the wait time for physical connectivity to be done. This feature also requires servers to be using Brocade HBAs/Adapters with HBA driver version 3.0.0.0 or higher which can be configured to use FAPWWN.

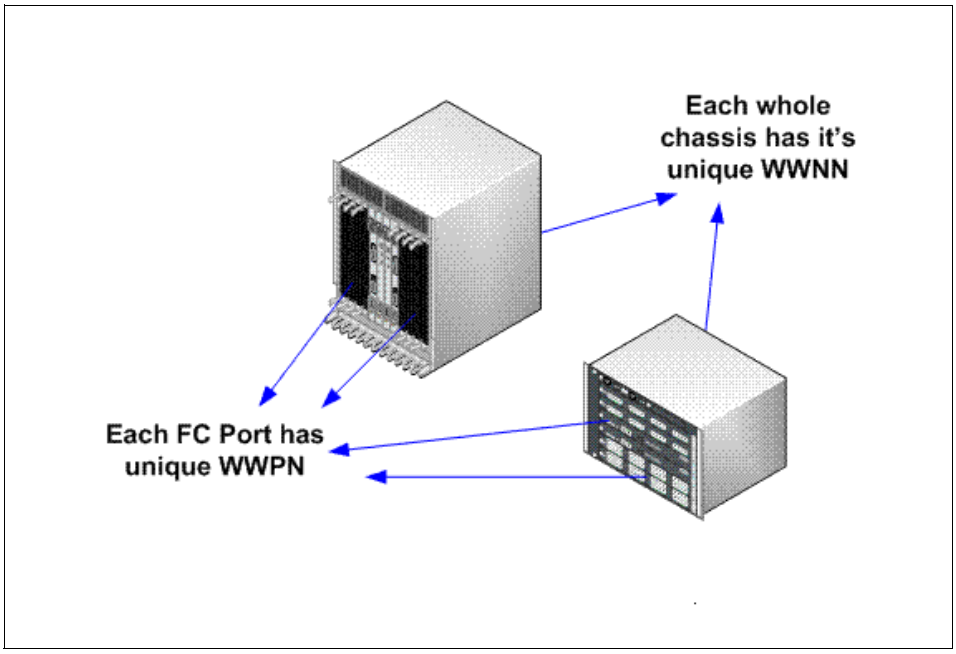

Figure 5-17 SAN switch WWNN and WWPN

Storage WWNN and WWPN

Disk storage has an individual WWNN for the entire storage system and the individual FC host ports will have a unique WWPN as indicated in Figure 5-18 on page 98 which shows a dual controller module.

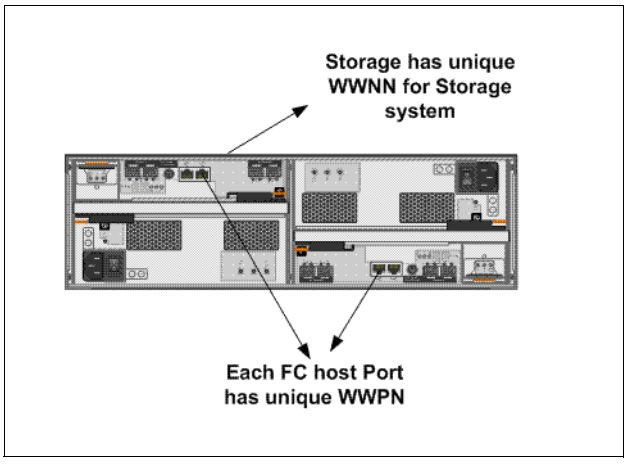

Figure 5-18 Storage WWNN and WWPN

Note: The IBM virtualization storage systems have a different WWNN usage, for example, each node in a SAN Volume Controller (SVC) or the IBM Storwize® V7000 will have an individual and unique WWNN.

For the DS8000® each Storage Facility Image will have a unique individual WWNN.

5.3.2 Tape Device WWNN and WWPN

For tape devices, each drive inside the tape library will have an individual WWPN and WWNN. Figure 5-19 on page 99 indicates multiple drive libraries will have a individual WWNN and WWPN for each drive.

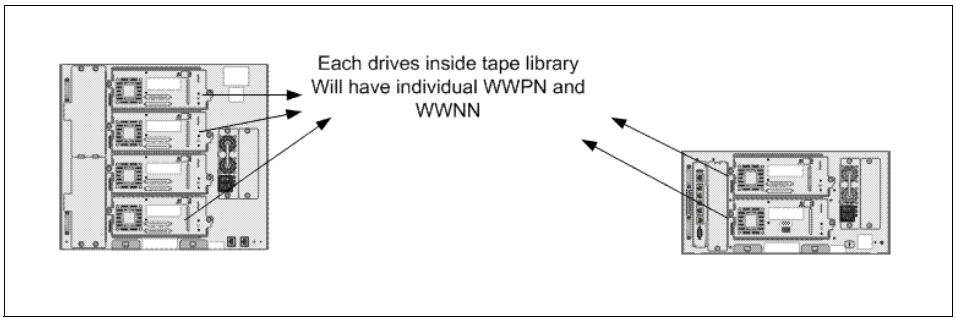

Figure 5-19 Tape Device WWNN and WWPN

5.3.3 Port address

Because of the potential impact on routing performance by using 64-bit addressing, there is another addressing scheme used in Fibre Channel networks. This scheme is used to address ports in the switched fabric. Each port in the switched fabric has its own unique 24-bit address. With this 24-bit address scheme, we get a smaller frame header, and this can speed up the routing process. With this frame header and routing logic, the Fibre Channel is optimized for high-speed switching of frames.

With a 24-bit addressing scheme, this allows for up to 16 million addresses, which is an address space larger than any practical SAN design in existence in today's world. There needs to be some relationship between this 24-bit address and the 64-bit address associated with World Wide Names. We will explain this in the section that follows.

5.3.4 24-bit port address

The 24-bit address scheme removes the overhead of manual administration of addresses by allowing the topology itself to assign addresses. This is *not* like WWN addressing where the addresses are assigned to manufacturers by the IEEE standards committee and are built into the device at the time of manufacture. If the topology itself is assigning the 24-bit addresses, then something has to be responsible for maintaining the addressing scheme from WWN addressing to port addressing.

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

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

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

A 24-bit port address consists of three parts:

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

We show how the address is built up in Figure 5-20 on page 100.

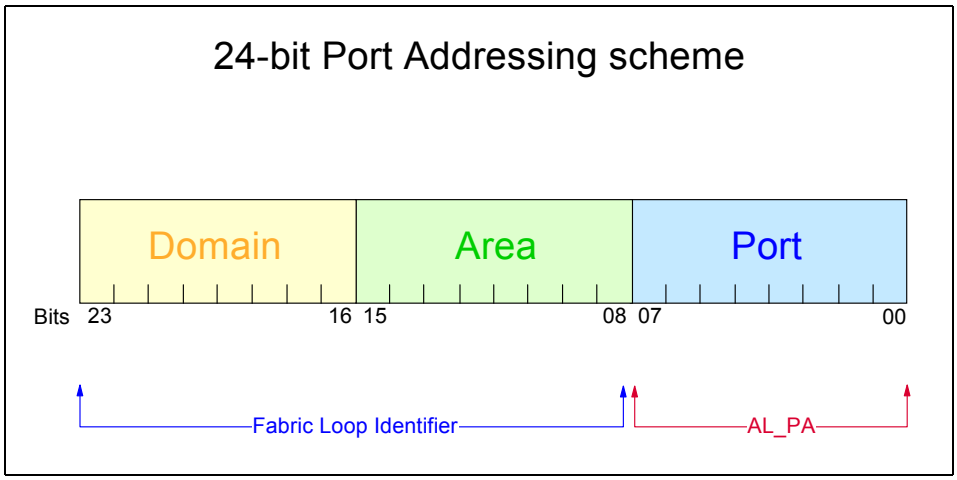

Figure 5-20 Fabric port address

The significance of some of the bits that make up the port address in the are:

- Domain

  The most significant byte of the port address is the domain. This is the address of the switch itself. A domain ID is a unique number that identifies the switch or director to a fabric. It can be either static or dynamic. Static (insistent) domain IDs are a requirement for FICON. Each manufacturer will have a range of numbers, and a maximum number of domain IDs that can be used in a fabric.

  One byte allows up to 256 possible addresses. Because some of these are reserved, as for the one for broadcast, there are only 239 addresses available. This means that you can theoretically have as many as 239 switches in your SAN environment. The domain number allows each switch to have a unique identifier if you have multiple interconnected switches in your environment.

- Area

  The area field provides 256 addresses. This part of the address is used to identify the individual Ports. Hence in order to have more than 256 ports in one switch in a director class of switches we have to follow the shared area addressing.

- Port

  The final part of the address provides 256 addresses for identifying attached N_Ports and NL_Ports.

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

```
Domain x Area x Ports
```

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

Depending on the fabric topology the fabric addressing format of device differs,

In a fabric topology, devices have an addressing format type of DDAA00. For example, the address 020300 indicates that the device belongs to the switch with domain id 02, that switch

is connected to port 03 and the ALPA address is 00 indicating this device is not a loop fabric device, ie. it is a switched fabric device. For any switched fabric device the ALPA ID will always be 00.

5.3.5 Loop address

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

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

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

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

5.3.6 b-type addressing modes

IBM b-type (IBM's OEM agreement with Brocade is referred to as b-type) has three different addressing modes: native mode, Core PID mode and shared area addressing mode.

Native mode was used in legacy switches which supported a maximum of 16 ports. This is because in native mode the fabric addressing format used is DD1A00, The area part of the fabric address will always have a prefix of 1 and hence it supports a port count from hexadecimal 10 to 1F (maximum of 16 ports).

Core PID mode is used to support a maximum of 256 ports per domain/switch. This is because in Core PID mode the area part of the fabric address will support addresses from hexadecimal 00 to FF (maximum of 256 ports). The fabric addressing format used for this is DDAA00.

Figure 5-21 on page 102 explains Native and core PID mode with example FC address of two devices.

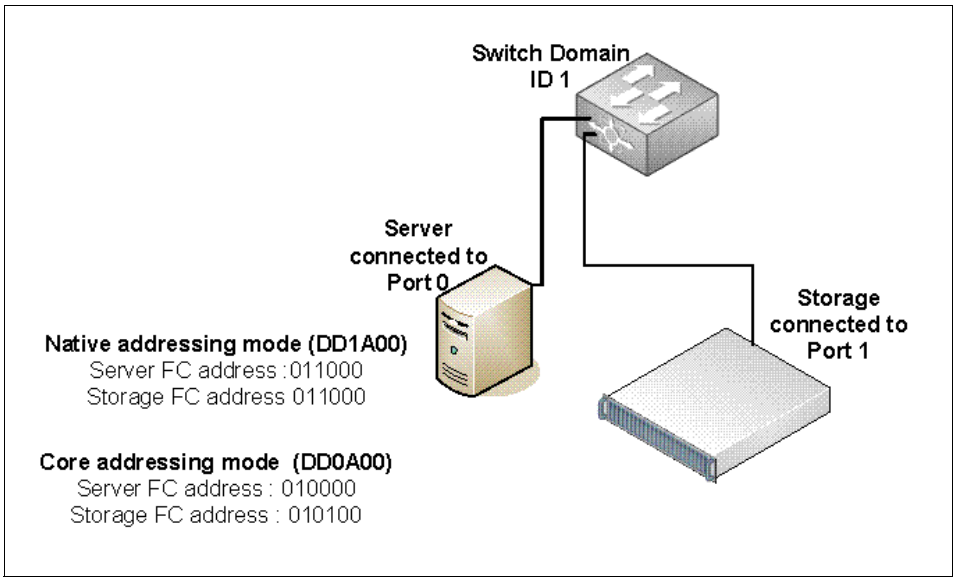

Figure 5-21 Native versus core addressing mode

Shared addressing mode is used when more than 256 ports are used in the same domain/switch. This mode is used in directors with high port density. The port addressing in these directors will use the same area numbers for two ports by having the third byte of the FC address (node addresses) as 80 for higher port numbers. By having the Area ID used more than once this mode enables more than 256 ports to exist in a single domain.

Figure 5-22 shows port 24, 25 shares same Area ID with the Port 32, 33 of FC4-48 port.

```
Index Slot Port Address Media Speed State
==============================================
168     3    24   01a800    --     N8   No_Module
169     3    25   01a900    --     N8   No_Module
<truncated output>

288     3    32   01a880    --     N8   No_Module
289     3    33   01a980    --     N8   No_Module
```

Figure 5-22 Shared addressing mode

5.3.7 FICON address

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

The Fibre Channel architecture does not specify how a server N_Port determines the destination port address of the storage device N_Port with which it requires communication. This is node and N_Port implementation dependent. Basically, there are two ways that a server can determine the address of the N_Port with which it wishes to communicate:

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

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

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

The FICON CU port will also perform a fabric login to obtain its 24-bit FC port address. Figure 5-23 shows the FC-FS 24-bit FC port address identifier is divided into three fields.

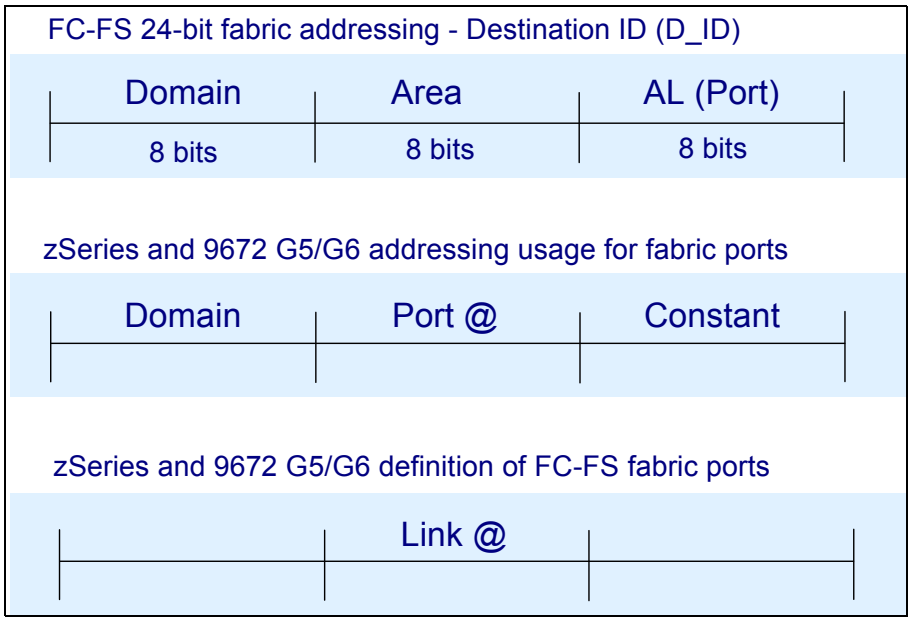

Figure 5-23 FICON port addressing

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

The zSeries and 9672 G5/G6 processors, when working with a switched point-to-point topology, require that the Domain and the AL_Port (Arbitrated Loop) field values be the same for all the FC F_Ports in the switch. Only the area field value will be different for each switch F_Port.

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

As shown in Figure 5-24, the eight bits for the domain address and the eight-bit constant field are provided from the Fabric Login initialization result, while the eight bits, one byte for the port address (1-byte Link address), are provided from the zSeries or 9672 G5/G6 CU link definition (using HCD and IOCP).

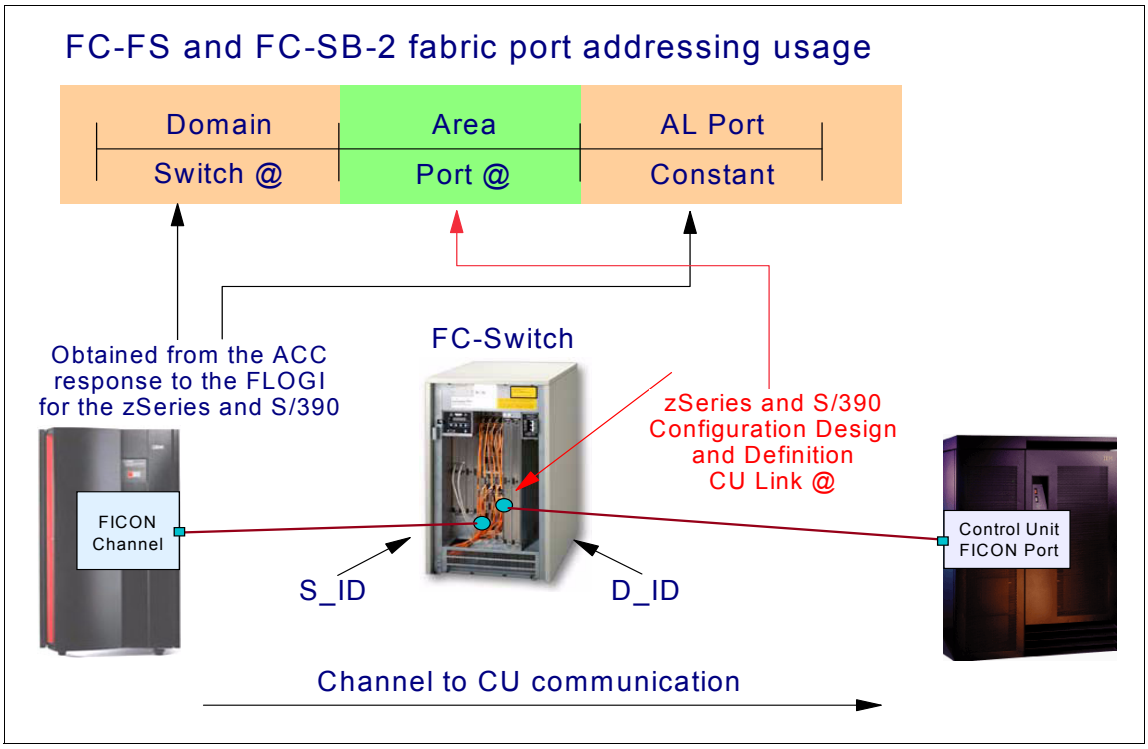

Figure 5-24 FICON single switch: Switched point-to-point link address

FICON address support for cascaded switches

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

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

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

Figure 5-25 on page 105 shows that the FC-FS 24-bit FC port address identifier is divided into three fields:

- Domain
- Area
- AL Port

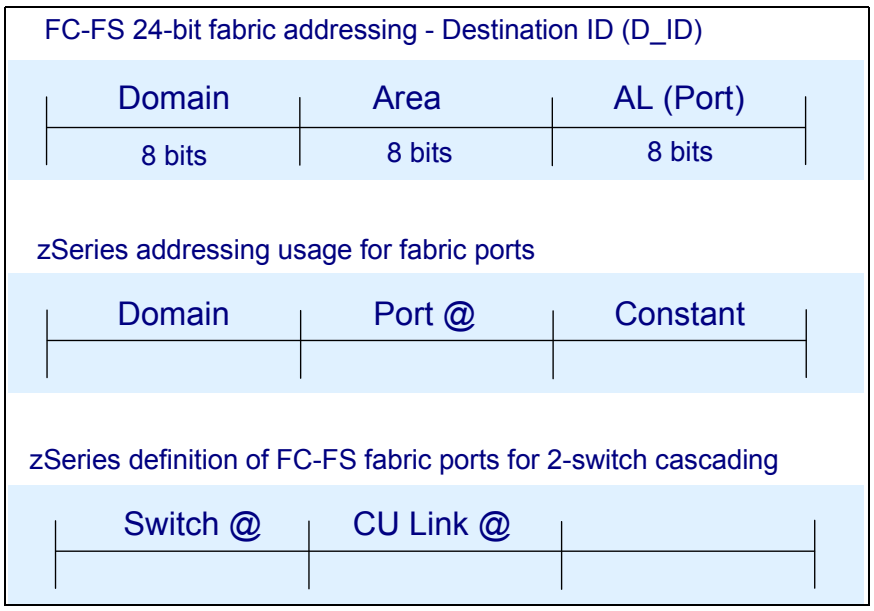

Figure 5-25 FICON addressing for cascaded directors

It shows the FC-FS 24-bit port address and the definition usage of that 24-bit address in a zSeries environment. Here, 16 bits making up the FC port address must be defined for the zSeries to access a FICON CU in a cascaded environment. The FICON channel in FICON native (FC) mode working with a cascaded FC topology, two-switch, provides the remaining byte making up the full three-byte FC port address of the CU to be accessed.

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

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

As shown in Figure 5-26 on page 106, the eight bits for the constant field are provided from the Fabric Login initialization result, while the 16 bits for the port address, two-byte Link address, are provided from the zSeries CU link definition using HCD and IOCP.

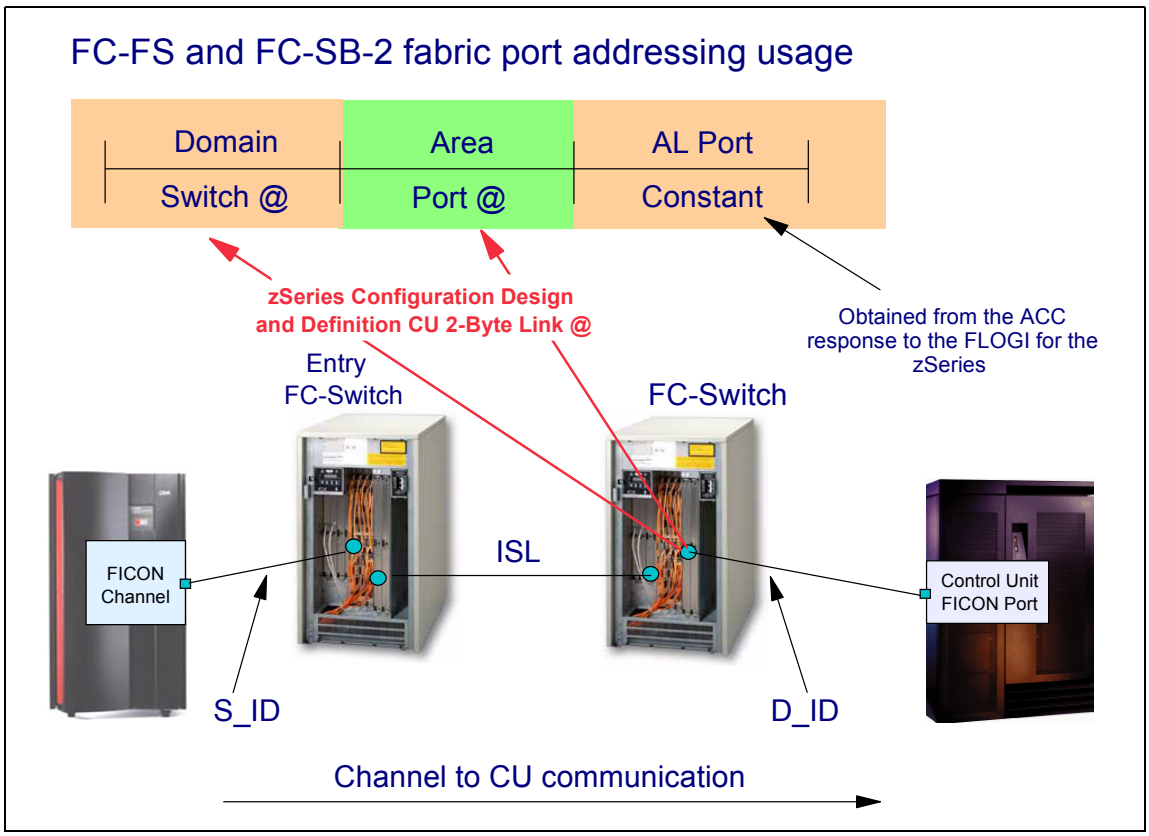

Figure 5-26 Two cascaded director FICON addressing

As a footnote, FCP connectivity is device-centric and is defined in the fabric using the WWPN of the devices that are allowed to communicate. When an FCP device attaches to the fabric, it queries the Name Server for the list of devices that it is allowed to form connections with (i.e. the zoning information). FICON devices do not query the Name Server for accessible devices because the allowable port/device relationships have been defined in the host, thus the zoning and Name Server information does not need to be retrieved.

5.4 Fibre Channel Arbitrated Loop protocols

To support the shared behavior of Fibre Channel Arbitrated Loop (FC-AL), a number of loop-specific protocols are used. These protocols are used to:

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

We discuss some of these topics in the sections that follow.

5.4.1 Fairness algorithm

The way that the fairness algorithm works is based around the IDLE ordered set, and the way that arbitration is carried out. In order to determine that the loop is not in use, an NL_Port

waits until it sees an IDLE go by and it can arbitrate for the loop by sending an RB Primitive Signal ordered set. If a higher priority device arbitrates before the first NL_Port sees its own ARB come by, then it loses the arbitration; but if it sees that its own ARB has gone all the way around the loop, then it has won arbitration. It can then open a communication to another NL_Port. When it has finished, it can close the connection and either arbitrate for the loop or send one or more IDLEs. If it complies with the fairness algorithm then it will take the option of sending IDLEs. That will force lower priority NL_Ports to successfully arbitrate for sending IDLEs, and that will allow lower priority NL_Ports to successfully arbitrate for the loop. However, there is no rule that forces any device to operate the fairness algorithm.

5.4.2 Loop addressing

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

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

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

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

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

Arbitrated Loop will assign priority to AL_PAs, based on numeric value. The lower the numeric value, the higher the priority is.

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

Note: System z9® and zSeries servers do not support the arbitrated loop topology.

5.5 Fibre Channel port initialization and fabric services

We had seen that there are different port types. At a very high level, port initialization starts with port type detection, then speed and active state detection where the speed is negotiated according to the device that is connected, and then the port initializes to an active state. In this active state every F port/FL port that has an N port/NL port connected, the Extended Link Service (ELS) and Fibre Channel Common Transport (FCCT) protocol are used for further switch port to node port communication. Only after this initialization completes can data flow happen.In this section let us see the services responsible for the port initialisation in a fabric switch.

The three login types for fabric devices are:

- Fabric login (FLOGI)
- Port login (PLOGI)
- Process login (PRLI)

Apart from these login types we will also describe the roles of other fabric services such as the fabric controller, management server and time server

5.5.1 Fabric login

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

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

The switch accepts the login and returns an accept (ACC) frame to the sender. If some of the service parameters requested by the N_Port or NL_Port are not supported the switch will set the appropriate bits in the ACC frame to indicate this.

NL_Ports derives their AL_PA during the loop initialization process (LIP). The switch then decides if it will accept this AL_PA, if it does not conflict with any previously assigned AL_PA on the loop. If not, a new AL_PA is assigned to the NL_Port, which then causes the start of another LIP.

Figure 5-27 on page 109 shows nodes performing FLOGI.

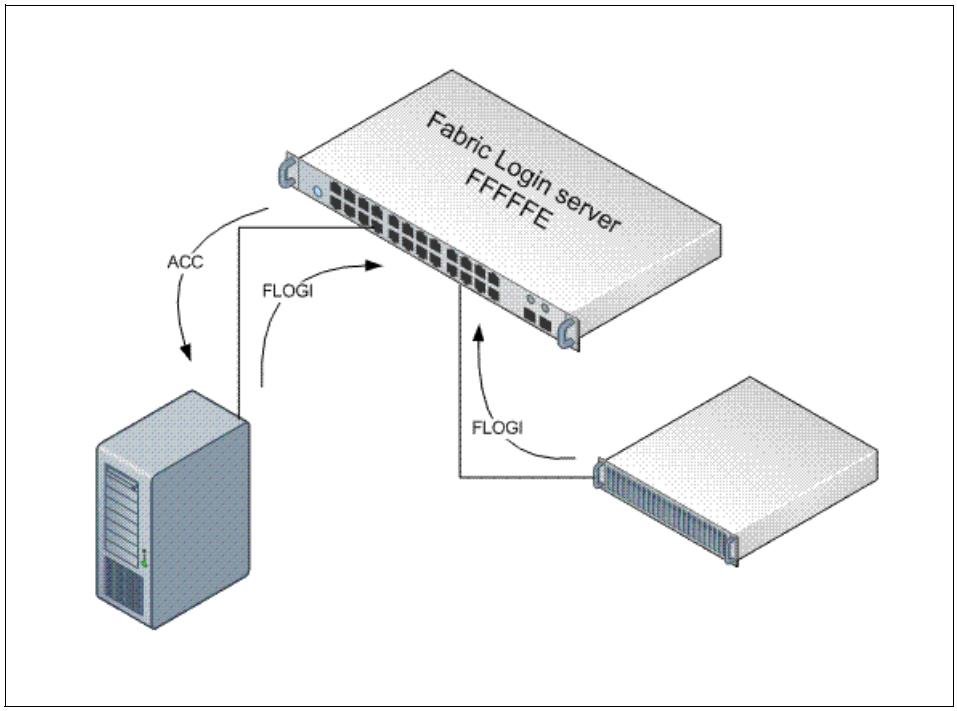

Figure 5-27 FLOGI of nodes

5.5.2 Port login (PLOGI)

Port login (PLOGI), is used to establish a session between two N_Ports and is necessary before any upper level commands or operations can be performed. During port login, two N_Ports (devices) swap service parameters and make themselves known to each other by performing a port login to a well known address of 0xFFFFFC. The device may register values for all or some of its objects but the most useful are:

- 24 bit port address
- 64 bit port name
- 64 bit node name
- Buffer to buffer credit capability
- Maximum frame size
- Class of service parameters
- FC-4 protocols supported
- Port type

Once the communication parameters and identities of other devices are discovered they are able to establish logical sessions between devices (initiator and targets).

Figure 5-28 on page 110 shows the PLOGI of a host.

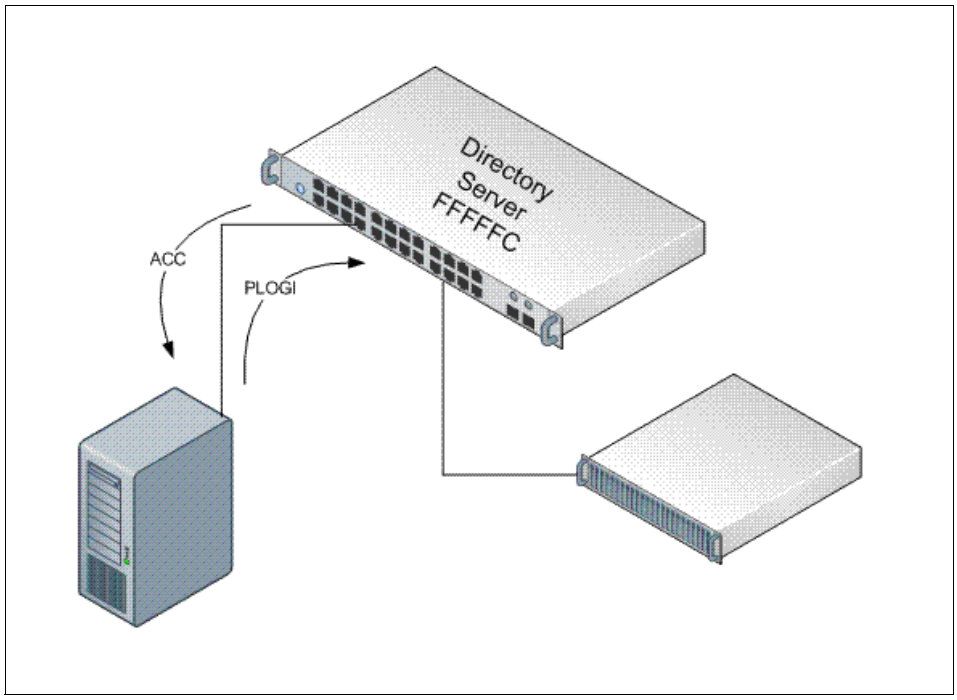

Figure 5-28 Node PLOGI to probe other nodes in fabric

5.5.3 Process login (PRLI)

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

Figure 5-29 on page 111 indicates the PRLI from server to storage.

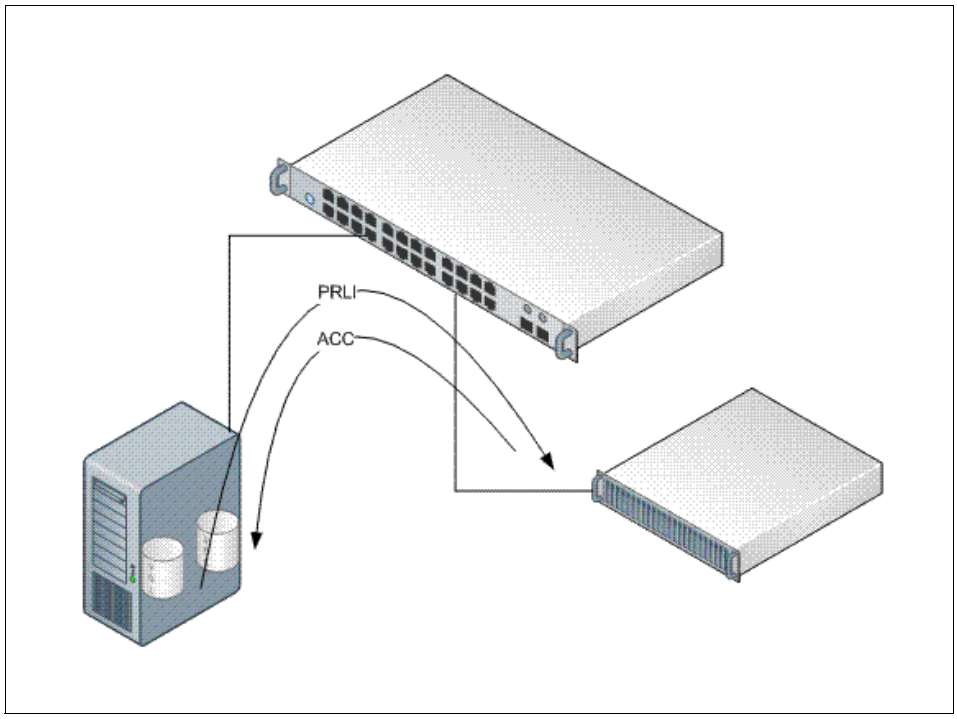

Figure 5-29 PRLI request from initiator to Target

5.6 Fabric services

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

- Management services
- Time services
- Simple name server
- Login services
- Registered State Change Notification (RSCN)

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

All these services are addressed by FC-2 frames and are accessed by so called well-known addresses.

5.6.1 Management server

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

5.6.2 Time server

The Time Service or Time Server is provided to serve time information that is sufficient for managing expiration time. This service is provided at the well-known address identifier, 0xFFFFFB.

The functional model of the Time Server consists of primarily two entities:

- Time Service Application: the entity representing a user accessing the Time Service
- Time Server: the entity that provides the time information through the Time Service.

There may be more than one distributed Time Server instance within the Fibre Channel network. However, from a user's perspective, the Time Service appears to come from the entity that is accessible at the Time Service well-known address identifier. If the Time Service is distributed, it will be transparent to the Application.

5.6.3 Simple name server

Fabric switches implement a concept known as the simple name server (SNS). All switches in the fabric keep the SNS updated, and are therefore aware of all devices in the SNS. After a node has successfully logged into the fabric, it performs a PLOGI into a well-known address of 0xFFFFFC. This allows it to register itself and pass on critical information such as class of service parameters, its WWN/address, and the upper layer protocols that it can support.

5.6.4 Fabric login server

In order to do a fabric login, a node communicates with the fabric login server at the well-known address 0xFFFFFE.

5.6.5 Registered State Change Notification service

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

The nodes registers to the fabric controller with a state change registration (SCR) frame. The fabric controller, which maintains the fabric state with all registered device details, alerts registered devices with a registered state change notification (RSCN). This alert is sent whenever there is any device added or removed, a zone changed, switch IP or name change, and so on. The fabric controller has a well-known address of 0xFFFFFD.

5.7 Routing mechanisms

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

5.7.1 Spanning tree

In case of failure, it is important to consider having an alternative path between source and destination available. This will allow data to still reach its destination. However, having different paths available could lead to the delivery of frames being out of the order, due to frame taking a different path and arriving earlier than one of its predecessors.

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

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

5.7.2 Fabric shortest path first

According to the FC-SW-2 standard, fabric shortest path first (FSPF) is a link state path selection protocol. The concepts used in FSPF were first proposed by Brocade, and have since been incorporated into the FC-SW-2 standard. Since then it has been adopted by most, if not all, manufacturers.

What FSPF is

FSPF keeps track of the links on all switches in the fabric and associates a cost with each link. The cost is always calculated as being directly proportional to the number of hops. The protocol computes paths from a switch to all other switches in the fabric by adding the cost of all links traversed by the path, and choosing the path that minimizes the cost.

How FSPF works

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

How FSPF helps

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

- Have the same latency

- Operate at the same speed
- Have no congestion

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

5.8 Zoning

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

For example, it might be desirable to separate a Microsoft Windows NT environment from a UNIX environment. This is very useful because of the manner in which Windows attempts to claim all available storage for itself. Because not all storage devices are capable of protecting their resources from any host seeking available resources, it makes sound business sense to protect the environment in another manner. We show an example of zoning in Figure 5-30 on page 114 where we have separated AIX from NT and created Zone 1 and Zone 2. This diagram also shows how a device can be in more than one zone.

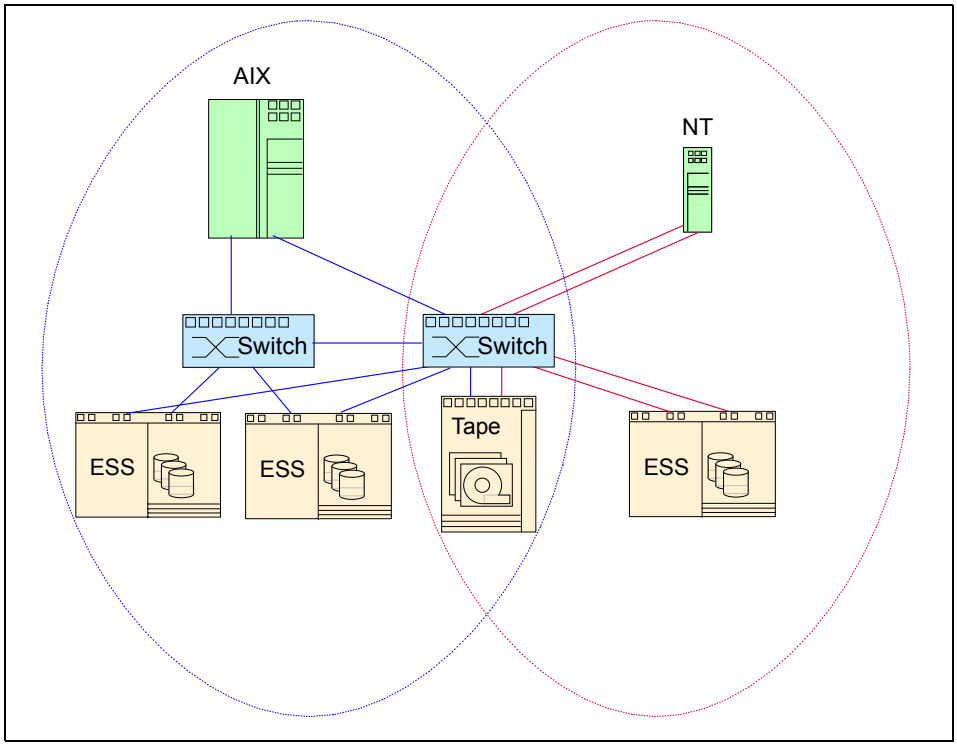

Figure 5-30 Zoning

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

An example of zoning is shown in Figure 5-31 on page 115. In this case:

- Server A and Storage A can communicate with each other.
- Server B and Storage B can communicate with each other.

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

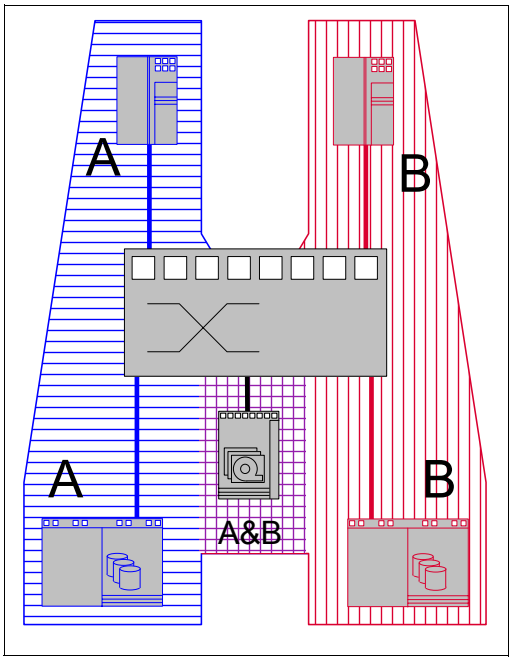

Figure 5-31 An example of zoning

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

Zoning can be implemented in two ways:

- Hardware zoning
- Software zoning

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

5.8.1 Hardware zoning

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

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

Figure 5-32 shows an example of zoning based on the switch port numbers.

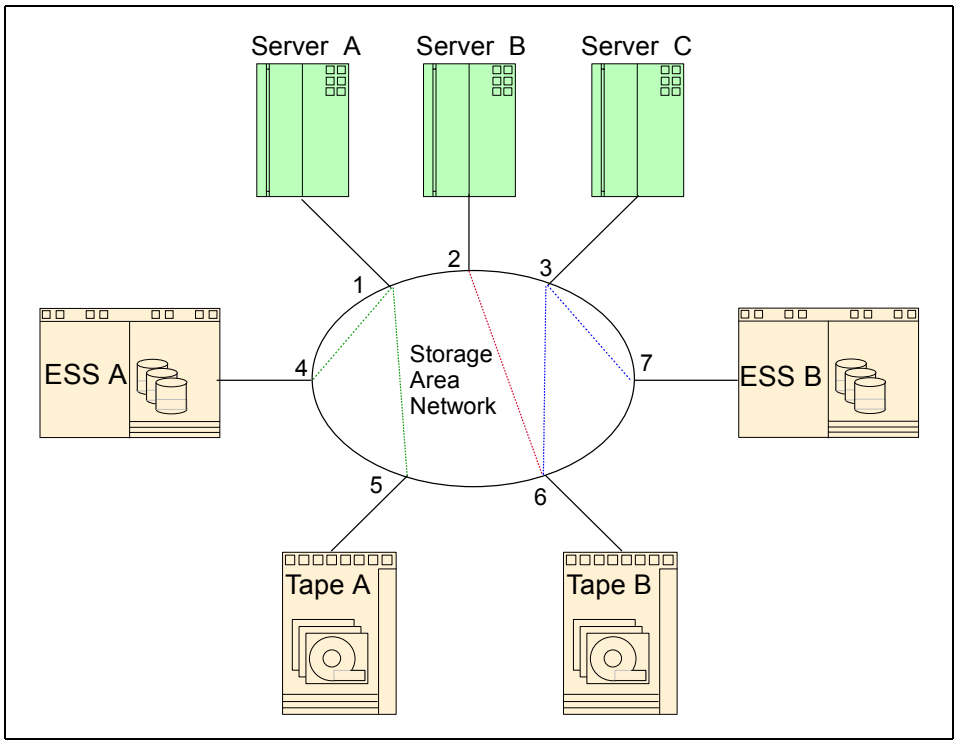

Figure 5-32 Zoning based on the switch port number

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

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

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

A single port can also belong to multiple zones.

We show an example of hardware zoning in Figure 5-33 on page 117. This example illustrates another way of considering the hardware zoning as an array of connections.

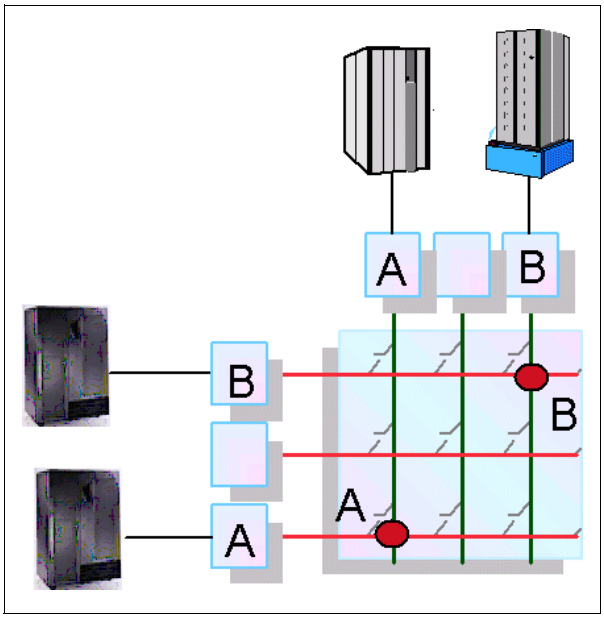

Figure 5-33 Hardware zoning

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

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

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

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

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

5.8.2 Software zoning

Software zoning is implemented by the fabric operating systems within the fabric switches. They are almost always implemented by a combination of the name server and the Fibre Channel Protocol. when a port contacts the name server, the name server will only reply with information about ports in the same zone as the requesting port. A soft zone, or software zone, is not enforced by hardware. What this means is that if a frame is incorrectly delivered (addressed) to a port that it was not intended to, then it will be delivered to that port. This is in contrast to hard zones.

When using software zoning the members of the zone can be defined using their World Wide Names:

- Node WWN
- Port WWN

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

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

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

Important: However, by stating this, it does not automatically mean that if you unplug a device, such as a disk subsystem, and plug it into another switch port, that your host will still be able to communicate with your disks (until you either reboot or unload and load your operating system device definitions), even if the device remains member of that particular zone. This depends on components you use in your environment, like operating system and multipath software.

Shown in Figure 5-34 is an example of WWN-based zoning. In this example, symbolic names are defined for each WWN in the SAN to implement the same zoning requirements, as shown in the previous Figure 5-32 on page 116 for port zoning:

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

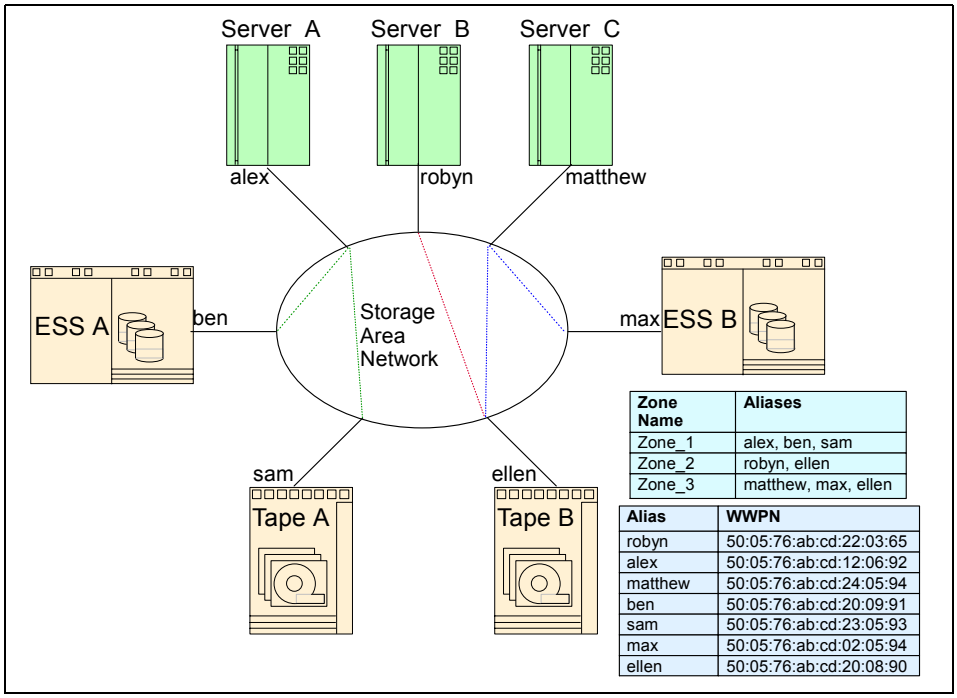

Figure 5-34 Zoning based on the devices' WWNs

There are some potential security leaks with software zoning:

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

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

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

Frame filtering

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

5.8.3 LUN masking

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

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

It is absolutely possible for more than one host to see the same storage device or LUN. This is potentially a problem, both from a practical and a security perspective. Another approach to securing storage devices from hosts wanting to take over already assigned resources is logical unit number (LUN) masking. Every storage device offers its resources to the hosts by means of LUNs.

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

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

6

SAN as a service for Cloud Computing

While information can be your greatest asset, it can also be your greatest challenge as you struggle to keep up with explosive data growth. More data means more storage and more pressure to install another rack into the data center.

Cloud Computing offers a new way of solution provisioning with significant cost savings and high reliability.

6.1 What is a Cloud?

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. Figure 6-1 shows an overview of cloud computing.

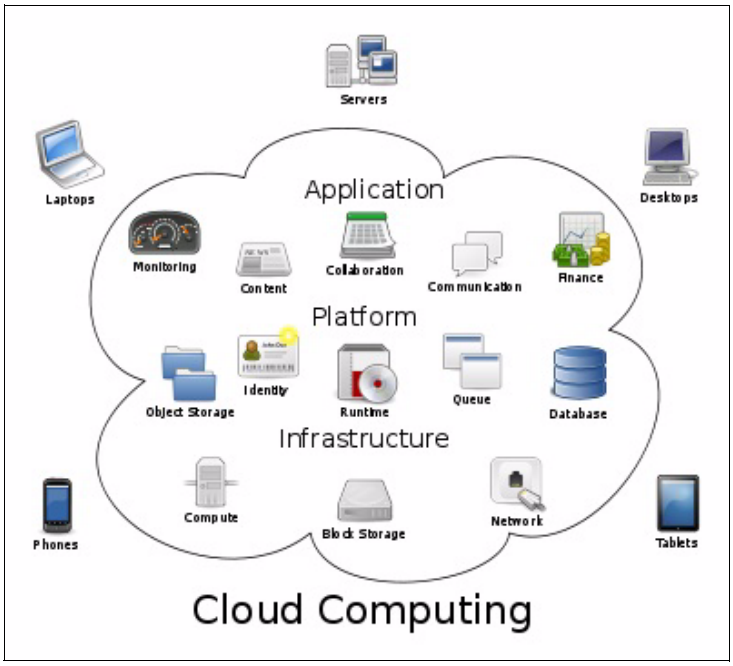

Figure 6-1 Cloud computing overview

Cloud computing provides computation, software, data access, and storage services that do not require end-user knowledge of the physical location and configuration of the system that delivers the services. Parallels to this concept can be drawn with the electricity grid, wherein end-users consume power without needing to understand the component devices or infrastructure required to provide the service.

Cloud computing describes a new consumption, and delivery model for IT services, and it typically involves provisioning of dynamically scalable and virtualized resources. The cloud introduces three key concepts, cost saving, service reliability, and infrastructure flexibility.

To cater to the increasing, on-demand needs of business, IT services and infrastructures are moving rapidly towards a flexible utility and consumer model by adopting new technologies.

One of these technologies is virtualization. Cloud computing is an example of a virtual, flexible delivery model. Inspired by consumer Internet services, cloud computing puts the user in the “driver’s seat,” that is, users can consume Internet offerings and services by using this self-service, on-demand model.

Cloud computing has the potential to make an enormous impact to your business by providing the following benefits:

- Reducing IT labor costs for configuration, operations, management, and monitoring
- Improving capital utilization and significantly reducing license costs
- Reducing provisioning cycle times from weeks to minutes

- Improving quality and eliminating many software defects
- Reducing end-user IT support costs

From a technical perspective cloud computing enables, amongst others, these capabilities:

- Abstraction of resources
- Dynamic right-sizing
- Rapid provisioning

6.1.1 Private and public cloud

A cloud can be private or public. A public cloud sells services to anyone on the Internet. A private cloud is a proprietary network or a data center that supplies hosted services to a limited number of people. When a service provider uses public cloud resources to create their private cloud, the result is called a virtual private cloud. Private or public, the goal of cloud computing is to provide easy, scalable access to computing resources and IT services. A cloud has four basic components as shown in Figure 6-2 on page 123.

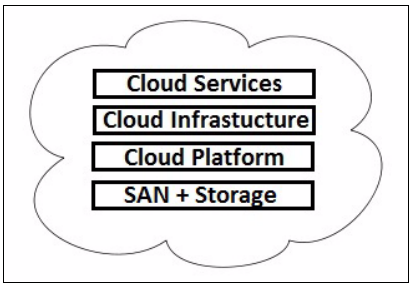

Figure 6-2 Cloud computing components

6.1.2 Cloud computing components

We describe the cloud computing components, or layers, in our model.

Cloud Services

This layer is the service delivered to the customer, it could be an application, a desktop, a server or disk storage space, and as we mentioned earlier the customer does not need to know where or how their service is running, they just use it.

Cloud Infrastructure

This layer could be difficult to visualize depending on the final delivered service. If the final service is a chat application the cloud infrastructure will be the servers on which the chat application is running. In the other case if the final service is a virtualized server the cloud infrastructure will be all the other servers required to provide “a server” as a service to the customer such as the DNS, security services and management.

Cloud Platform

This layer consists of the selected platform to build the cloud, and there are many vendors like IBM Smart Business Storage Cloud, VMware vSphere, Microsoft Hyper V, Citrix Xen Server which are well known cloud solutions in the market.

SAN + Storage

This layer is where information flows and lives. Without it nothing could happen. Depending on the cloud design the storage can be any of the previously presented solutions such as

DAS, NAS, iSCSI, SAN, and/or FCoE. For the purposes of this book we will talk about Fibre Channel or Fibre Channel over Ethernet for networking and compatible storage devices.

6.1.3 Cloud Computing Models

While cloud computing is still a relatively new technology, there are generally three cloud service models, each with a unique focus. The American National Institute of Standards and Technology (NIST) has defined the following cloud service models:

- *Software-as-a-Service (SaaS):* "The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through a thin client interface, such as a web browser (for example, web-based email). The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings."
- *Platform-as-a-Service (PaaS):* "The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure, including network, servers, operating systems, or storage, but has control over the deployed applications and possibly application hosting environment configurations."
- *Infrastructure-as-a-Service (IaaS):* "The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components (for example, hosts)."

Figure 6-3 shows these cloud models.

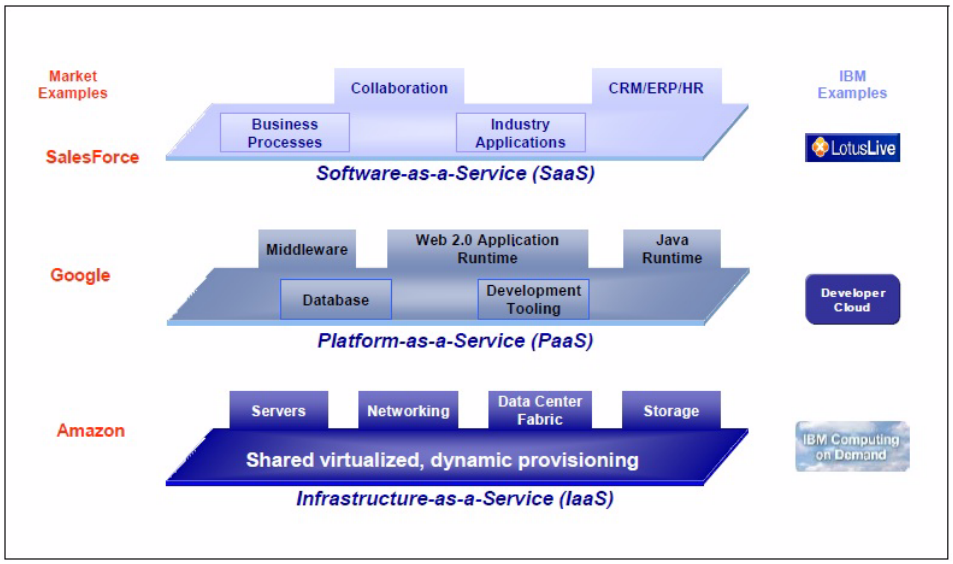

Figure 6-3 Examples of SaaS, PaaS, and IaaS services

In addition, NIST has also defined the following models for deploying cloud services:

- *Private cloud:* "The cloud infrastructure is owned or leased by a single organization and is operated solely for that organization."

- *Community cloud:* "The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (for example, mission, security requirements, policy, and compliance considerations)."
- *Public cloud:* "The cloud infrastructure is owned by an organization selling cloud services to the general public or to a large industry group."
- *Hybrid cloud:* "The cloud infrastructure is a composition of two or more clouds (internal, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability (for example, cloud bursting)."

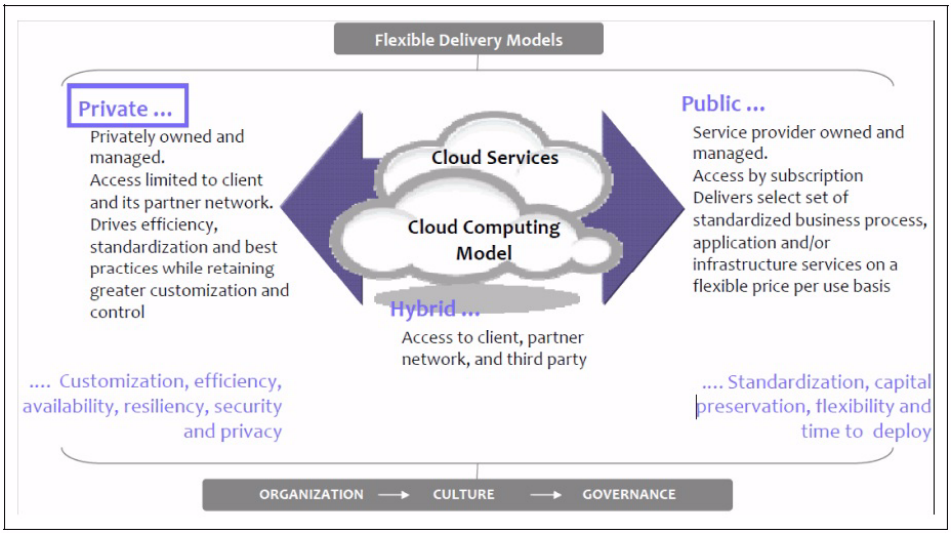

Figure 6-4 Cloud computing deployment models

From a storage perspective, IBM clients, based on their business requirements, can choose to adopt either a pubic or private storage cloud. These storage clouds are defined as follows:

- *Public storage cloud:* This is designed for clients who do not want to own, manage, or maintain the storage environment, thus reducing their capital and operational expenditures for storage. IBM dictates the choice of technology and cloud location, shared infrastructure with variable monthly charges, dynamic physical capacity at the customer level, and security measures to isolate customer data. The public storage cloud allows for variable billing options and shared tenancy of the storage cloud, giving customers the flexibility to manage the use and growth of their storage needs. This is the industry standard view of a storage cloud offering and is comparable to storage cloud offerings by other vendors.
- *Private storage cloud:* With a private storage cloud, customers have the choice of technology and location on a dedicated infrastructure with fixed monthly charges and physical capacity manageable by the customer. Each application can utilize dynamic capacity by sharing the cloud storage among multiple applications.

Private storage cloud solution technology and services from IBM address multiple areas of functionality. For more information visit:

http://www.ibm.com/cloud-computing/us/en/

6.2 Virtualization and the Cloud

When people talk about virtualization, they're usually referring to server virtualization, which means partitioning one physical server into several virtual servers, or machines. Each virtual machine can interact independently with other devices, applications, data and users as though it were a separate physical resource.

Different virtual machines can run different operating systems and multiple applications while sharing the resources of a single physical computer. And, because each virtual machine is isolated from other virtualized machines, if one crashes, it doesn't affect the others.

Hypervisor software is the secret sauce that makes virtualization possible. This software sits between the hardware and the operating system, and de-couples the operating system and applications from the hardware. The hypervisor assigns the amount of access that the operating systems and applications have with the processor and other hardware resources, such as memory and disk input/output.

In addition to using virtualization technology to partition one machine into several virtual machines, you can also use virtualization solutions to combine multiple physical resources into a single virtual resource. A good example of this is storage virtualization, where multiple network storage resources are pooled into what appears as a single storage device for easier and more efficient management of these resources. Other types of virtualization you may hear about include.

- Network virtualization splits available bandwidth in a network into independent channels that can be assigned to specific servers or devices.
- Application virtualization separates applications from the hardware and the operating system, putting them in a container that can be relocated without disrupting other systems.
- Desktop virtualization enables a centralized server to deliver and manage individualized desktops remotely. This gives users a full client experience, but lets IT staff provision, manage, upgrade and patch them virtually, instead of physically

Virtualization was first introduced in the 1960s by IBM to boost utilization of large, expensive mainframe systems by partitioning them into logical, separate virtual machines that could run multiple applications and processes at the same time. In the 1980s and 1990s, this centrally shared mainframe model gave way to a distributed, client-server computing model, in which many low-cost x86 servers and desktops independently run specific applications.

6.2.1 Cloud infrastructure virtualization

This consists of virtualizing three key parts: servers, desktops or applications. The virtualization concept used for servers and desktops is almost the same but for applications is different.

Virtualizing servers and desktops basically takes physical computers and makes them virtual. In order to make this possible a cloud platform is required, and we show the traditional physical environment in Figure 6-5 where one application maps to one operating system, and one OS to one physical server, and one physical server to one storage.

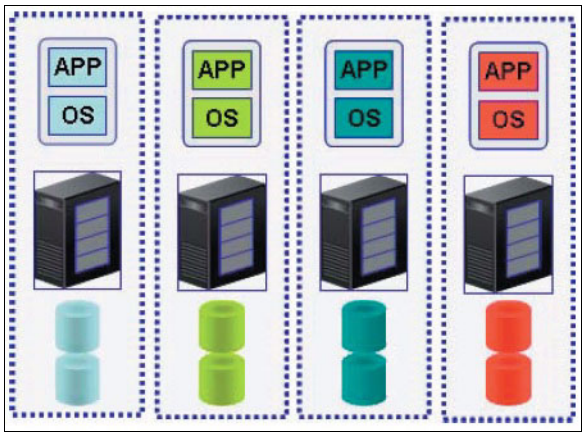

Figure 6-5 Traditional physical environment model

6.2.2 Cloud Platforms

There needs to be a platform that can handle putting multiple virtual servers into a single physical computer, and this platform is called the hypervisor, This is a layer in the computer stack between the virtual and physical components.

There are four core concepts in virtualization: encapsulation, isolation, partitioning, and hardware independence.

- Encapsulation: the entire machine becomes a set of files, and these files contain the operating system and application files plus the virtual machine configuration. The virtual machine files can be managed the same way you manage other files.
- Isolation: VMs running on a hardware platform cannot see or affect each other, so multiple applications can be run securely on a single server.
- Partitioning: VMware, for example, divides and actively manages the physical resources in the server to maintain optimum allocation.
- Hardware Independence: the hypervisor provides a layer between the operating systems and hardware, allowing hardware from multiple vendors to run on the same physical resource, as long as the server is on Hardware Compatibility List.

Figure 6-6 shows the virtualized environment.

Figure 6-6 Virtualized environment model

Server Virtualization

There are three popular approaches to server virtualization: the virtual machine model, the paravirtual machine model, and virtualization at the operating system layer.

Virtual machines are based on the host/guest paradigm. Each guest runs on a virtual implementation of the hardware layer. This approach allows the guest operating system to run without modifications. It also allows the administrator to create guests that use different operating systems. The guest has no knowledge of the host's operating system because it is not aware that it's not running on real hardware. It does, however, require real computing resources from the host so it uses a hypervisor to coordinate instructions to the CPU.

The paravirtual machine (PVM) model is also based on the host/guest paradigm and it uses a virtual machine monitor (VMM) too. In the paravirtual machine model, however, the VMM actually modifies the guest operating system's code. This modification is called porting. Porting supports the VMM so it can utilize privileged systems calls sparingly. Like virtual machines, paravirtual machines are capable of running multiple operating systems. Xen and UML both use the paravirtual machine model.

Virtualization at the OS level works a little differently. It isn't based on the host/guest paradigm. In the OS level model, the host runs a single OS kernel as its core and exports operating system functionality to each of the guests. Guests must use the same operating system as the host, although different distributions of the same system are allowed. This distributed architecture eliminates system calls between layers, which reduces CPU usage overhead. It also requires that each partition remain strictly isolated from its neighbors so that a failure or security breach in one partition isn't able to affect any of the other partitions. In this model, common binaries and libraries on the same physical machine can be shared, allowing an OS level virtual server to host thousands of guests at the same time. IBM AIX VIO and Solaris Zones both use OS-level virtualization.

Desktop Virtualization

This is sometimes referred to as client virtualization, and is defined as a virtualization technology that is used to separate a computer desktop environment from the physical computer. Desktop virtualization is considered a type of client-server computing model

because the virtualized desktop is stored on a centralized, or remote, server and not the physical machine being virtualized.

Desktop virtualization virtualizes desktop computers and these virtual desktop environments are “served” to users on the network. Users interact with a virtual desktop in the same way that a physical desktop is accessed and used. Another benefit of desktop virtualization is that is lets you remotely log in to access your desktop from any location.

One of the most popular uses of desktop virtualization is in the data center, where personalized desktop images for each user is hosted on a data center server.

There are also options for using hosted virtual desktops, where the desktop virtualization services is provided to a business through a third-party. The service provider will provide the managed desktop configuration, security, and storage-area network.

Application Virtualization

Just like desktop virtualization, where individual desktop sessions (OS and Applications) are virtualized and run from a centralized server, Application Virtualization virtualizes the Applications so that It can either be run from a centralized server or It can be streamed from a central server and run in an isolated environment in the desktop itself.

In the first type of application virtualization, the application image is loaded on to a central server and when a user requests the application it is streamed to an isolated environment on the user’s computer for execution. The application starts running shortly after it gets sufficient data to start running, and since the application is isolated from other applications, there may not be any conflicts. The applications that can be downloaded can be restricted based on the User ID which is established by logging in to corporate directories such as Active Directory (AD) or Lightweight Directory Access Protocol (LDAP).

In the second type of application virtualization, the applications are loaded as an image in remote servers and they are run (executed) in the servers itself, and only the on-screen information that is required to be seen by the end user is sent over the LAN. This is closer to desktop virtualization, but here only the application is virtualized instead of both the application and the operating system. The biggest advantage of this type of application virtualization is that it does not matter what the underlying Operating System is in the user’s computer as the applications are processed in the server. Another advantage is the effectiveness of mobile devices (mobile phones, tablet computers, and so on) that have lesser processing power while running processor hungry applications, as these are processed in the powerful processors of the servers.

6.2.3 Storage virtualization

Storage virtualization refers to the abstraction of storage systems from applications or computers. It is a foundation for the implementation of other technologies, such as thin provisioning, tiering, and data protection, which are transparent to the server.

These are some of the advantages of storage virtualization:

- Improved physical resource utilization: By consolidating and virtualizing storage systems, we can make more efficient use of previously wasted white spaces.
- Improved responsiveness and flexibility: De-coupling physical storage to virtual storage provides the ability to re-allocate resources dynamically, as required by the applications or storage subsystems.
- Lower total cost of ownership: Virtualized storage allows more to be done with the same or less storage.

Several types of storage virtualization are available.

Block level storage virtualization

This refers to provisioning storage to your operating systems or applications in the form of virtual disks. Fibre Channel (FC) and Internet Small Computer System Interface (iSCSI) are examples of protocols used by this type of storage virtualization.

There are two types of block level virtualization:

- Disk level virtualization: This is an abstraction process from a physical disk to a Logical Unit Number (LUN) that is presented as of it were a physical device
- Storage level virtualization: Unlike disk level virtualization, storage level virtualization hides the physical layer of Redundant Array of Independent Disks (RAID) controllers and disks, and hides and virtualizes the entire storage system.

File level storage virtualization

File level storage virtualization refers to provisioning storage volumes to operating systems or applications in the form of files and directories. Access to storage is by network protocols, such as Common Internet File Systems (CIFS) and Network File Systems (NFS). It is a file presentation in a single global namespace, regardless of the physical file location.

Tape virtualization

This refers to virtualization of tapes and tape drives using specialized hardware and software. This type of virtualization can enhance backup and restore flexibility and performance, because disk devices are used in the virtualization process, rather than tape media.

6.3 SAN Virtualization

For SAN virtualization, we describe the virtualization features available in the IBM System Networking portfolio. These features enable the SAN infrastructure to support the requirements of scalability, and consolidation and combine this with a lower TCO and a higher ROI.

- IBM b-type Virtual Fabrics
- CISCO Virtual SAN (VSAN)
- NPIV support for virtual nodes

6.3.1 IBM b-type Virtual Fabrics

The Virtual Fabric of the IBM b-type switches is a licensed feature which enables the logical partitioning of SAN switches. When Virtual Fabric is enabled a default logical switch using all ports is formed and this default logical switch can be then divided in to multiple logical switches by grouping them together at a port level. Figure 6-7 on page 131 indicates the flow of Virtual Fabric creation.

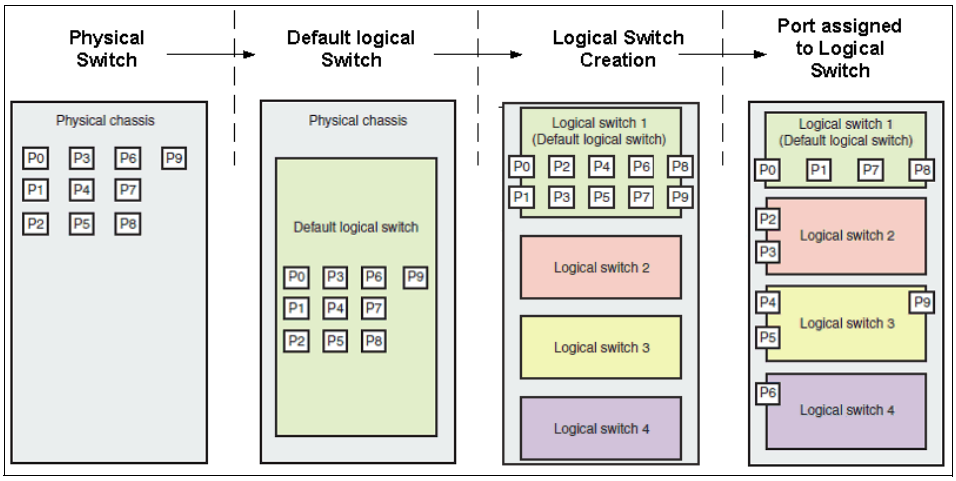

Figure 6-7 Virtual fabric creation

Logical fabric

When the fabric is formed with at least one logical switch it is called a logical fabric. The logical fabric has two different ways of fabric connectivity.

- A logical fabric connected with a dedicated ISL to another switch or a logical switch. Figure 6-8 shows a logical fabric formed between logical switches through a dedicated ISL for logical switches.

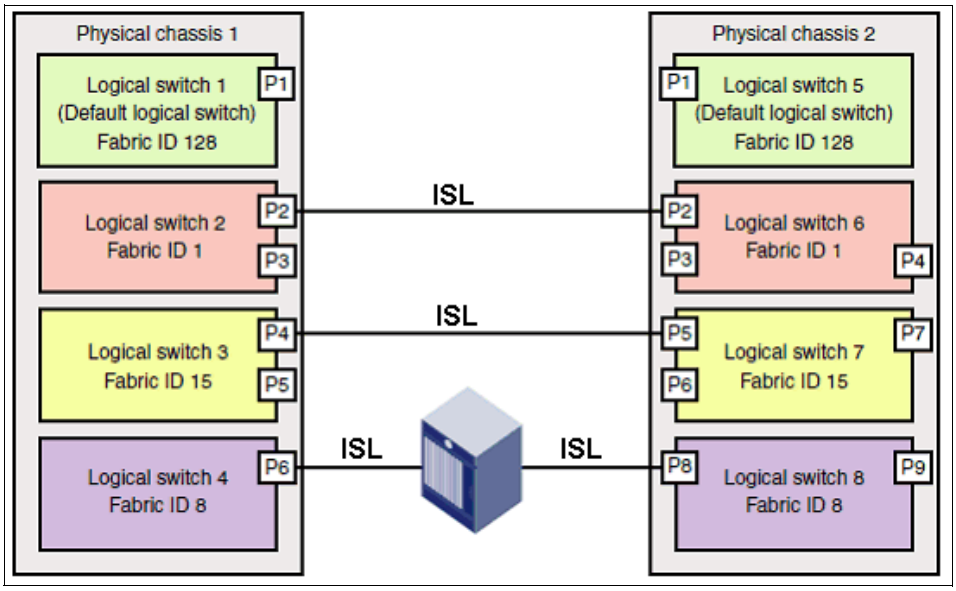

Figure 6-8 Logical fabrics with dedicated ISL

- Logical fabrics connected using a shared ISL (called Extended ISL (XISL)) from a base logical switch. In this case a separate logical switch is configured to be a base switch and will be used only for XISL connectivity and not for device connectivity. Figure 6-9 indicates a logical fabric formed through the XISL in the base switch.

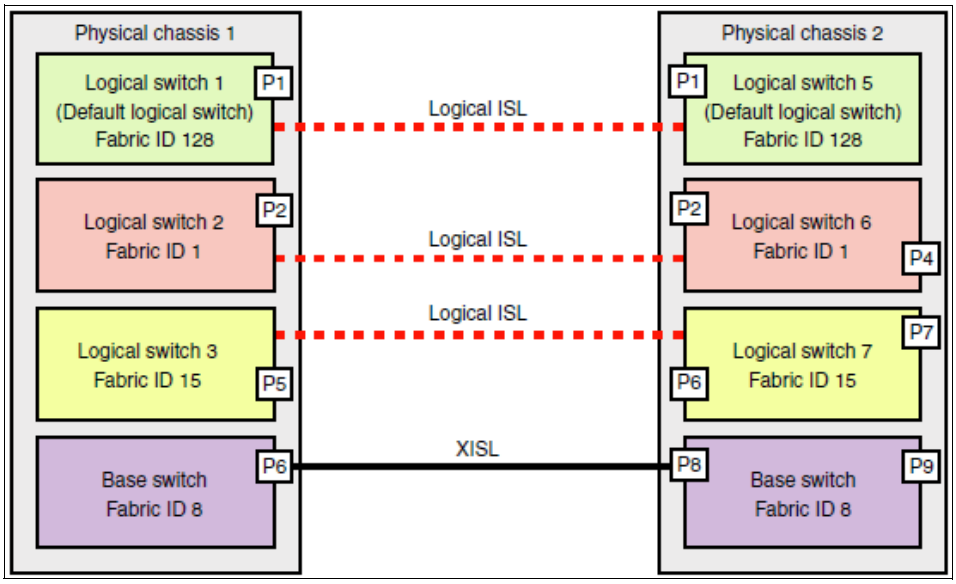

Figure 6-9 Logical ISL through XISL in base switch

6.3.2 Cisco Virtual SAN

Cisco Virtual SAN is a feature which enables the logical partition of SAN switches. A VSAN provides the flexibility to partition, for example, a dedicated VSAN for disk and tape, or to have production and test devices in separate VSANs on the same chassis. Also the VSAN can scale across the chassis which allows it to overcome the fixed port numbers on the chassis.

VSAN in a single SAN switch

VSAN brings the ability to consolidate small fabrics into the same chassis. This can also enable additional security by logical separation of the chassis in to two individual VSANs. Figure 6-10 on page 133 shows a single chassis divided into two logical VSANs.

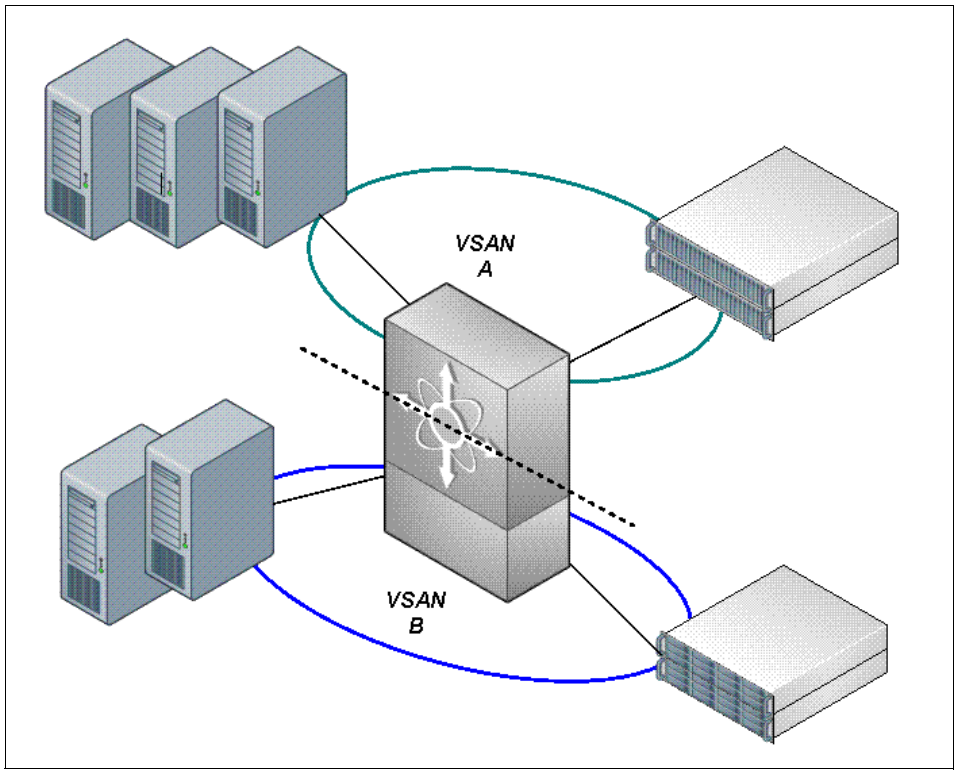

Figure 6-10 Two VSANs in a single chassis

VSAN across multiple chassis

In multiple chassis the VSAN can be formed with devices in one chassis to devices in another switch chassis through the Extended Inter Switch Link. Figure 6-11 on page 134 shows the VSAN across chassis with an EISL for VSAN communication.

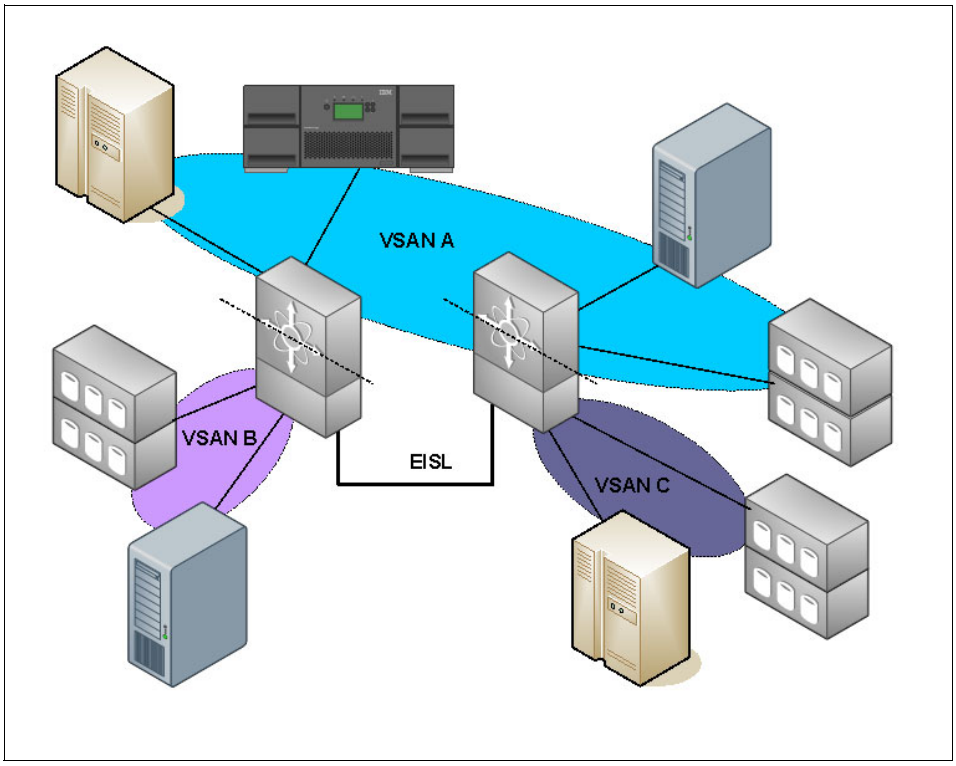

Figure 6-11 VSAN across multiple chassis

6.3.3 NPIV

Server virtualization with blade servers provides enhanced scalability of servers and this scalability is supported equally in the SAN with something called N_Port ID virtualization (NPIV). NPIV allows SAN switches to have one port shared by many virtual nodes, which in turn supports a single HBA having many virtual nodes.

Figure 6-12 on page 135 shows the sharing of a single HBA by multiple virtual nodes. In this case the same HBA will be defined with multiple virtual WWWNs and WWPNs.

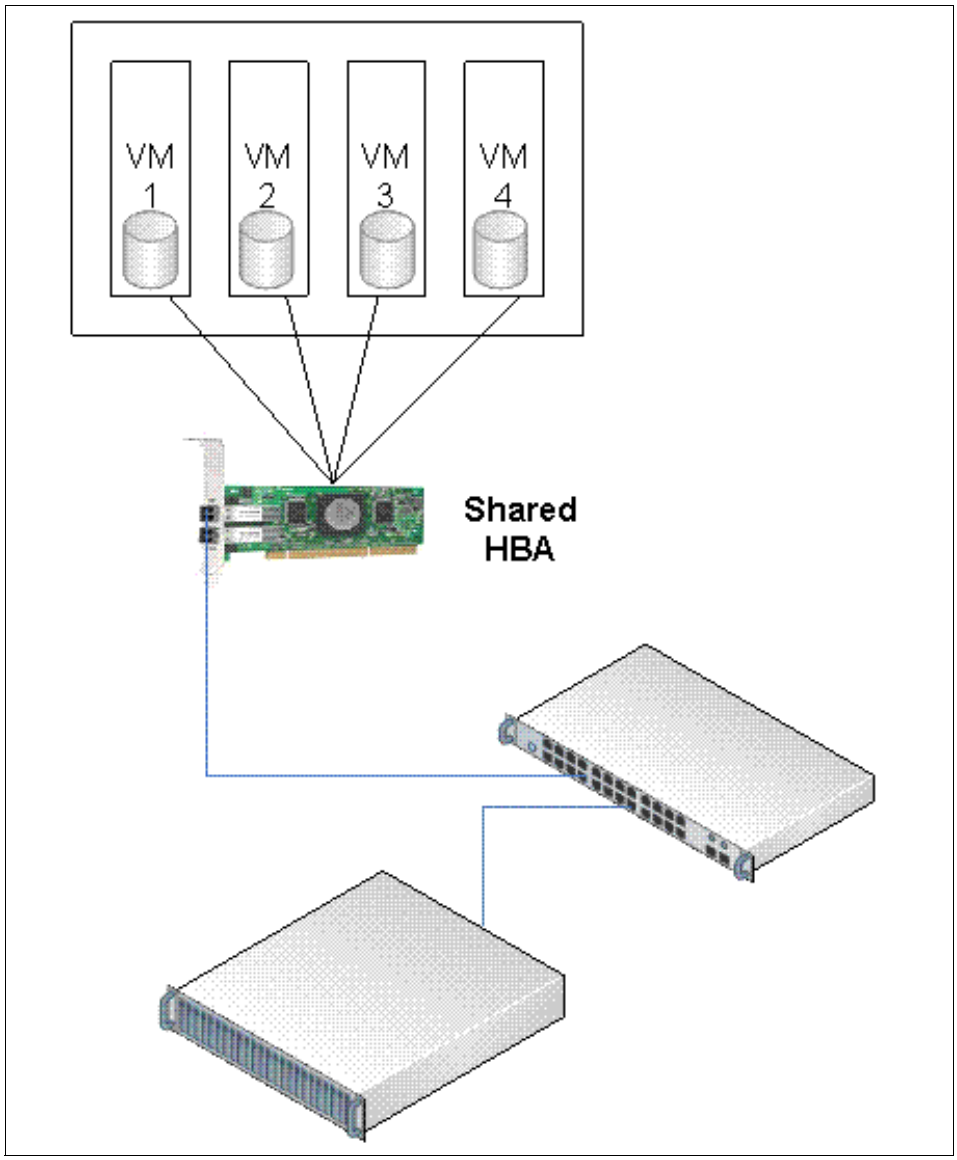

Figure 6-12 Single HBA with multiple virtual nodes

NPIV mode of blade server switch modules

Blade servers, when enabled with NPIV mode, the FC switch modules connected to an external SAN switch for access to storage will act as an HBA N_Port instead of a switch E_Port in this case. The back-end ports will be F_Ports which are connected to server blade modules. Figure 6-13 on page 136 shows the switch module in NPIV mode.

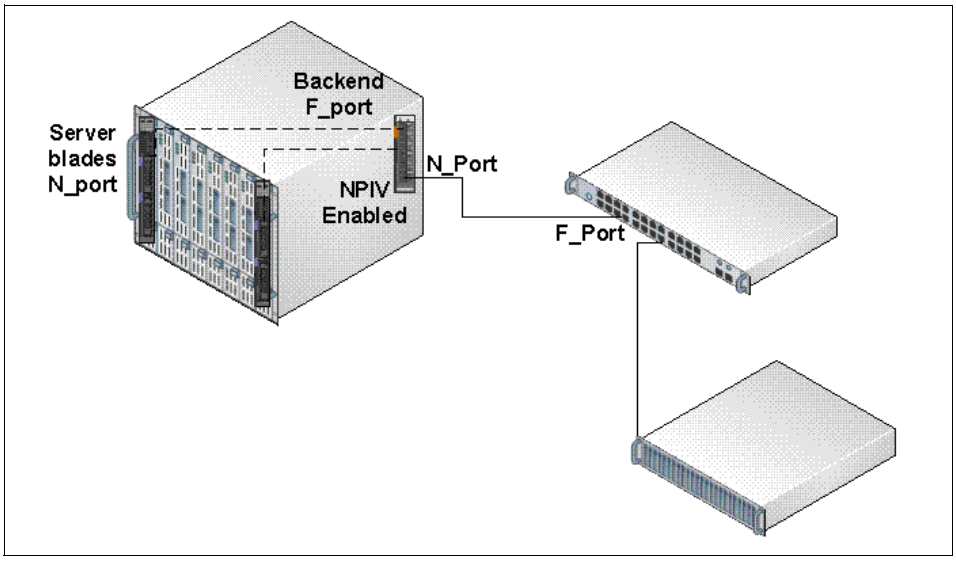

Figure 6-13 Blade server with FC switch module in NPIV mode

With NPIV mode we can overcome the interoperability issues of merging external switches to the bade server switch module which may be from different vendors. Also we have the benefit of easy management as the blade switch module becomes a node in the fabric, and we can overcome the scalability limitations of having many switch domains for a switch module in blade servers.

6.4 Building a Smarter Cloud

Storage-as-a-Service is a business model in which a large company rents space in their storage infrastructure to a smaller company or individual. It is generally seen as a good alternative when small business lacks the capital budget and/or technical personnel to implement and maintain their own storage infrastructure. in some circumstances it is being promoted as a way for all businesses to mitigate risks in disaster recovery, provide long-term retention for records and enhance both business continuity and availability.

6.4.1 Automated tiering

In modern and complex application environments, the increasing and often unpredictable demands for storage capacity and performance leads to relevant issues in terms of planning and optimization of storage resources.

Most of these issues can be managed by having spare resources available and by moving data, using data mobility tools, or using operating systems features (such as host level mirroring). However, all these corrective actions are expensive in terms of hardware resources, labor, and service availability. Relocating data among the physical storage resources dynamically, that is, transparently to hosts, is becoming increasingly important.

IBM Storage Solutions offer two types of automated tiering.

Automated tiering to optimize performance

The Easy Tier® feature, available with the DS8000, SAN Volume Controller, and Storwize V7000, provides performance optimization. Easy Tier is a built-in, dynamic data relocation

feature that provides optimal performance at the lowest cost. Easy Tier is designed to determine the appropriate tier of storage to use, based on data access patterns. Then, it automatically and non-disruptively moves data, at the sub-LUN or sub-volume level, to the appropriate disk tier.

Automated tiering to optimize space management:

The ability to optimize space management is an Information Lifecycle Management (ILM) function that is available, for instance, with Scale Out Network Attached Storage (SONAS) and with Hierarchical Storage Management, such as that provided by Tivoli Storage Manager and the IBM Data Facility Storage Management System (DFSMS) Hierarchical Storage Management (DFSMShsm).

Policy-based automation is used to migrate less active data to lower cost storage.

6.4.2 Thin provisioning

Traditional storage provisioning pre-allocates and dedicates physical storage space for use by the application or host. However, often not all space allocated to applications is needed resulting in wasted “white space”.

Thin provisioning allows a server to see logical volume sizes that are larger than the physical capacity actually dedicated to the volumes on the storage system. From the server or application perspective, thinly provisioned volumes appear and function just the same as fully provisioned volumes, but physical disk drive capacity is allocated only as needed (on demand) for write activity to the volumes. Unallocated physical capacity is available for use as needed by all volumes in a storage pool or even across an entire storage system.

Some of the advantages of thin provisioning are:

- It allows higher storage systems utilization which in turn leads to a reduction in the amount of storage you need, lowering your direct capital expenses (capex).
- It lowers operating expenses (opex) because your storage occupies less data center space and requires less electricity and cooling.
- It postpones the need to buy more storage, and as storage prices continue to drop over time, when additional capacity is required, it will likely cost less.
- Capacity planning is simplified because you are able to manage a single pool of free storage. Multiple applications or users can allocate storage from the same free pool, avoiding the situation in which some volumes are capacity constrained while others have capacity to spare.
- Your storage environment becomes more agile and it becomes easier to react to change.

Thin Provisioning Increases Utilization Ratios

Thin provisioning increases storage efficiency by increasing storage utilization ratios. Real physical capacity is provided only as it is actually needed for writing data. This results in large potential savings in both storage acquisition and operational costs, including infrastructure costs such as power, space and cooling.

Storage utilization is measured by comparing the amount of physical capacity actually used for data with the total amount of physical capacity allocated to a server. Historically, utilization ratios have been well under 50%, indicating a large amount of allocated but unused physical storage capacity. Often neither users nor storage administrators are certain how much capacity is needed, but they need to ensure that they won’t run out of space, and they also need to allow for growth. As a result, users may request more than they need and storage

administrators may allocate more than is requested, resulting in significant over-allocation of storage capacity.

Thin provisioning increases storage utilization ratios by reducing the need to over-allocate physical storage capacity to prevent out of space conditions. Large logical or virtual volume sizes may be created and presented to applications without dedicating an equivalent amount of physical capacity. Physical capacity can be allocated on demand as needed for writing data. Unallocated physical capacity is available for multiple volumes in a storage pool or across the entire storage system.

Thin provisioning also increases storage efficiency by reducing the need to resize volumes or add volumes and re-stripe data as capacity requirements grow. Without thin provisioning, if an application requires capacity beyond what is provided by its current set of volumes, there are two options:

- Existing volumes may be increased in size
- Additional volumes may be provisioned

In many environments these options are undesirable due to the steps and potential disruption required to make the larger or additional volumes visible and optimized for the application.

With thin provisioning, large virtual or logical volumes may be created and presented to applications while the associated physical capacity grows only as needed, completely transparent to the application.

Without thin provisioning, physical capacity was dedicated at the time of volume creation, and storage systems didn't typically display or report how much of the dedicated physical capacity was actually used for data. As storage systems have implemented thin provisioning, physical allocation and usage has been made visible. Thin provisioning increases storage efficiency by making it easy to see the amount of physical capacity actually needed and used, because physical space is not allocated until it is needed for data.

6.4.3 Deduplication

Data deduplication has emerged as a key technology in an effort to dramatically reduce the amount and the cost associated with storing large amounts of data. Deduplication is the art of intelligently reducing storage needs an order of magnitude better than common data compression techniques – through the elimination of redundant data so that only one instance of a data set is actually stored. IBM has the broadest portfolio of deduplication solutions in the industry giving IBM the freedom to solve client issues with the most effective technology. Whether its source or target, inline or post, hardware or software, disk or tape, IBM has a solution with the technology that best solves the problem.

- IBM ProtecTIER® Gateway and Appliance
- IBM System Storage N series Deduplication
- IBM Tivoli Storage Manager

Data deduplication is a technology that reduces the amount of space required to store data on disk. It achieves this space reduction by storing a single copy of data that is backed up repetitively.

Data deduplication products read data while they look for duplicate data. Data deduplication products break up data into elements, using their respective technique to create a signature or identifier for each data element. Then, they compare the data element signature to identify duplicate data. After they identify duplicate data, they retain one copy of each element. They create pointers for the duplicate items, and discard the duplicate items.

The effectiveness of data deduplication is dependent upon many variables, including the rate of data change, the number of backups, and the data retention period. For example, if you back up the exact same incompressible data once a week for six months, you save the first copy and do not save the next 24, which would provide a 25 to 1 data deduplication ratio. If you back up an incompressible file on week one, back up the exact same file again on week two and never back it up again, you have a 2 to 1 deduplication ratio. A more likely scenario is that some portion of your data changes from backup to backup so that your data deduplication ratio will change over time. With data deduplication you can minimize your storage requirements.

Data deduplication can provide greater data reduction and storage space savings than other existing technologies.

Figure 6-14 shows the basic concept of data deduplication.

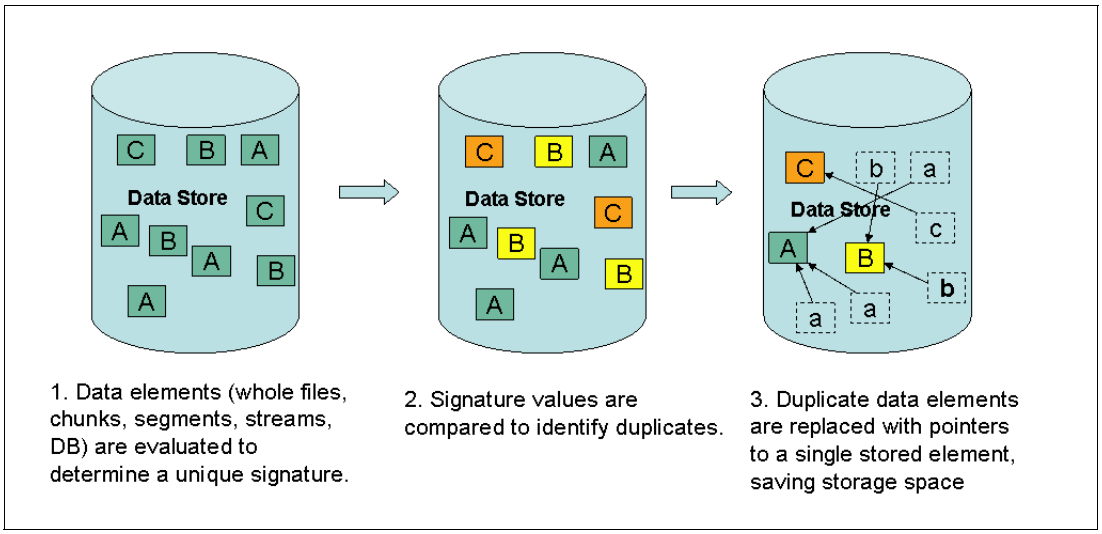

Figure 6-14 The basic concept of data deduplication

Data deduplication can reduce your storage requirements but the benefit you derive is determined by your data and your backup policies. Workloads with a high database content generally have the highest deduplication ratios; however, product functions like Tivoli Storage Manager Progressive Incremental, Oracle RMAN, or Light Speed, can reduce the deduplication ratio. Compressed, encrypted, or otherwise scrambled workloads typically do not benefit from deduplication. Good candidates for deduplication are typically text files, log files, uncompressed and non-encrypted database files, email files (PST, DBX, Domino®), Snapshots (Filer Snaps, BCVs, VMware images).

Types of data deduplication and HyperFactor

Many vendors offer products that perform deduplication. Various methods are used for de-duplicating data. Three methods frequently used:

- **Hash based** deduplication uses a hashing algorithm to identify chunks of data. Commonly used process is Secure Hash Algorithm 1 (SHA-1) or Message-Digest Algorithm 5 (MDA-5). The details of each technique are beyond the intended scope of this publication.
- **Content aware** deduplication methods are aware of the structure of common patterns of data used by applications. It assumes the best candidate to de-duplicate against is an object with the same properties, such as a file name. When a file match is found, a bit by bit comparison is performed to determine if data has changed and saves the changed data.

- **HyperFactor®** is a patented technology which is used in IBM System Storage ProtecTIER Enterprise Edition higher software. HyperFactor takes an approach that reduces the phenomenon of missed factoring opportunities, providing a more efficient process. With this approach, HyperFactor is able to surpass the reduction ratios attainable by any other data reduction method. HyperFactor can reduce any duplicate data, regardless of its location or how recently it was stored. HyperFactor data deduplication uses a 4 GB Memory Resident Index to track similarities for up to 1 petabyte (PB) of physical disk in a single repository.

HyperFactor technology uses a pattern algorithm that can reduce the amount of space required for storage by up to a factor of 25, based on evidence from existing implementations. The capacity expansion that results from data deduplication is often expressed as a ratio, essentially the ratio of nominal data to the physical storage used.

Figure 6-15 on page 140 shows the HyperFactor technology.

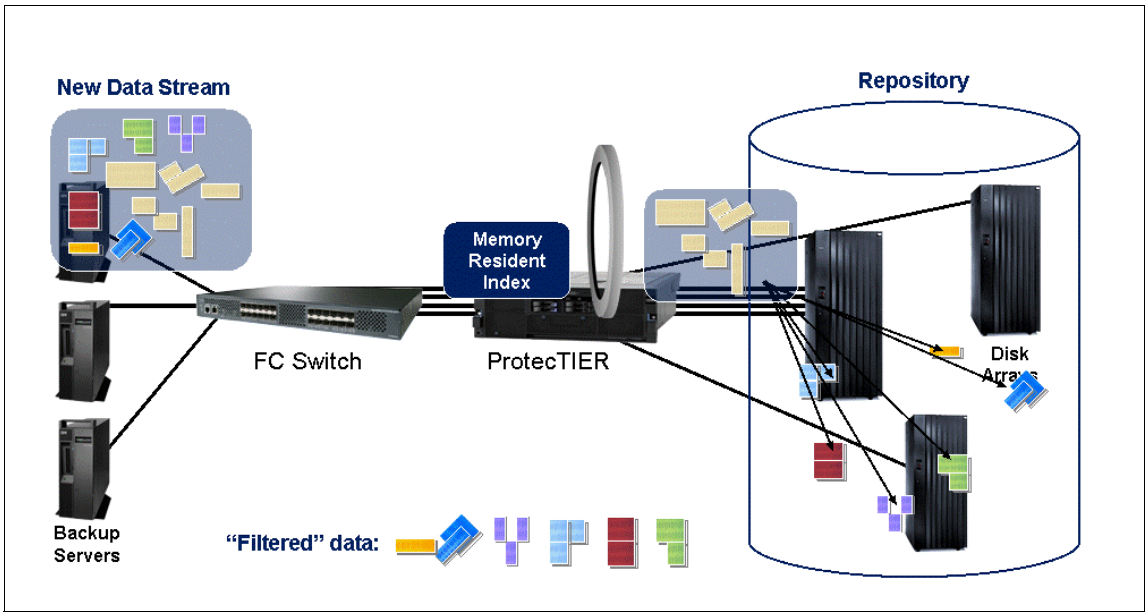

Figure 6-15 IBM HyperFactor technology

Data deduplication processing

Data deduplication can either be performed while the data is being backed up to the storage media (real-time or inline) or after the data has been written to the storage media (post-processing). Each method certainly brings positive and negative aspects, those must be evaluated by the engineer or technical specialist responsible for the concrete solution architecture and deployment. IBM decided to use inline deduplication processing as it offers larger target storage space without any need of temporary disk cache pool for post processed deduplication data.

Bit comparison techniques such as the one used by ProtecTIER were designed to provide 100% data integrity by avoiding the risk of hash collisions.

6.4.4 New generation management tools

It is paramount that this new virtualized infrastructure be managed by new generation management tools, because older tools generally lack the required features.

When used properly, these tools can make the adoption of virtualization technology easier and more cost effective. The tools have the following additional benefits:

- Enable line of business insight into storage utilization and allocation, enabling easier departmental charge back.
- Allow for more intelligent business decisions about storage efficiency, enabling you to respond faster to changing business demands, yet reduce costs.
- Provide better understanding of application storage performance and utilization patterns enabling better Information Lifecycle Management (ILM) of application data.
- Allow an organization to perform infrastructure management pro-actively, through proper capacity management, rather than reactively.
- Improve operational efficiency leading to cost savings for the organization.

6.4.5 Business Continuity and Disaster Recovery

The importance of business continuity and disaster recovery remains at the forefront of thought for many executives and IT technical professionals, and the most important factor to consider is how the choice of the technology impacts the recovery time objective (RTO). For SAN there are lot of possible solutions available and the cloud design will drive the selections that are capable of meeting the requirements. A smart cloud must be capable of guaranteeing business continuity and any disaster recovery plans.

Note: For more details of DR lessons learned and solutions refer to the IBM Storage Infrastructure for Business Continuity.

http://www.redbooks.ibm.com/abstracts/redp4605.html?Open

6.4.6 Storage On Demand

Scalable, pay-per-use cloud storage can help to manage massive data growth and storage budget. This leads to a costly procurement cycle with setup costs and implementation delays every time that more storage is needed. Cloud storage allows the ability to expand storage capacity on the spot, and later, shrink storage consumption if needed.

Cloud storage provides a ready-made data storage solution helping in these areas:

- Reduce up front capital expenses
- Meet demands without expensive over-provisioning
- Supplement other storage systems more cost-effectively
- Align data storage costs with business activity
- Scale dynamically

7

Fibre Channel products and technology

In this chapter we describe some of the most common Fibre Channel SAN products and technology that are encountered. For a description of the IBM products that have been inducted into the IBM System Storage and TotalStorage portfolio, refer to Chapter 12, “The IBM product portfolio” on page 239.

7.1 The environment

The SNIA definition of a SAN, Fibre Channel and Storage is as follows

- Storage Area Network (SAN)

  [Network] A network whose primary purpose is the transfer of data between computer systems and storage elements and among storage elements.

A SAN consists of a communication infrastructure, which provides physical connections, and a management layer, which organizes the connections, storage elements, and computer systems so that data transfer is secure and robust. The term SAN is usually (but not necessarily) identified with block I/O services rather than file access services.

- Fibre Channel

  A serial I/O interconnect capable of supporting multiple protocols, including access to open system storage (FCP), access to mainframe storage (FICON), and networking (TCP/IP). Fibre Channel supports point to point, arbitrated loop, and switched topologies with a variety of copper and optical links running at speeds from 1 Gb/s to 10 Gb/s. The committee standardizing Fibre Channel is the INCITS Fibre Channel (T11) Technical Committee

- Storage System

  A storage system consisting of storage elements, storage devices, computer systems, and/or appliances, plus all control software, communicating over a network.

Storage subsystems, storage devices, and server systems can be attached to a Fibre Channel SAN. Depending on the implementation, several different components can be used to build a SAN. It is, as the name suggests, a network so any combination of devices that are able to interoperate are likely to be utilized.

Given this, a Fibre Channel network may be composed of many different types of interconnect entities, including directors, switches, hubs, routers, gateways, and bridges.

It is the deployment of these different types of interconnect entities that allow Fibre Channel networks of varying scale to be built. In smaller SAN environments you can employ hubs for Fibre Channel arbitrated loop topologies, or switches and directors for Fibre Channel switched fabric topologies. As SANs increase in size and complexity, Fibre Channel directors can be introduced to facilitate a more flexible and fault tolerant configuration. Each of the components that compose a Fibre Channel SAN should provide an individual management capability, as well as participate in an often complex end-to-end management environment.

Figure 7-1shows generic SAN connection

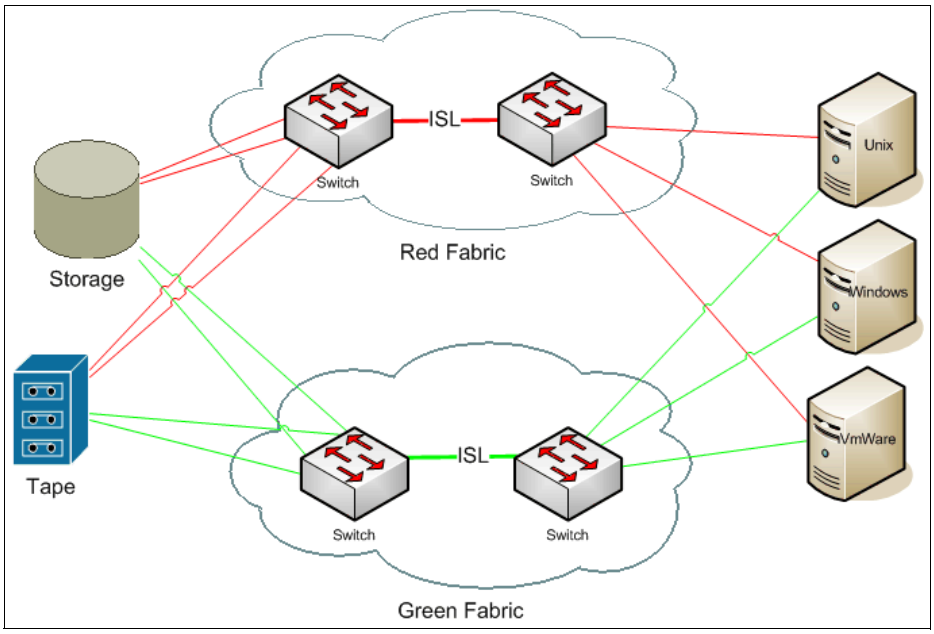

Figure 7-1 Generic SAN

7.2 SAN devices

A Fibre Channel SAN employs a fabric to connect devices, or end points. A fabric can be as simple as a single cable connecting two devices, akin to server attached storage. However, the term is most often used to describe a more complex network to connect servers and storage utilizing switches, directors, and gateways.

Independent from the size of the fabric, a good SAN environment starts with good planning, and always includes an up-to-date map of the SAN.

Some of the items to consider are:

- How many ports do I need now?
- How fast will I grow in two years?
- Are my servers and storage in the same building?
- Do I need long distance solutions?
- Do I need redundancy for every server or storage?
- How high are my availability needs and expectations?
- Will I connect multiple platforms to the same fabric?
- What technology to I want to use, FC - FCoE - iScsi?

7.2.1 Fibre Channel bridges

Fibre Channel bridges allow the integration of legacy SCSI devices in a Fibre Channel network. Fibre Channel bridges provide the capability for Fibre Channel and SCSI interfaces to support both SCSI and Fibre Channel devices seamlessly. Therefore, they are often referred to as FC-SCSI routers.

Note: Fibre Channel bridges are not to be confused with ***Data Center Bridging (DCB)***, though fundamentally they serve the same purpose, to interconnect different protocols.

A bridge is a device that converts signals and data from one form to another. We can imagine these devices in a similar way as the bridges we use to cross rivers with. They act as a translator/a bridge between two different protocols. These protocols can be:

- Fibre Channel
- iSCSI
- SSA
- FCIP

We do not see many of these devices today and they are considered legacy devices.

7.2.2 Arbitrated loop hubs and switched hubs

Arbitrated loop, also known as FC-AL, is a Fibre Channel topology in which devices are connected in a one-way loop fashion in a ring topology, this is also described in Chapter 5, “Topologies and other fabric services” on page 83.

Figure 7-2 on page 146 shows FC-AL topology

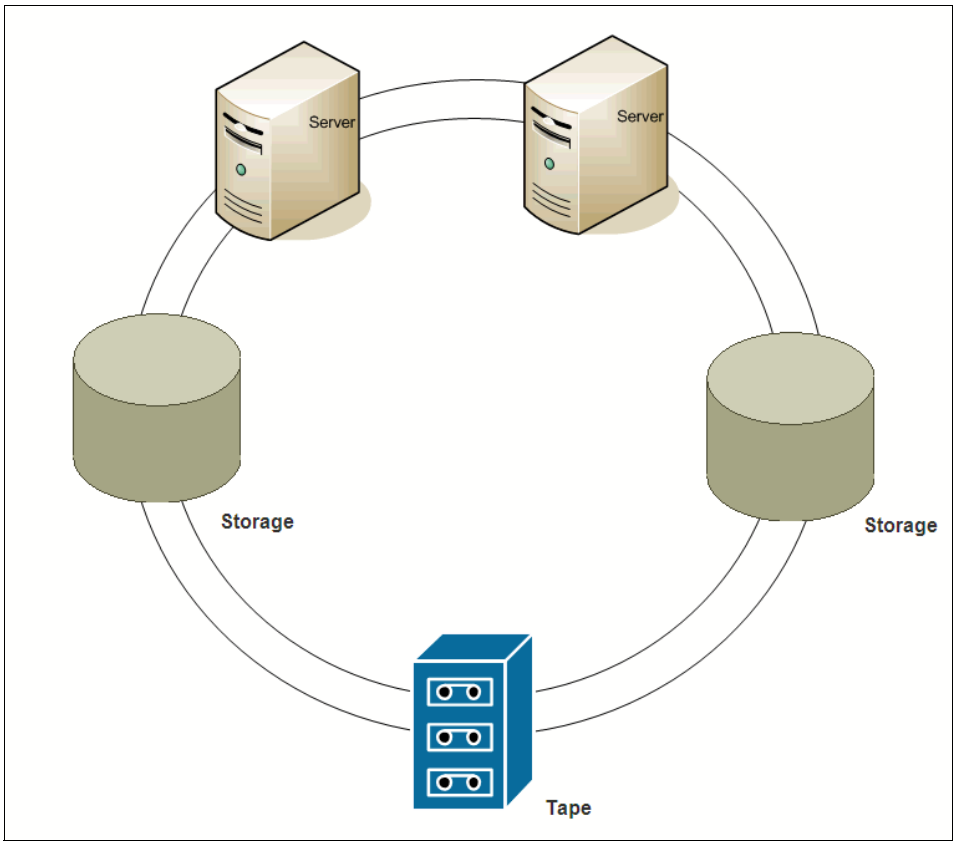

Figure 7-2 Arbitrated loop

In FC-AL all devices on the loop share the bandwidth. The total number of devices that may participate in the loop is 126, without using any hubs or fabric. For practical reasons, however, the number tends to be limited to no more than 10 and 15.

Hubs are typically used in a SAN to attach devices or servers that do not support switched fabric only FC-AL. They may be unmanaged hubs, managed hubs, or switched hubs.

Unmanaged hubs serve as cable concentrators and as a means to configure the Arbitrated Loop based on the connections it detects. When any of the hub's interfaces, usually GBIC, senses no cable connected, that interface shuts down and the hub port is bypassed as part of the Arbitrated Loop configuration.

Managed hubs offer all the benefits of unmanaged hubs, but in addition offer the ability to manage them remotely, using SNMP.

It used to be a an alternative to a fabric topology but FC switches and directors. By using FC-AL you could connect many servers and storage devices without using then very costly Fibre Channel switches. So FC-AL is not used much today since switched fabrics have taken the lead in the Fibre Channel market.

Switched hubs

Switched hubs allow devices to be connected in its own Arbitrated Loop. These loops are then internally connected by a switched fabric.

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

Switched hubs are usually managed hubs.

Note: In its early days, FC-AL was described as "SCSI on steroids". Although FC-AL has the bandwidth advantage over SCSI, it does not come anywhere close to the speeds that can be achieved and sustained on a per port basis in a switched fabric. For this reason, FC-AL implementations are, by some observers, considered as legacy SANs.

7.2.3 Switches and directors

Switches and directors allow Fibre Channel devices to be connected (cascaded) together, implementing a switched fabric topology between them. The switch intelligently routes frames from the initiator to responder and operates at full Fibre Channel bandwidth.

It is possible to connect switches together in cascades and meshes using inter-switch links (E_Ports). It should be noted that devices from different manufacturers may not interoperate fully.

The switch also provides a variety of fabric services and features such as:

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

It is common to refer to switches as either *core* switches or *edge* switches depending on where they are located in the SAN. If the switch forms, or is part of the SAN backbone, then it is the core switch. If it is mainly used to connect to hosts or storage then it is called an edge switch. Like it or not, directors are also sometimes referred to as switches, since they are switches in all essence. Directors are large switches with higher redundancy than most normal switches. Whether this is appropriate or not is a matter for debate outside of this book.

7.2.4 Multiprotocol routing

There are also devices that are multiprotocol routers and devices. These provide improved scalability, security, and manageability by enabling devices in separate SAN fabrics to communicate *without* merging fabrics into a single, large Meta-SAN fabric. Depending on the manufacturer, they support a number of protocols and have their own features, such as zoning. As their name suggests, the protocols supported include:

- FCP
- FCIP
- iFCP
- iSCSI
- IP

7.2.5 Service modules

Increasingly, with the demand for the intermix of protocols and the introduction to the marketplace of new technologies, SAN vendors are starting to adopt a modular system approach to their components. What this means is that service modules can be plugged into a slot on the switch or director to provide functions and features such as virtualization, the combining of protocols, storage services, and so on.

7.2.6 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 based on either time or wavelength:

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

When using multiplexers in a SAN environment additional parameters in the SAN switch configuration could be needed to ensure correct load balancing, so check with your SAN switch vendor for best practices.

Note: Usually multiplexers are transparent to the SAN fabric. If you are troubleshooting an ISL link that covers some distance, keep in mind that the multiplexer, if installed, plays an important role in that path.

7.3 Componentry

There are, of course, a number of components that have to come together to make a SAN, well, a SAN. We will identify some of the components that are likely to be encountered.

7.3.1 ASIC

The fabric electronics utilize personalized application-specific integrated circuits (ASIC or ASICs) and its predefined set of elements, such as logic functions, I/O circuits, memory arrays, and backplane to create specialized fabric interface components.

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

Frame filtering

Frame filtering is a feature that enables devices to provide zoning functions with finer granularity. Frame filtering can used to set up port level zoning, world wide name zoning, device level zoning, protocol level zoning, and LUN level zoning. Frame filtering is commonly carried out by an ASIC. This has the result that, after the filter is set up, the complicated function of zoning and filtering can be achieved at wire speed.

7.3.2 Fibre Channel transmission rates

Sometimes referred to as feeds and speeds. The current set of vendor offerings for switches, host bus adapters, and storage devices is constantly increasing. Currently the 16Gb FC port has the fastest line rate supported for an IBM SAN. The 16Gb FC port uses 14.025 Gbps transfer rate and uses 64b/66b encoding giving approximately 1600 MB/sec in throughput. The 8Gb FC port has a line rate of 8.5 Gbps using 8b/10b encoding resulting in approximately 800 MB/sec. It should also be mentioned that when comparing feeds and speeds then FC ports are sometimes referred to as “full duplex” and then the transceiver and receive parts of the FC port are added together therefore “doubling” the MB/sec.

Note: It is worth mentioning that by introducing the 64b/66b encoding to Fibre Channel the encoding overhead is reduced from approximately 20% using 8b/10b encoding down to approximately 3% with the 64b/66b encoding.

The new 16Gb FC port is approved by FCIA (Fibre Channel Industry Association) and that ensures that each port speed is able to communicate with at least two previous approved port speeds, for example 16Gb is able to communicate with 8Gb and 4Gb.

The FCIA has also created roadmap for future feeds and speeds and it can be found at:

http://www.fibrechannel.org/

7.3.3 SerDes

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

7.3.4 Backplane and blades

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

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

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

7.4 Gigabit transport technology

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 that can convert other communication transport into gigabit transport. The bit error rate (BER) only allows for a single bit error to occur once in every trillion bits in the Fibre Channel standard. Gigabit transport can be used in copper or fiber optic infrastructure.

Layer 1 of the OSI model is the layer at which the physical transmission of data occurs. The unit of transmission at Layer 1 is a *bit*. This section explains some of the common concepts that are at the Layer 1 level.

7.4.1 FC cabling

Fibre Channel cabling is one of two forms: fiber optic cabling or copper cabling. Fiber optic cabling is the usual cabling type but with the introduction of FCoE copper cabling is introduced.

Fiber optic cabling is more expensive than copper cabling. The optical components for devices and switches and the cost of any customer cabling is typically more expensive to install. However, the higher costs are often easily justified by the benefits of fiber optic cabling.

Fiber optic cabling provides for longer distance and is resistant to the signaling being distorted by electromagnetic interference.

Fiber optic cabling

In copper cabling, electric signals are used to transmit data through the network. The copper cabling was the medium for that electrical transmission. In fiber optic cabling, light is used to transmit the data. Fiber optic cabling is the medium for channeling the light signals between devices in the network.

Two modes of fiber optic signaling are explained in this chapter, single-mode and multimode. The difference between the modes is the wavelength of the light used for the transmission as illustrated in Figure 7-3.

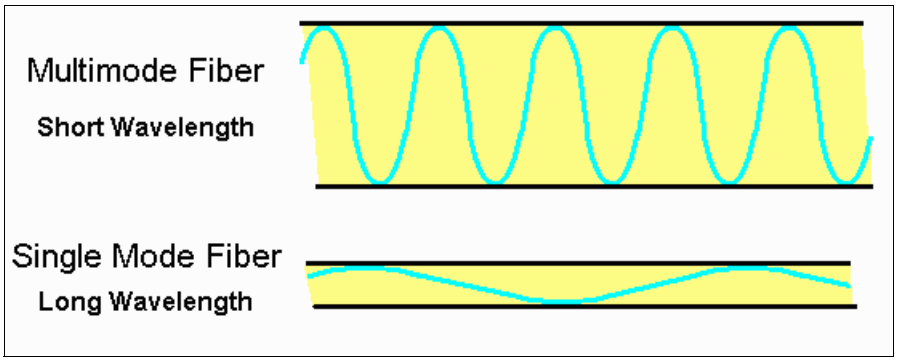

Figure 7-3 Multimode versus single-mode optic signaling

Single-mode fiber

Single-mode fiber (SMF) uses long wavelength light to transmit data and requires a cable with a small core for transmission (Figure 7-3). The core diameter for single-mode cabling is 9 microns in diameter (Figure 7-4).

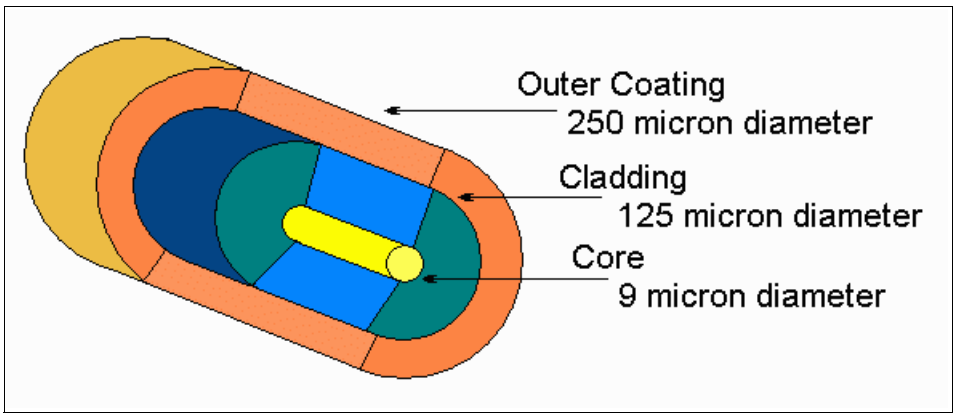

Figure 7-4 Single-mode fiber cable

Multimode fiber

Multimode fiber (MMF) uses short wavelength light to transmit data and requires a cable with a larger core for transmission (Figure 7-3 on page 150). The core diameter for multimode cabling can be 50 or 62.5 microns in diameter as illustrated in Figure 7-5.

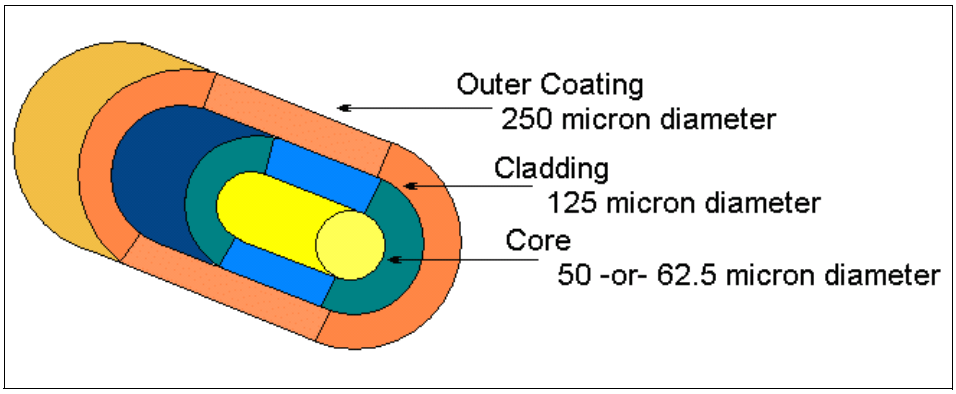

Figure 7-5 Multimode fiber cable

The color of the outer coating is sometimes used to identify if a cable is a multimode or single-mode fiber cable, but the color is not a reliable method. The TIA-598C standard suggests the outer coating to be yellow for single mode fiber and orange for multimode fiber for civilian applications. This guideline is not always implemented as illustrated in Figure 7-6, which shows a blue cable. The reliable method is to look at the specifications of the cable printed on the outer coating of the cabling. See also Figure 7-7 on page 152 and Figure on page 152.

Figure 7-6 Blue 62.5 micron MMF cable

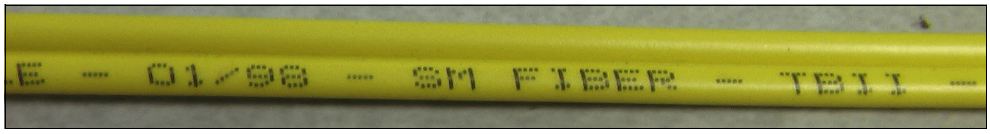

Figure 7-7 Yellow SMF cable

Orange 50 micron MMF cable

Copper Cabling

By copper cabling we mean that the material used to transfer the signals are made of copper. The most common copper wire is the twisted pair cable used for normal ethernet. This is explained in more depth in the following section.

Twisted-pair cabling

Twisted-pair copper cabling is a common media for Ethernet networking installations. Twisted-pair cabling is available as Unshielded Twisted-Pair (UTP) or Shielded Twisted-Pair (STP). This shielding helps prevent electromagnetic interference.

Several different categories of twisted-pair cabling are available as listed in Table 7-1. These categories indicate the signaling capabilities of the cabling.

Table 7-1 TIA/EIA cabling categories

TIA/EIA cabling category	Maximum network speeds supported
Cat 1	Telephone or ISDN
Cat 2	4 Mb Token Ring
Cat 3	10 Mb Ethernet
Cat 4	16 Mb Token Ring
Cat 5	100 Mb Ethernet
Cat 5e	1 Gb Ethernet
Cat 6	10 Gb Ethernet Short Distance - 55 m (180 ft.)
Cat 6a	10 Gb Ethernet

The connector used for Ethernet twisted-pair cabling is likely the one most people recognize and associate with networking, the RJ45 connector, which is shown in Figure 7-8.

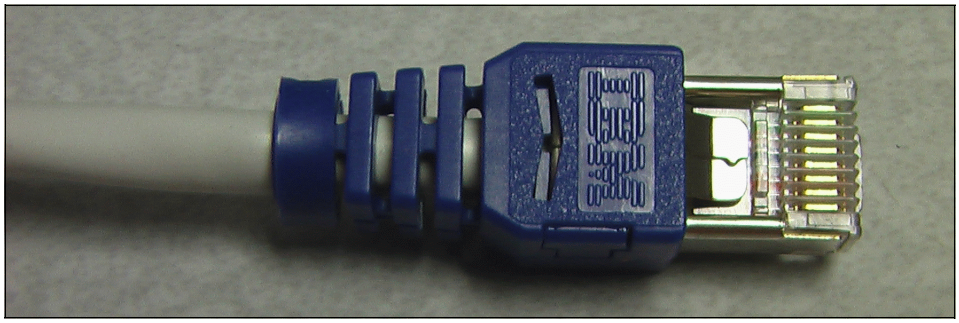

Figure 7-8 RJ45 Copper Connector

Twisted-pair cabling contains four pairs of wire inside the cable, as illustrated in Figure 7-9 on page 153.

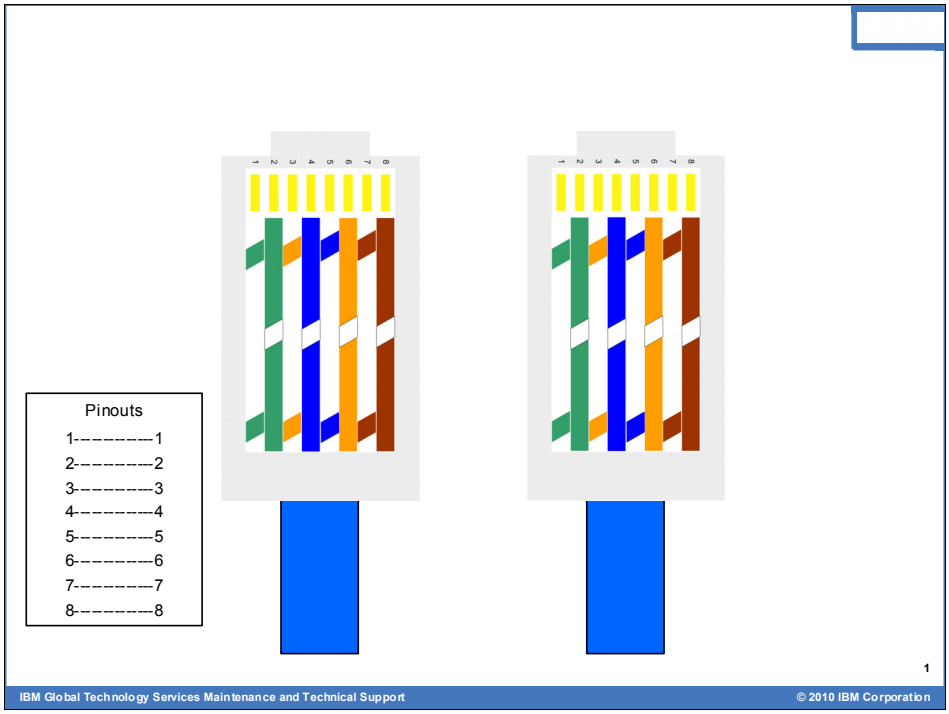

Figure 7-9 Straight through Ethernet cable

An Ethernet operating in 10/100 Mb mode only uses two pairs, pairs 1-2 and 3-6. An Ethernet operating in 1 Gb mode uses all four pairs: pairs 1-2, 3-6, 4-5, and 7-8. Distances up to 100 meters are supported.

Damaged twisted pair: If a twisted-pair cable is damaged, so that pair 4-5 or pair 7-8 is unable to communicate, the link is unable to communicate in 1 Gbps mode. If the devices are set to auto negotiate speed, the devices successfully operate in 100 Mbps mode.

Supported maximum distances of cabling segment: The actual maximum distances of a cabling segment that are supported vary on multiple factors such as vendor support, cabling type, electromagnetic interference, and number physical connections in the segment.

Twinax cabling

Twinax cables have been used by IBM for many years but they are recently being re-introduced to the market as a transport media for 10Gb Ethernet. One of the biggest benefits of twinax cable is the low power consumption and they are lower cost than standard fiber cables. The downside is the limited capability to connect over long distance.

Connector types

The most common connector type for fiber optic media used in networking today is the LC connector, which is shown in Figure 7-10.

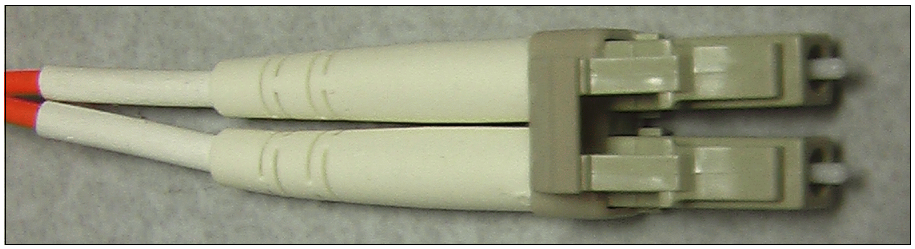

Figure 7-10 LC fiber connector

Other type of connectors are the SC connector (Figure 7-11), and the ST connector (Figure 7-11).

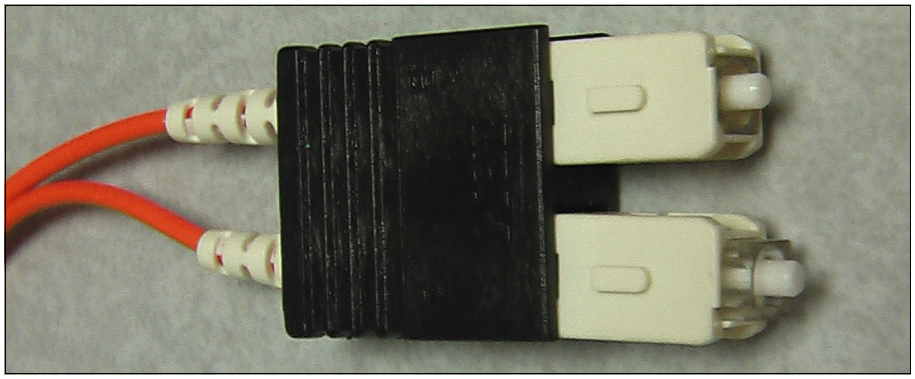

Figure 7-11 SC fiber connector

7.4.2 Transceivers

A *transceiver* or *transmitter/receiver* is the fiber optic port of a device. It is where the fiber optic cables connect. Occasionally a device might have an integrated transceiver, which limits the flexibility in the type of cabling that can be used. Most devices provide a slot for a modular transceiver to be inserted, providing flexibility for single or multimode implementations to be selected.

Some equipment might use a larger transceiver known as a *Gigabit Interface Converter* (GBIC), which is shown in Figure 7-12. As technology advancements have been made, smaller transceivers have been introduced providing much higher port density, such as small form-factor pluggable (SFP), 10-Gigabit SFP+, 10-Gigabit SFP-XFP and Quad SFP (QSFP). See also Figure 7-13 to compare the different transceivers.

Figure 7-12 Gigabit Interface Converter

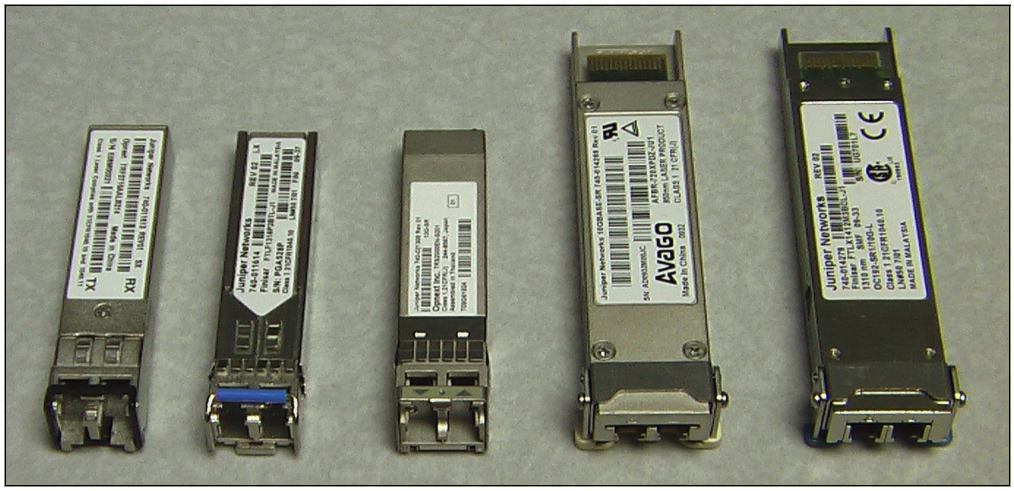

Figure 7-13 From left to right: SFP-MMF, SFP-SMF, SFP+-MMF, XFP-MMF, and XFP-SMF

Figure 7-14 on page 156 shows a QSFP or Quad SFP and cable.

Figure 7-14 QSFP and cable

7.4.3 Host bus adapters

The device that acts as an interface between the fabric of a SAN and either a host or a storage device is a host bus adapter (HBA).

FC HBA

The HBA connects to the bus of the host or storage system. Some devices offer more than one Fibre Channel connection and even have a built in SFP that can be replaced. The function of the HBA is to convert the parallel electrical signals from the bus into a serial signal to pass to the SAN.

A host bus adapter is shown in Figure 7-15 on page 157.

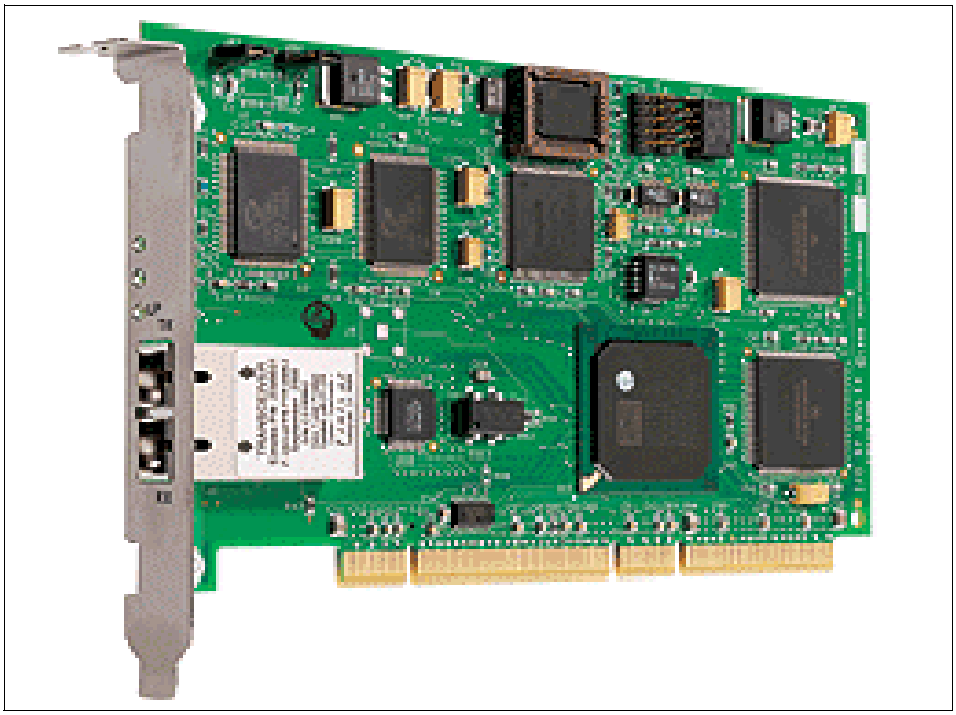

Figure 7-15 HBA

There are several manufacturers of HBAs, and an important consideration when planning a SAN is the choice of HBAs. HBAs may have more than one port, may be supported by some equipment and not others, may have parameters that can be used to tune the system, and many other features.It is important to know that HBA's also have certain amount of buffer-to-buffer credits so if you are thinking of using a certain HBA with multiple virtual machines behind it, you should be aware of the choice of an HBA is a critical one.

CNA HBA

CNA is short for Converged Network Adapter. These adapters are capable of running both Converged Enhanced Ethernet (CEE) and Fibre Channel (FC) traffic at the same time. These CNA's combine the functions of a Host Bus Adapter (HBA) and Network Interface Card (NIC) on one card. These CNA's fully support FCoE protocols and allow Fibre Channel traffic to converge onto 10 Gbps Converged Enhanced Ethernet (CEE) networks. These adapters play a critical role in the FCoE implementation.

When implementing Converged Network Adapters (CNA's) they are big enablers in reducing data center costs by converging data and storage networking. Standard TCP/IP and Fibre Channel (FC) traffic can both run on the same high-speed 10 Gbps Ethernet wire, resulting in cost savings through reduced adapter, switch, cabling, power, cooling and management requirements. CNA's have gained rapid market traction because they deliver excellent performance, help reduce data center TCO, and protect current data center investment. Cutting–edge 10 Gbps bandwidth can eliminate performance bottlenecks in the I/O path with a 10X data rate improvement versus existing 1 Gbps Ethernet solutions. Additionally, full hardware off load for FCoE protocol processing reduces system CPU utilization for I/O operations, which leads to faster application performance and higher levels of consolidation in virtualized systems.

Figure 7-16 on page 158 shows a dual port CNA adapter.

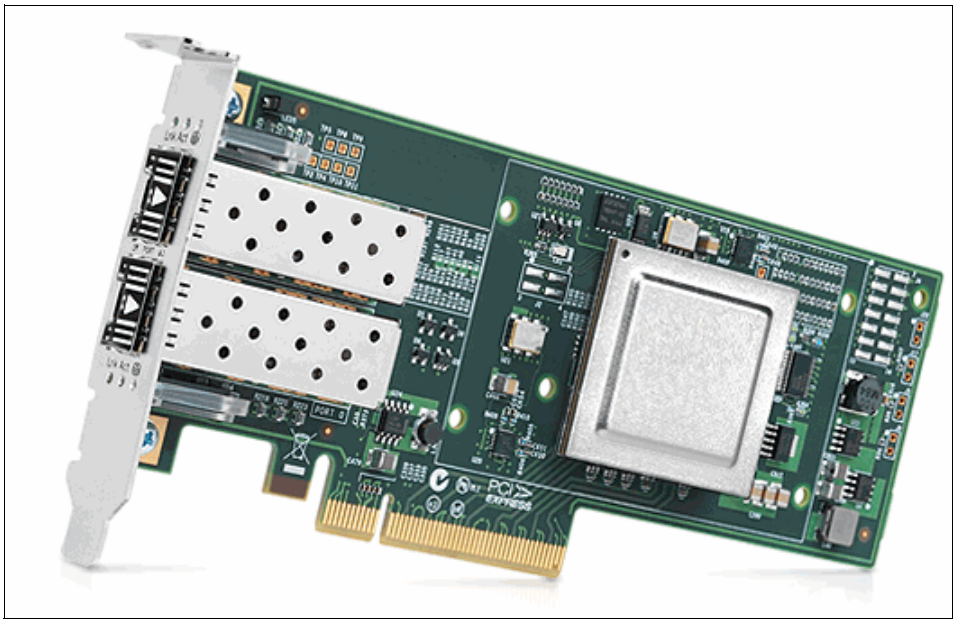

Figure 7-16 Dual port Converged Network Adapter

7.5 Inter-switch links

A link that joins a port on one switch to a port on another switch (referred to as E_Ports) is called an inter-switch link (ISL).

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

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

There is also the possibility to connect an E_Port to a Fibre Channel router or a switch with embedded routing capabilities, and then the ports becomes an EX-port on the router side. Brocade calls these ports an IFL (inter-fabric link); however Cisco uses TE_Port (trunked E_Port) also known as an EISL that allows traffic (from multiple VSAN's) to be routed through that link.

7.5.1 Cascading

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

- The fabric can be seamlessly extended. Additional switches can be added to the fabric, without powering down existing fabric.
- You can easily increase the distance between various SAN participants.
- By adding more switches to the fabric, you increase connectivity by providing more available ports.
- Cascading provides high resilience in the fabric.

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

7.5.2 Hops

When FC traffic traverses an ISL, this is known as a *hop*. Or, to put it another way, traffic going from one E_Port over an ISL to another E_Port is one hop. As stated in 7.5, “Inter-switch links” on page 158 that ISL’s are made from connecting an E_Port to an E_Port. In Figure 7-17 on page 160 we see an illustration of the hop count from server to storage.

There is a hop count limit. This is set by the fabric operating system and is used to derive a frame holdtime value for each switch. This holdtime value is the maximum amount of time that a frame can be held in a switch before it is dropped, or if the fabric indicates that it is too busy. The hop count limits need to be investigated and considered in any SAN design work as it will have a major effect on the proposal.

7.5.3 Fabric shortest path first

Although not strictly speaking a physical component, it makes sense to introduce fabric shortest path first (FSPF) at this stage. According to the FC-SW-2 standard, fabric shortest path first is a link state path selection protocol. FSPF keeps track of the links on all switches in the fabric (in routing tables) and associates a cost with each link. The protocol computes paths from a switch to all the other switches in the fabric by adding the cost of all links traversed by the path, and choosing the path that minimizes the cost i.e shortest link.

For example, as shown in Figure 7-17 on page 160, if a server needs to connect to its storage through multiple switches, FSPF will route all traffic from this server to its storage through switch A directly to switch C since it has a lower cost than travelling through additional hop via switch B.

Figure 7-17 on page 160 showing hops in a fabric.

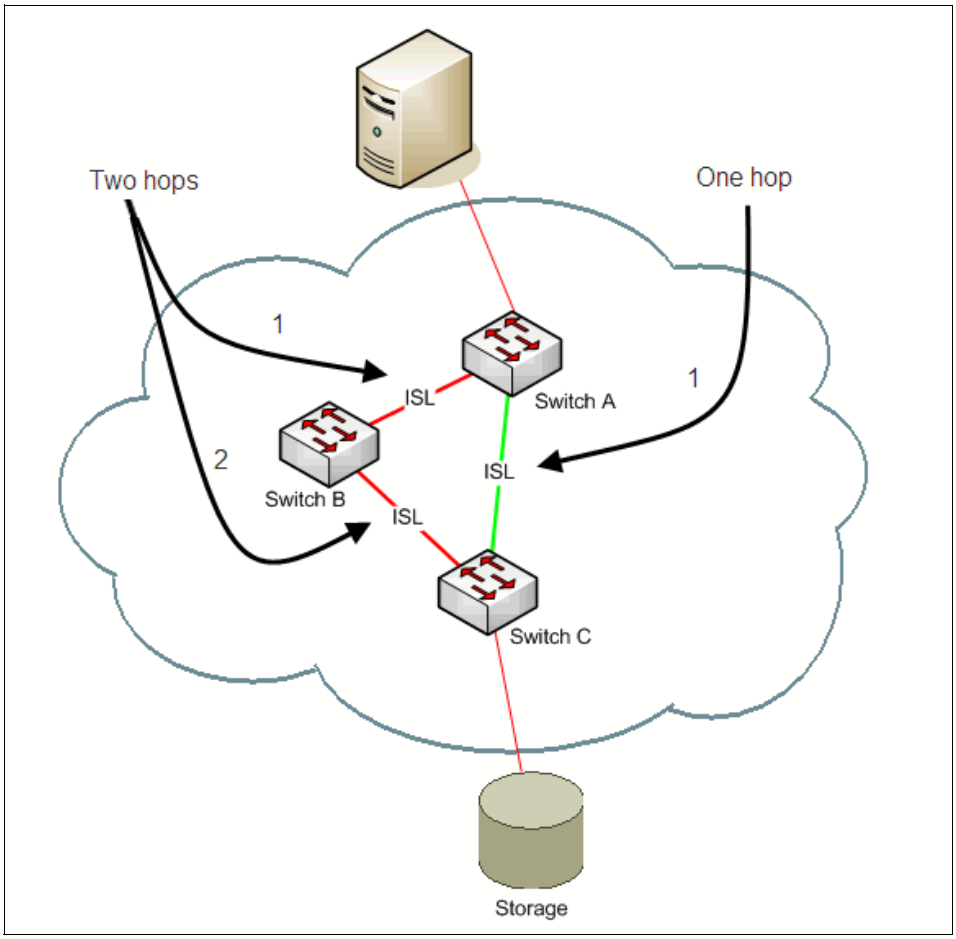

Figure 7-17 Hops explained

FSPF is currently based on the hop count cost.

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

7.5.4 Non-blocking architecture

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

Blocking

Blocking means that the data does not get to the destination. This is not the same as congestion, since data will still be delivered, but with a delay. Currently almost all FC switches are created using non-blocking architecture.

Non-blocking

A non-blocking architecture is now most commonly used by switch vendors. Non-blocking switches enable multiple connections travelling through the switch at the same time. We illustrate this concept in Figure 7-18.

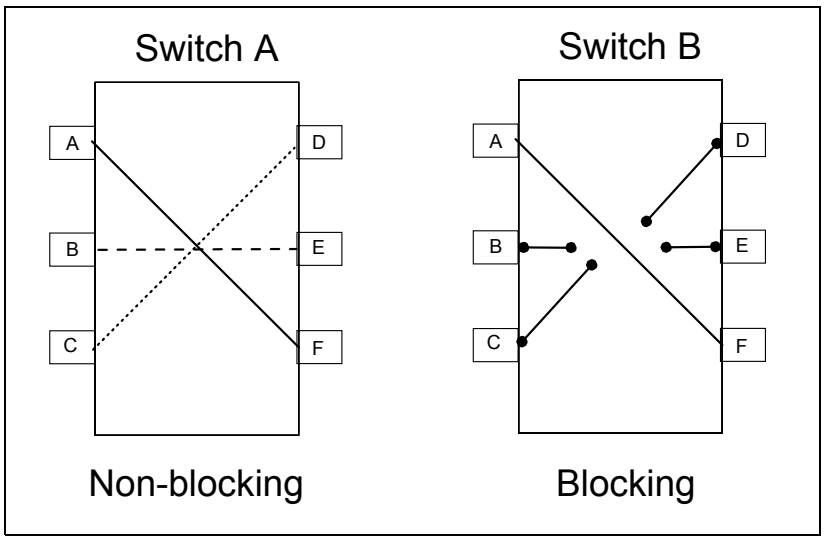

Figure 7-18 Non-blocking and blocking switching

In this example, nonblocking Switch A, port A speaks to port F, port B speaks to E, and C speaks to D without any form of suspension of communication or delay. That is to say, the communication is not blocked. In the blocking switch, Switch B, while port A is speaking to F, all other communication has been stopped or blocked and will not continue until A has finished talking to F.

7.5.5 Latency

Typically, in the SAN world, latency is the time that it takes for a FC frame to traverse the fabric. When we discuss SAN then latency in SAN is rarely taken into the equation since it is in the low microsecond range. This is sometimes confused with disk latency which is the measure on how fast/slow a storage target completes a read or write request sent from server. However when we talk very long distances then all latency, both storage and SAN, plays a significant role.

The more ISLs, the more the latency since the FC frame has to traverse the fabric using ISLs. By fabric, we mean the FC components, and in any latency discussion related to the SAN. Usually the time taken is expressed in microseconds, which gives an indication as to the performance characteristics of the SAN fabric. It will often be given at a switch level, and sometimes a fabric level.

7.5.6 Oversubscription

Another aspect of data flow is ***fan-in ratio*** (also called the oversubscription ratio and frequently the ***fan-out ratio*** from the storage device perspective), both in terms of host ports to target ports and device to ISL. This ratio is the number of device ports that need to share a single port.

For example, two servers each equipped with 4Gb port (4+4=8Gb) are both communicating with a storage device through a single 4Gb port, giving a 2:1 ratio. In other words the total

theoretical input is higher than what that port can provide. Figure 7-19 on page 163 shows a typical oversubscription through an ISL.

This can happen on storage device ports and ISLs. When designing a SAN it is important to consider the possible traffic patterns to determine the possibility of oversubscription, which may result in degraded performance. Oversubscription of an ISL may be overcome by adding a additional ISL between the switches to increase the bandwidth. Oversubscription to a storage device may be overcome by adding additional ports from the storage device to the fabric.

Note: There is a difference on how vendors practice utilization on their stated overall bandwidth per chassis, though both use storage port and ISL oversubscription. Verify with your switch vendor for details on oversubscription best practices.

7.5.7 Congestion

When oversubscription occurs, it leads to a condition called congestion. When a node is unable to utilize as much bandwidth as it would like to, due to contention with another node, then there is a congestion. A port, link, or fabric can be congested. This normally has direct impact to the application in forms of poor performance.

Congestion can be tricky to detect since it can as well be directly related to buffer-to-buffer credit starvation in the switch port, so when looking at the data throughput from the switch it seems like normal or less traffic is flowing through the ports but actually we are keeping the server I/O as hostage since we cannot transport their data due to lack of buffer-to-buffer credits.

7.5.8 Trunking or port-channeling

One means of delivering high availability at the network level is aggregation of multiple physical Inter-Switch Links (ISLs) into a single logical interface. This aggregation allows us to provide link redundancy, greater aggregated bandwidth, and load balancing. Cisco calls this technology *port channeling* others call it simply *trunking*.

We illustrate the concepts of trunking in Figure 7-19.

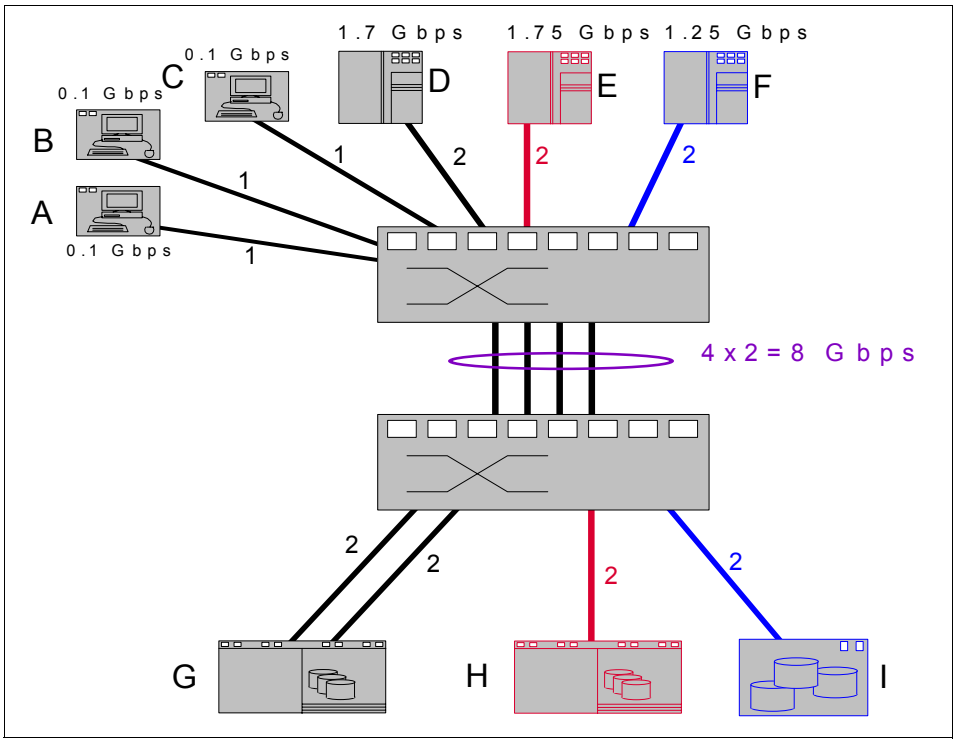

Figure 7-19 Trunking

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

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

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

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

8

Management

Management is one of the key issues behind the concept of infrastructure simplification. The ability to manage heterogeneous systems at different levels as though they were a fully-integrated infrastructure, and offering the system administrator a unified view of the whole SAN, is a goal that many vendors and developers have been striving to achieve.

In this chapter, we look at some of the initiatives that have been, and are being, developed in the field of SAN management, and these will incrementally smooth the way towards infrastructure simplification.

8.1 Management principles

SAN management systems typically comprise a set of multiple-level software components that provide tools for monitoring, configuring, controlling (performing actions), diagnosing, and troubleshooting a SAN. In this section, we briefly describe the different types and levels of management that can be found in a typical SAN implementation, as well as the efforts that are being made towards the establishment of open and general-purpose standards for building interoperable, manageable components.

In this section it is also shown that, despite these efforts, the reality of a “one pill cures all” solution is a long way off. Typically, each vendor and each device has its own form of software and hardware management techniques. These are usually independent of each other, and to pretend that there is one SAN management solution that will provide a single point of control, capable of performing every possible action, would be premature at this stage.

This book does not aim at fully describing each vendor’s own standard(s), but at presenting the reader with an overview of the myriad of possibilities that they might find in the IT environment. That stated, the high-level features of any SAN management solution are likely to include most, if not all, of the following:

- Capacity management
- Device management
- Fabric management
- Pro-active monitoring
- Fault isolation and troubleshooting
- Centralized management
- Remote management
- Performance management
- Security and standard compliant

8.1.1 Management types

There are essentially two philosophies used for building management mechanisms: *in-band management* and *out-of-band management*. They can be defined as:

In-band management
: This means that the management data, such as status information, action requests, events, and so on, flows through the same path as the storage data itself.

Out-of-band management
: This means that the management data flows through a dedicated path, therefore not sharing the same physical path used by the storage data.

In-band and out-of-band models can be illustrated as shown in Figure 8-1. These models are not mutually exclusive. In many environments a combination of both may be desired.

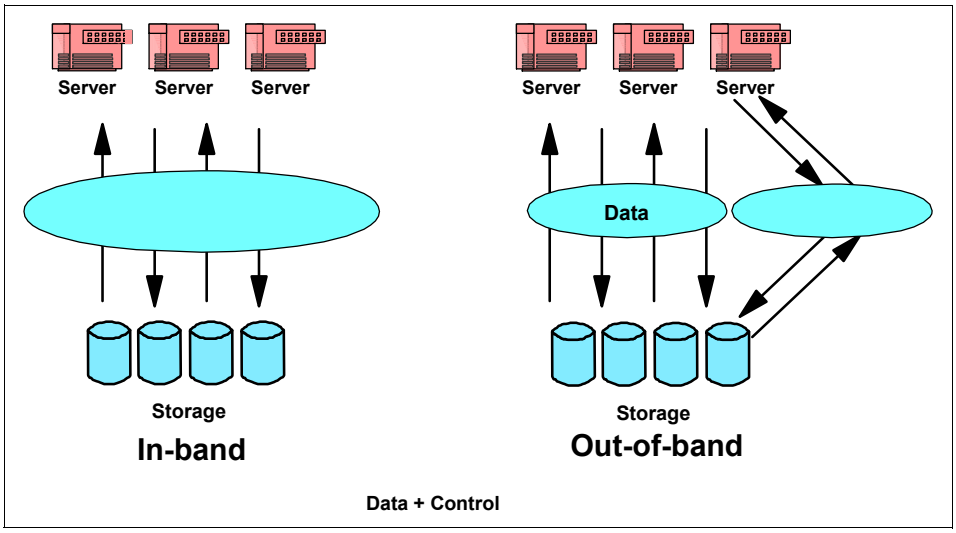

Figure 8-1 In-band and out-of-band models

The in-band approach is simple to implement, requires no dedicated channels (other than LAN connections) and has inherent advantages, such as the ability for a switch to initiate a SAN topology map by means of queries to other fabric components. However, in the event of a failure of the Fibre Channel transport itself, the management information cannot be transmitted. Therefore access to devices is lost, as is the ability to detect, isolate, and recover from network problems. This problem can be minimized by a provision of redundant paths between devices in the fabric.

In-band management allows attribute inquiries on storage devices and configuration changes for all elements of the SAN. Since in-band management is performed over the SAN itself, administrators are not required to manage any additional connections.

On the other hand, out-of-band management does not rely on the storage network; its main advantage is that management commands and messages can be sent even if a loop or fabric link fails. Integrated SAN management facilities are more easily implemented. However, unlike in-band management, it cannot automatically provide SAN topology mapping.

In summary, we can say that In-band management has these main advantages:

- Device installation, configuration, and monitoring
- Inventory of resources on the SAN
- Automated component and fabric topology discovery
- Management of the fabric configuration, including zoning configurations
- Health and performance monitoring

Out-of-band management has these main advantages:

- It keeps management traffic out of the FC path, so it does not affect the business-critical data flow on the storage network.
- It makes management possible, even if a device is down.
- It is accessible from anywhere in the routed network.

8.1.2 Connecting to SAN management tools

A usual way of connecting to a SAN device (by SAN device we mean Fibre Channel switches and storage devices connected to SAN) is by connecting through the ethernet to a storage management device located on a network segment intended for storage devices. This is shown in Figure 8-2 below.

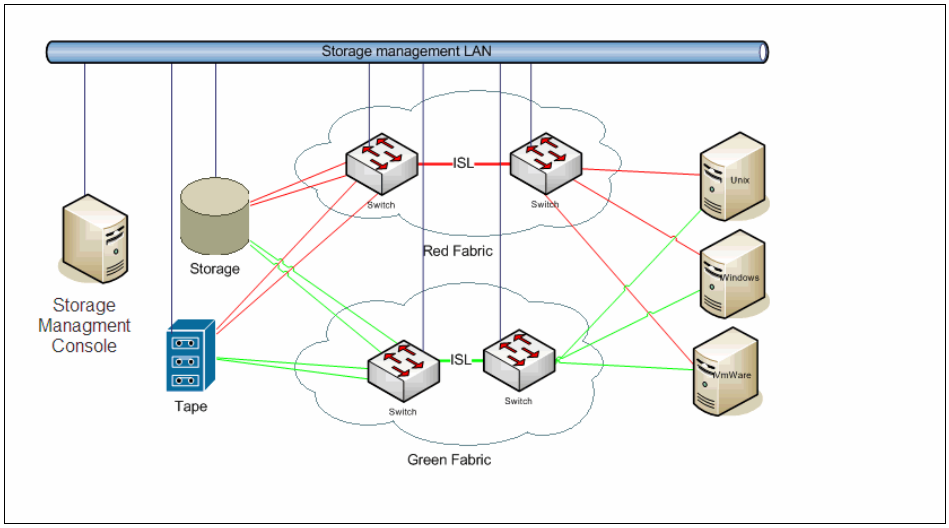

Figure 8-2 Storage Management network

The SAN storage level is comprised of the storage devices that integrate the SAN, such as disks, disk arrays, tapes, and tape libraries. As the configuration of a storage resource must be integrated with the configuration of the server's logical view of them, the SAN storage level management may also span both storage resources and servers.

8.1.3 SAN fault isolation and troubleshooting

In addition to providing tools for monitoring and configuring a SAN, one of the key benefits that a well-designed management mechanism can bring is the ability to efficiently detect, diagnose and solve problems in a SAN.

There are many tools to collect the necessary data in order to perform problem determination and problem source identification (PD/PSI) in a SAN. Generally speaking, these tools offer the ability to:

- Monitor SAN health
- Report failures
- Monitor and identify storage devices
- Monitor the fabric for failures or imminent bottlenecks
- Interpret message and error logs
- Send SNMP traps or syslog messages

Although a well-designed management system can provide invaluable facilities, an easy-to-troubleshoot SAN still relies heavily on a good design, and on good documentation; in terms of PD/PSI, this means that configuration design information is understandable, available at any support level, and is always updated with respect to the latest configuration. There must also be a database where all the information about connections, naming conventions, device serial numbers, WWN, zoning, system applications, and so on, is safely

stored. Last, but not least, there should be a responsible person in charge of maintaining this infrastructure, and monitoring the SAN health status.

8.2 Management interfaces and protocols

In this section we present the main protocols and interfaces that have been developed to support management mechanisms.

8.2.1 SNIA initiative

The Storage Networking Industry Association (SNIA) is using its Storage Management Initiative (SMI) to create and promote adoption of a highly functional interoperable management interface for multivendor storage networking products. The SNIA strategic imperative is to have all storage managed by the SMI interface. The adoption of this interface will allow the focus to switch to the development of value-add functionality. IBM is one of the industry vendors promoting the drive towards this vendor-neutral approach to SAN management.

In 1999, the SNIA and Distributed Management Task Force (DMTF) introduced open standards for managing storage devices. These standards use a common protocol called the Common Information Model (CIM) to enable interoperability. The Web-based version of CIM (WBEM) uses XML to define CIM objects and process transactions within sessions. This standard proposes a CIM Object Manager (CIMOM) to manage CIM objects and interactions. CIM is used to define objects and their interactions. Management applications then use the CIM object model and XML over HTTP to provide for the management of storage devices. This enables central management through the use of open standards.

SNIA uses the xmlCIM protocol to describe storage management objects and their behavior. CIM allows management applications to communicate with devices using object messaging encoded in xmlCIM.

The Storage Management Interface Specification (SMI-S) for SAN-based storage management provides basic device management, support for copy services, and virtualization. As defined by the standard, the CIM services are registered in a directory to make them available to device management applications and subsystems.

For more information about SMI-S go to:

http://www.snia.org

Open storage management with CIM

SAN management involves configuration, provisioning, logical volume assignment, zoning, and Logical Unit Number (LUN) masking, as well as monitoring and optimizing performance, capacity, and availability. In addition, support for continuous availability and disaster recovery requires that device copy services are available as a viable failover and disaster recovery environment. Traditionally, each device provides a command line interface (CLI) as well as a graphical user interface (GUI) to support these kinds of administrative tasks. Many devices also provide proprietary APIs that allow other programs to access their internal capabilities.

For complex SAN environments, management applications are now available that make it easier to perform these kinds of administrative tasks over a variety of devices.

The CIM interface and SMI-S object model adopted by SNIA provide a standard model for accessing devices, which allows management applications and devices from a variety of vendors to work with each other's products. This means that customers have more choice as to which devices will work with their chosen management application, and which management applications they can use with their devices.

IBM has embraced the concept of building open standards-based storage management solutions. IBM management applications are designed to work across multiple vendors' devices, and devices are being CIM-enabled to allow them to be controlled by other vendors' management applications.

CIM Object Manager

The SMI-S standard designates that either a proxy or an embedded agent may be used to implement CIM. In each case, the CIM objects are supported by a CIM Object Manager (CIMOM). External applications communicate with CIM via HTTP to exchange XML messages, which are used to configure and manage the device.

In a proxy configuration, the CIMOM runs outside of the device and can manage multiple devices. In this case, a *provider* component is installed into the CIMOM to enable the CIMOM to manage specific devices.

The providers adapt the CIMOM to work with different devices and subsystems. In this way, a single CIMOM installation can be used to access more than one device type, and more than one device of each type on a subsystem.

The CIMOM acts as a catcher for requests that are sent from storage management applications. The interactions between catcher and sender use the language and models defined by the SMI-S standard. This allows storage management applications, regardless of vendor, to query status and perform command and control using XML-based CIM interactions.

IBM has developed its storage management solutions based on the CIMOM architecture, as shown in Figure 8-3.

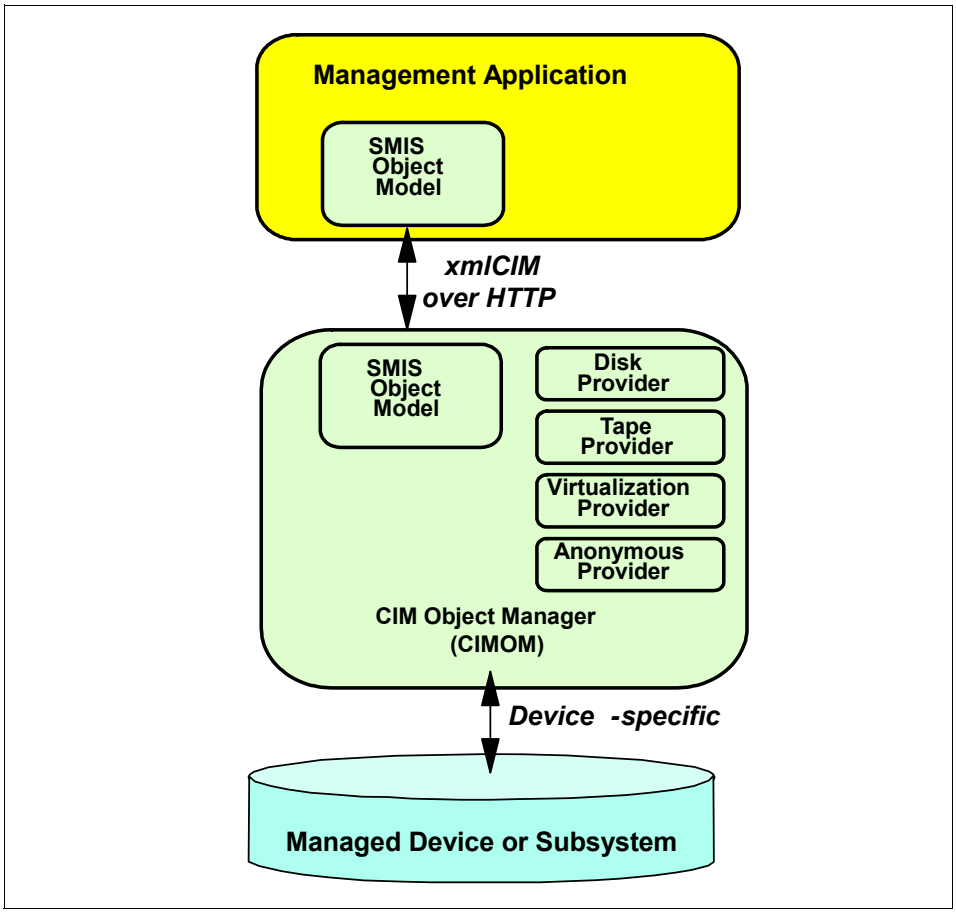

Figure 8-3 CIMOM component structure

8.2.2 Simple Network Management Protocol

SNMP, which is an IP-based protocol, has a set of commands for obtaining the status and setting the operational parameters of target devices. The SNMP management platform is called the SNMP manager, and the managed devices have the SNMP agent loaded. Management data is organized in a hierarchical data structure called the Management Information Base (MIB). These MIBs are defined and sanctioned by various industry associations. The objective is for all vendors to create products in compliance with these MIBs, so that inter-vendor interoperability at all levels can be achieved. If a vendor wants to include additional device information that is not specified in a standard MIB, then that is usually done through MIB extensions.

This protocol is widely supported by LAN/WAN routers, gateways, hubs, and switches, and is the predominant protocol used for multivendor networks. Device status information (vendor, machine serial number, port type and status, traffic, errors, and so on) can be provided to an enterprise SNMP manager. A device can generate an alert by SNMP, in the event of an error condition. The device symbol, or icon, displayed on the SNMP manager console, can be made to turn red or yellow, or any warning color, and messages can be sent to the network operator.

Out-of-band developments

SNMP MIBs are being implemented for SAN fabric elements that allow out-of-band monitoring. The ANSI Fibre Channel Fabric Element MIB provides significant operational and

configuration information on individual devices. The emerging Fibre Channel Management MIB provides additional link table and switch zoning information that can be used to derive information about the physical and logical connections between individual devices.

8.2.3 Service Location Protocol

The Service Location Protocol (SLP) provides a flexible and scalable framework for providing hosts with access to information about the existence, location, and configuration of networked services. Traditionally, users have had to find devices by knowing the name of a network host that is an alias for a network address. SLP eliminates the need for a user to know the name of a network host supporting a service. Rather, the user supplies the desired type of service and a set of attributes that describe the service. Based on that description, the Service Location Protocol resolves the network address of the service for the user.

SLP provides a dynamic configuration mechanism for applications in local area networks. Applications are modeled as clients that need to find servers attached to any of the available networks within an enterprise. For cases where there are many different clients and/or services available, the protocol is adapted to make use of nearby Directory Agents that offer a centralized repository for advertised services.

8.2.4 Vendor-specific mechanisms

These are some of the vendor-specific mechanisms that have been deployed by major SAN device providers.

Application Program Interface

As we know, there are many SAN devices from many different vendors and everyone has their own management and/or configuration software. In addition to this, most of them can also be managed via a command line interface (CLI) over a standard telnet connection, where an IP address is associated with the SAN device, or they can be managed via a RS-232 serial connection.

With different vendors and the many management and/or configuration software tools, we have a number of different products to evaluate, implement, and learn. In an ideal world there would be one product to manage and configure all the actors on the SAN stage.

Application Program Interfaces (APIs) are one way to help this become a reality. Some vendors make the API of their product available for other vendors in order to make it possible for common management in the SAN. This allows for the development of upper level management applications capable of interacting with multiple-vendor devices and offering the system administrator a single view of the SAN infrastructure.

Tivoli Common Agent Services

The Tivoli Common Agent Services is a component designed to provide a way to deploy agent code across multiple end-user machines or application servers throughout an enterprise. The agents collect data from and perform operations on managed resources for Fabric Manager.

The Tivoli Common Agent Services agent manager provides authentication and authorization and maintains a registry of configuration information about the agents and resource managers in the SAN environment. The resource managers are the server components of products that manage agents deployed on the common agent. Management applications use the services of the agent manager to communicate securely with and to obtain information

about the computer systems running the Tivoli common agent software, referred to in this document as the agent.

Tivoli Common Agent Services also provide common agents to act as containers to host product agents and common services. The common agent provides remote deployment capability, shared machine resources, and secure connectivity.

Tivoli Common Agent Services is comprised of these subcomponents:

- Agent manager

  The agent manager is the server component of the Tivoli Common Agent Services that provides functions that allow clients to get information about agents and resource managers. It enables secure connections between managed endpoints, maintains the database information about the endpoints and the software running on those endpoints, and processes queries against that database from resource managers. It also includes a registration service, which handles security certificates, registration, tracking of common agents and resource managers, and status collection and forwarding.

- Common agent

  The common agent is a common container for all the subagents to run within. It enables multiple management applications to share resources when managing a system.

- Resource manager

  Each product that uses Tivoli Common Agent Services has its own resource manager and subagents. For example, Tivoli Provisioning Manager has a resource manager and subagents for software distribution and software inventory scanning.

Figure 8-4 on page 174 shows the Common Agent topology

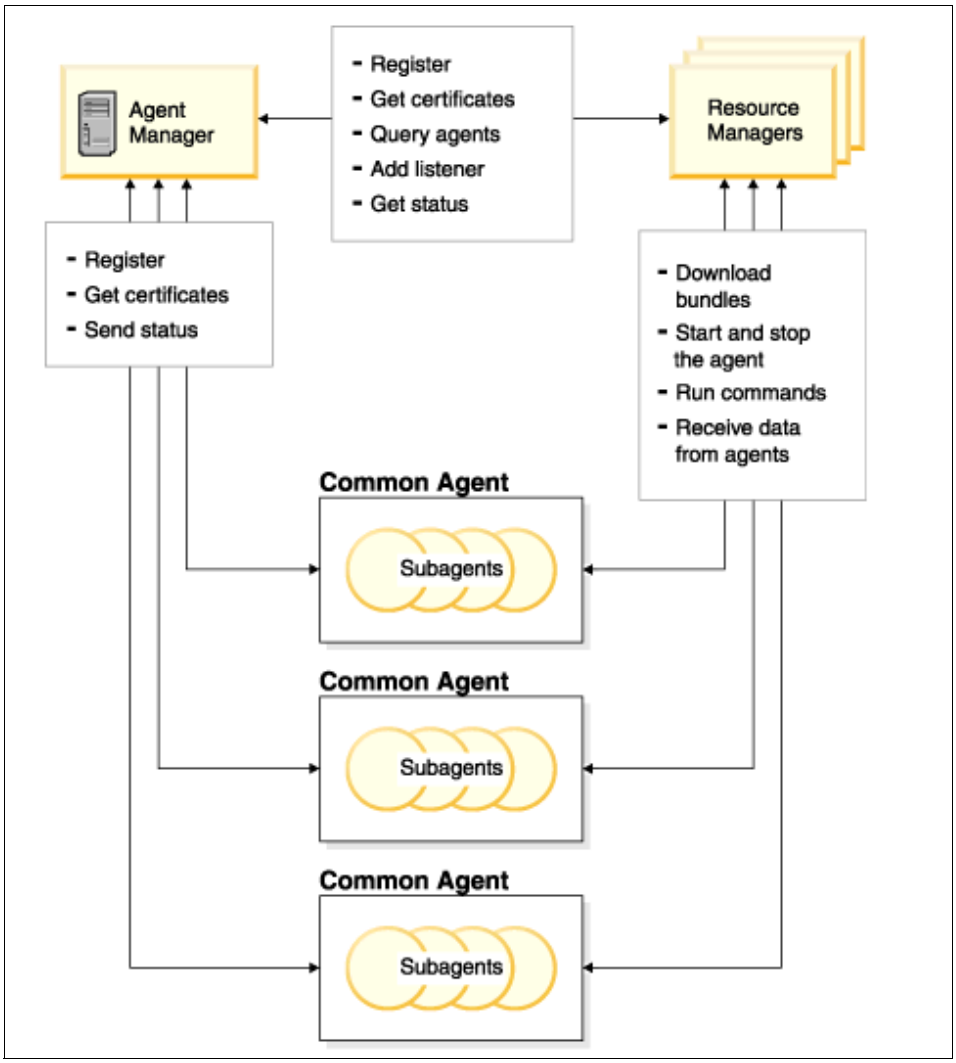

Figure 8-4 Tivoli Common Agent Services

8.3 Management features

SAN management requirements are typified by having a common purpose, but are implemented in different fashions by the different vendors. Some prefer to use Web browser interfaces, some prefer to use embedded agents, and some prefer to use the CLI, and some use a combination of all. There is no right or wrong way. Usually the selection of SAN components is based on a combination of what the hardware and software will provide, not on the ease of use of the management solution. As we have stated previously, the high-level features of any SAN management solution are likely to include most of the following:

- Cost effectiveness
- Open approach
- Device management
- Fabric management
- Pro-active monitoring
- Fault isolation and troubleshooting
- Centralized management
- Remote management
- Adherence to standards
- Resource management
- Secure access
- Standards compliant

8.3.1 Operations

When we talk about management it automatically includes the operational aspects of the environment. The SAN administrators are responsible for all configuration of the SAN switches.

It is quite common that the initial design and creation of a SAN environment includes only a handful of servers and few storage systems but then the environment grows and new technology needs to be added - and at this stage it tends to get more complicated. That is why it is absolutely necessary to ensure that there is comprehensive documentation that documents all aspects of the environment, and it needs to be reviewed on a regular basis to ensure that it is current.

Some of the standards and guidelines that need to be documented are;

- Zoning standards:
  - How to create zones using best practices
  - Naming standards used in the SAN configuration
  - Aliases used
- Volume / LUN allocation standards:
  - Volume characteristics and their uses
  - Allocation rules
- Incident and problem guidelines:
  - How to react in case of an incident
- Roles and responsibilities:
  - Roles and responsibilities with in the team
- SAN and storage installation best practices:

  - Agreed process to install and configure the equipment
- SAN and storage software / firmware upgrade roadmaps:
  - High level overview of how to ensure that the environment is kept current
  - Change schedules
- Monitoring and performance guidelines:
  - What is monitored and how are exceptions handled

8.4 IBM Tivoli Storage Productivity Center

The IBM Tivoli Storage Productivity Center is an integrated hardware and software solution that provides a single point of entry for managing storage devices and other components of your data storage infrastructure and we present a basic product explanation.

This product comes in several different flavors and we will briefly discuss them in this chapter.

- IBM Tivoli Storage Productivity Center Basic Edition
- IBM Tivoli Storage Productivity Center for Data
- IBM Tivoli Storage Productivity Center for Disk
- Tivoli Storage Productivity Center for Disk Select
- Tivoli Storage Productivity Center for Replication
- IBM Tivoli Storage Productivity Center Standard Edition
- SSPC

For detailed information about each product visit:

http://www.ibm.com/systems/storage/software/center/index.html

The IBM TotalStorage Productivity Center is an open storage infrastructure management solution designed to help:

- Reduce the effort of managing complex, heterogeneous storage infrastructures
- Improve storage capacity utilization
- Improve administrative efficiency

TPC provides reporting capabilities, identifying data usage and its location, and provisioning. It also provides a central point of control to move the data based on business needs to more appropriate online or offline storage, and centralizes the management of storage infrastructure capacity, performance and availability.

Figure 8-5 shows the TPC architecture overview

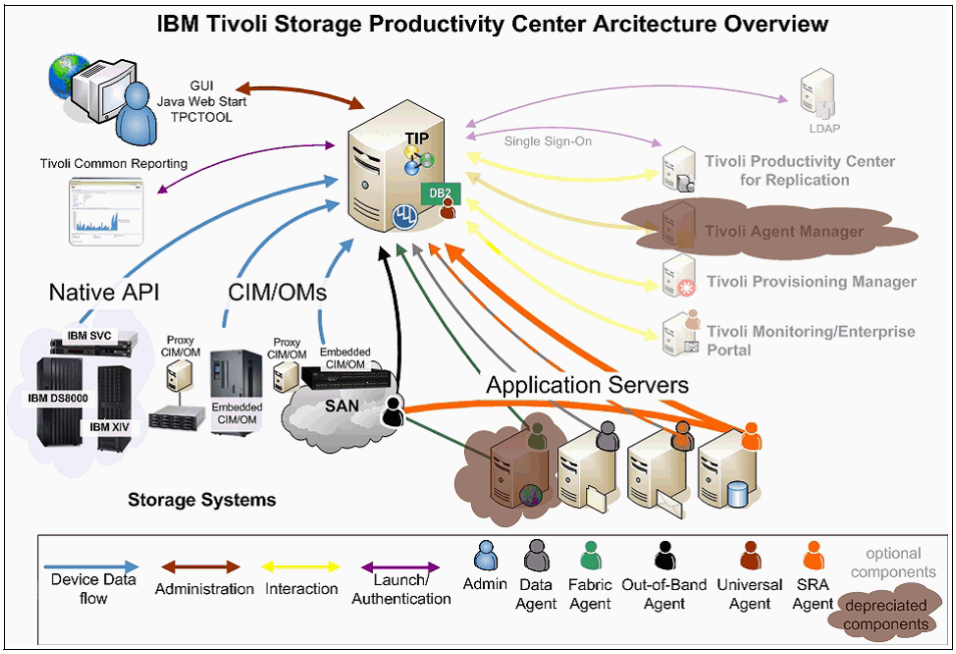

Figure 8-5 TPC architecture

8.4.1 Tivoli Storage Productivity Center for Data

Tivoli Storage Productivity Center for Data provides over 400 enterprise-wide reports, monitoring and alerts, policy-based action and file-system capacity automation in a heterogeneous environment. Tivoli Storage Productivity Center for Data is designed to help improve capacity utilization of file systems and databases and add intelligence to data protection and retention practices.

8.4.2 Tivoli Storage Productivity Center for Disk

Tivoli Storage Productivity Center for Disk is designed to provide storage device configuration and management from a single console. It includes performance capabilities to help monitor and manage performance, and measure service levels by storing received performance statistics into database tables for later use. Policy-based automation enables event action based on business policies. It sets performance thresholds for the devices based on selected performance metrics, generating alerts when those thresholds are exceeded. Tivoli Storage Productivity Center for Disk helps simplify the complexity of managing multiple SAN-attached storage devices.

8.4.3 Tivoli Storage Productivity Center for Disk Select

Tivoli Storage Productivity Center for Disk Select is designed help reduce the complexity of managing storage devices by allowing administrators to configure, manage and monitor performance of their entire storage infrastructure from a single console. Tivoli Storage Productivity Center for Disk Select provides the same features and functions as Tivoli Storage Productivity Center for Disk, but is limited to managing IBM System Storage DS3000, DS4000®, DS5000 and Storwize V7000 and IBM XIV® devices. It provides performance management, monitoring and reporting for these devices.

8.4.4 Tivoli Storage Productivity Center Basic Edition

Tivoli Storage Productivity Center Basic Edition is designed to provide device management services for IBM System Storage DS3000, DS4000, DS5000, and DS8000 products, IBM Storwize V7000, IBM SAN Volume Controller, IBM XIV and heterogeneous storage environments. Productivity Center Basic Edition is a management option available with IBM Storage hardware acquisitions. This tool provides storage administrators a simple way to conduct device management for multiple storage arrays and SAN fabric components from a single integrated console that also is the base of operations for the IBM Tivoli Storage Productivity Center suite.

- Contains a storage topology viewer for a “big picture” perspective
- Offers asset and capacity reporting to improve storage utilization
- Assists with problem determination
- Can reduce storage complexity and improve interoperability
- Automates device discovery
- Extends existing device utilities
- Aids with server consolidation
- Storage Topology Viewer
- The ability to monitor, alert, report, and provision storage
- Status dashboard
- IBM System Storage DS8000 GUI integration with TPC Basic Edition

8.4.5 Tivoli Storage Productivity Center Standard Edition

Tivoli Storage Productivity Center Standard Edition is one of the industry’s most comprehensive storage resource management solutions combining the consolidated benefits of the following three components as one bundle at a reduced price:

- Tivoli Storage Productivity Center for Data
- Tivoli Storage Productivity Center for Disk
- Tivoli Storage Productivity Center Basic Edition

In addition to the benefits and features of Data, Disk and Basic Edition, Tivoli Productivity Center Standard Edition offers additional management, control and performance reporting for the Fibre Channel SAN infrastructure.

Figure 8-6 shows the difference between the basic and standard edition of TPC.

Function	DS Storage Manager	SVC Admin Console	TPC Basic Edition	TPC Standard Edition
Storage Infrastructure Configuration/Status Reporting				
Device Discovery/Configuration	☑	☑	☑	☑
Manage multiple DS8000s / SVCs from 1 User Interface	☑	☑	☑	☑
Topology Viewer and Storage Health Management			☑	☑
Provisioning, including Fabric zoning and Disk LUN assignment			☑	☑
Configuration Management – Highlight configuration changes overtime periods, Best Practice recommendations, Storage configuration planning and recommendations, Security planner				☑
Storage Reporting				
Basic Asset & Capacity Reporting			☑	☑
Storage reporting on the relationships of computers, file systems and DS8000 LUNs/volumes			☑	☑
Capacity Analysis/Predictive Growth				☑
Customized and Detailed Capacity Reporting – including Chargeback and Database Reporting				☑
Performance Management				
Performance Reporting/Thresholds				☑
Volume Performance Advisor – Recommend DS8000 configuration based on performance workloads				☑
Fabric performance reporting and monitor				☑

Figure 8-6 Basic and standard edition matrix

8.4.6 Tivoli Storage Productivity Center for Replication

Tivoli Storage Productivity Center for Replication improves time to value, improves continuous availability of open systems servers, and reduces the downtime of critical applications. It provides the ability to manage the advanced copy services provided by the IBM System Storage DS8000, DS6000™ series, the IBM Enterprise Storage Server® (ESS) Model 800 and the IBM System Storage SAN Volume Controller

8.4.7 What is SSPC?

IBM System Storage Productivity Center (SSPC) is a consolidated storage-management solution designed as a new management console. It provides a single point of management by integrating the functionality of the IBM Tivoli Productivity Center (TPC) with the storage devices element managers in an easy-to-use user interface for management.

Developed as a solution to reduce complexity and improve overall interoperability between SAN components, SSPC also offers a storage topology viewer that acts as a dynamic graphical map of the overall SAN environment and includes information on SAN device relationships, general system health and detailed configuration data that can help storage administrators to better streamline operations.

SSPC is an appliance, it is important to understand that the Operating System where SSPC runs requires regular maintenance no matter if it was originally sold by IBM or deployed on your own.

SSPC represents a simplification as well as an improvement in the way customers can manage their systems. Architecturally, SSPC is basically a Windows technology-based component of IBM System x hardware, delivered pre-loaded with the TPC Basic Edition.It also comes preloaded with TPC Standard Edition software, which requires a separate license purchase.

8.4.8 What can be done from the SSPC?

The complete SSPC offers the following capabilities:

- Pre-installed and tested console: IBM has designed and tested SSPC to support interoperability between server, software, and supported storage devices.
- IBM System Storage DS8000 GUI integration: With TPC V4.2.1, the DS Storage Manager GUI for the DS8000 is integrated with TPC for remote Web access.
- IBM System Storage SAN Volume Controller (console and CIM agent V5.5): These management components of SAN Volume Controller (SVC) are pre-installed on the SSPC along with TPC Basic Edition, which together are designed to reduce the number of management servers.
- Automated device discovery: DS8000 and SVC storage devices can be automatically discovered and configured into TPC environments. These devices are displayed in TPC through a storage topology.
- Asset and capacity reporting: TPC collects asset and capacity information from storage devices on the SAN, which can be kept for historical reporting, forecasting, and used for other tasks, such as analysis and provisioning.

Advanced Topology Viewer: Provides a linked graphical and detailed view of the overall SAN, including device relationships and visual notifications.

8.5 Vendor management applications

Each vendor in the IBM SAN portfolio brings their own bespoke applications to manage and monitor the SAN. In the topics that follow we give a high-level overview of each of them.

8.5.1 b-type

The b-type family switch management framework is designed to support the widest range of solutions, from the very small workgroup SANs up to very large enterprise SANs. The software that Brocade (IBM's valued OEM partner) provides is called Data Center Fabric Manager (DCFM) and Brocade Network Advisor (BNA). This software has been added to the IBM portfolio as IBM Data Center Fabric Manager and IBM Network Advisor.

The following tools can be used with b-type SANs to centralize control and enable automation of repetitive administrative tasks:

- Web Tools:

  A built-in Web-based application that provides administration and management functions on a per switch basis.
- Data Center Fabric Manager (DCFM):

  A client/server-based external application that centralizes management of IBM/Brocade multiprotocol fabrics within and across data centers, including support for FCoE and CEE.
- Fabric Watch:

  A Fabric OS built-in tool that allows the monitoring of key switch elements: power supplies, fans, temperature, error counters, and so on.
- SNMP:

A feature that enables storage administrators to manage storage network performance, find and solve storage network problems, and plan for storage network growth

The following management interfaces allow you to monitor fabric topology, port status, physical status, and other information to aid in system debugging and performance analysis:

- Command-line interface (CLI) through a Telnet connection
- Advanced Web Tools
- SCSI Enclosure Services (SES)
- SNMP applications
- Management server

You can use all these management methods either in-band (Fibre Channel) or out-of-band (Ethernet), except for SES, which can be used for in-band only.

More information about this product can be found here:

http://www.brocade.com/products/all/management-software/product-details/dcfm-enterprise/index.page

8.5.2 Cisco

Fabric Manager and Device Manager are the centralized tools used to manage the Cisco SAN fabric and the devices connected to it. Fabric Manager can be used to manage fabric-wide settings such as zoning but it can manage settings at an individual switch level as well.

Note: As of NX-OS 5.2, Cisco Fabric Manager and FMS will be known as Cisco Data Center Network Manager for SAN, and the LAN-focused Data Center Network Manager will become Data Center Network Manager for LAN. Data Center Network Manager (DCNM) now refers to the converged product.

Cisco DCNM is advanced management software that provides comprehensive life cycle management for the data center LAN and SAN.

Cisco DCNM Release 5.2 combines Cisco Fabric Manager, which previously managed SANs, and Cisco DCNM, which previously managed only LANs, into a unified product that can manage a converged data center fabric. As a part of the product merger in Cisco DCNM Release 5.2, the name Cisco DCNM for SAN replaces the name Cisco Fabric Manager. The name Cisco Fabric Manager still applies to Cisco Fabric Manager Release 5.0(x) and all earlier versions.

Cisco DCNM Release 5.2 supports the Cisco Nexus product family, Cisco MDS 9000 product family, Catalyst 6500 Series, and the Cisco UCS product family.

Fabric Manager provides a high-level summary information about all the switches in a fabric, automatically launching the Web Tools interface when more detailed information is required. In addition, Fabric Manager provides improved performance monitoring over Web Tools alone.

Some of the capabilities of Fabric Manager are:

- Configures and manages the fabric on multiple efficient levels
- Intelligently groups multiple SAN objects and SAN management functions to provide ease and time efficiency in administering tasks

- Identifies, isolates, and manages SAN events across multiple switches and fabrics
- Provides drill-down capability to individual SAN components through tightly coupled Web Tools and Fabric Watch integration
- Discovers all SAN components and views the real-time state of all fabrics
- Provides multi-fabric administration of secure Fabric OS SANs through a single encrypted console
- Monitors ISLs
- Manages switch licenses
- Performs fabric stamping

More detailed information about this product can be found here:

http://www.cisco.com/en/US/products/ps9369/index.html

8.6 SAN multipathing software

In a well-designed SAN, your device will be accessed by the host application over more than one path in order to potentially obtain better performance, and to facilitate recovery in the case of controller, adapter, SFP, cable, switch failure.

Multipathing software provides the SAN with an improved level of fault-tolerance and performance as it provides more than one physical path between the server and storage.

Traditionally, multipathing software would be supplied by each vendor to support its storage arrays. These days there is also the added option to use the multipathing software that often is embedded in the operating system itself. This approach has led to a server-centric approach to multipathing and is independent of the storage array itself. An approach such as this is often easier to implement from a testing and migration viewpoint.

Note: It is important to understand the is a key difference between SAN and storage, though sometime they are referred to as one.

Storage is where you keep your data.

SAN is the network the data travels through between your server and storage.

Figure 8-7 on page 183 shows an example of a dual fabric environment where hosts have multipathing software and can access the storage should a path fail, or should a fabric fail.

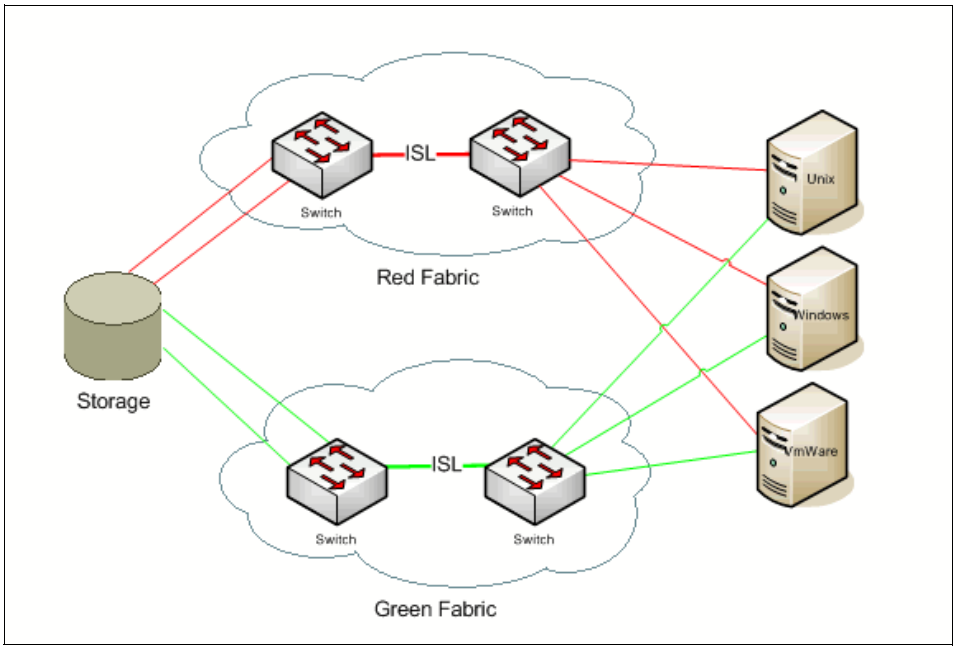

Figure 8-7 SAN overview

In IBMs case the IBM Subsystem Device Driver (SDD) is used and has the following benefits:

- Enhances data availability
- Dynamic input/output (I/O) load-balancing across multiple paths
- Automatic path failover protection
- Concurrent download of licensed machine code

When we think about how many paths should be configured to each volume you should never exceed the supported level given by the storage device. When implementing zoning to a storage device you must make a decision on how many paths you should have. Detailed information about multipath drivers for IBM storage can be found at:

http://www.ibm.com/support/docview.wss?rs=540&context=ST52G7&q=ssg1*&uid=ssg1S7000303&loc=en_US&cs

This is an example of how many paths you will have under different scenarios. In Figure 8-7 on page 183 we have servers that are connected to the SAN with two HBA's and they access their volumes through two storage ports on the storage device. This is controlled by zoning and in our case gives them four working paths for their volumes: two from the Red Fabric and two from the Green Fabric for each server.

Figure 8-8 indicates that we have experienced a single path failure as indicated by the STOP sign.

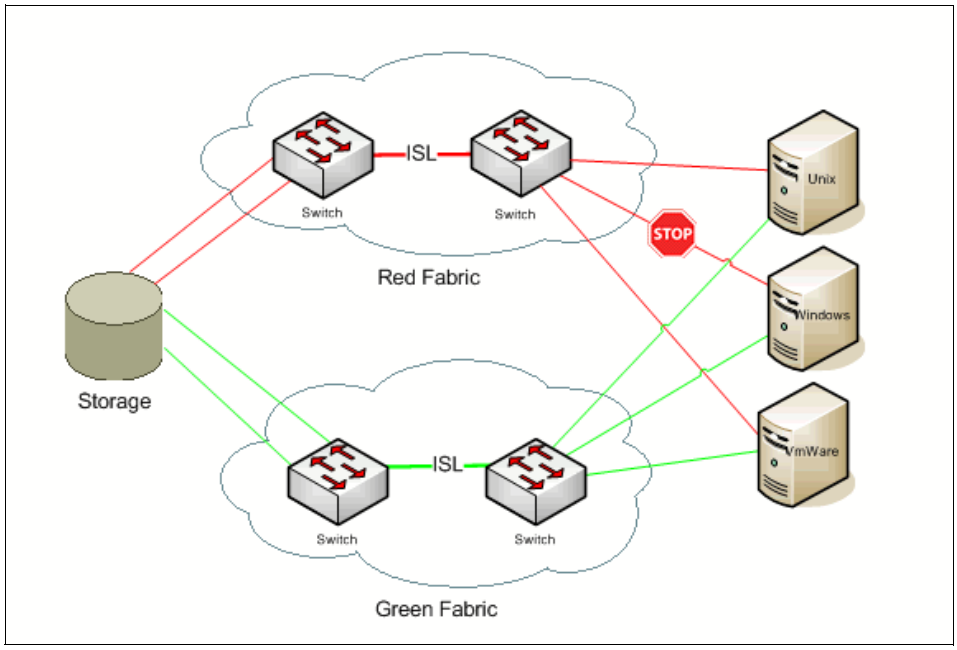

Figure 8-8 HBA failure in a single server

In Figure 8-8 the Windows server has lost connectivity to the SAN and it does not have access to the Red Fabric leaving any longer but we do have working paths through the Green Fabric. All other servers running are running without any issues.

Figure 8-9 on page 184 shows a switch in the Red Fabric has failed.

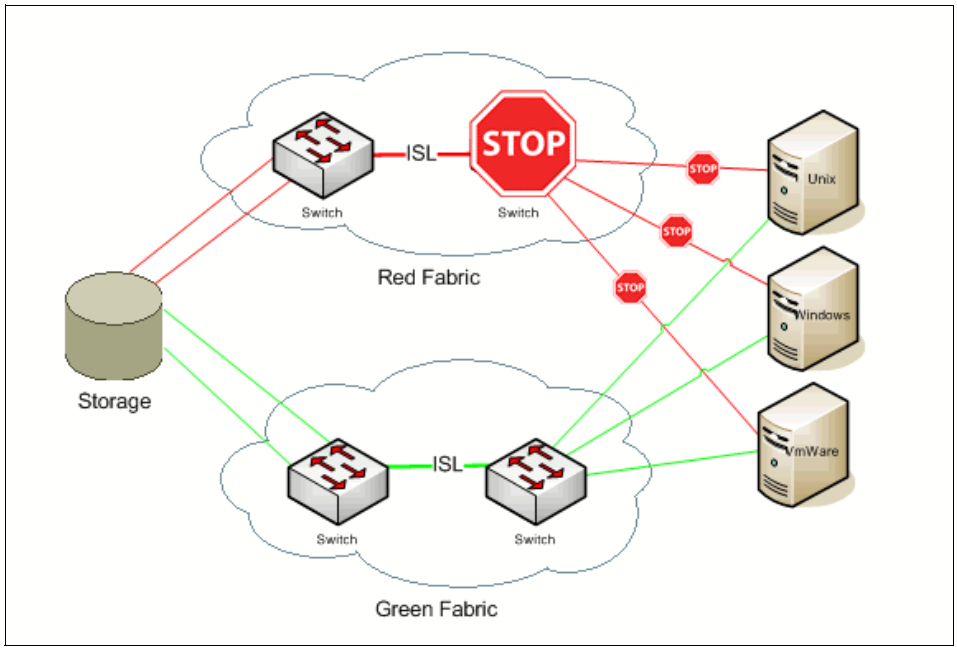

Figure 8-9 A switch is not functional and impacts all attached devices

In Figure 8-9 we have no access to that switch from our servers. The servers will still have working paths through the Green Fabric.

Figure 8-10 on page 185 shows a link from the storage device to a switch has failed in the Red Fabric.

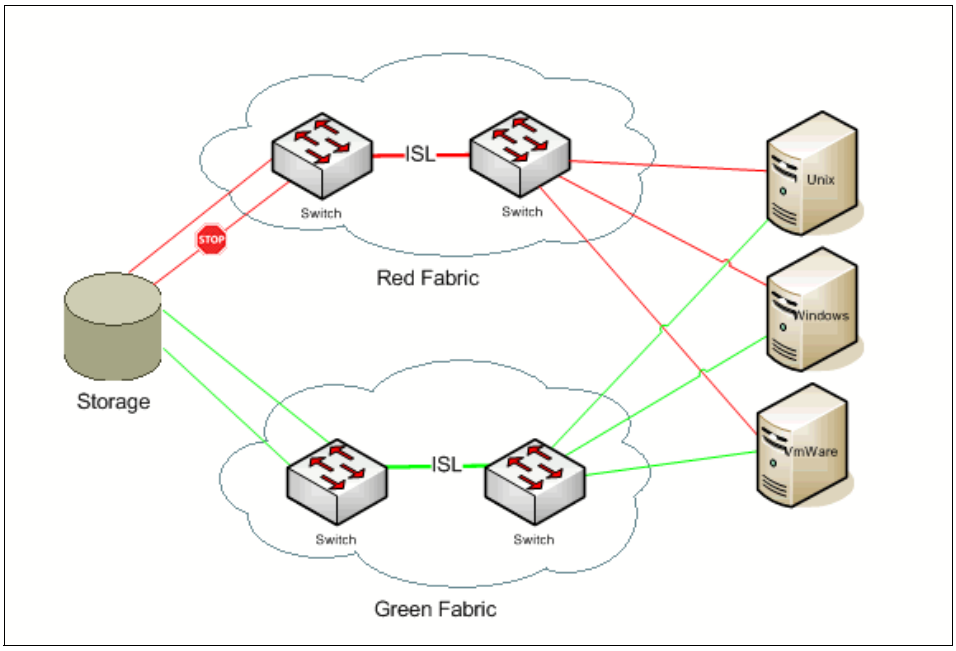

Figure 8-10 Storage device with a single failed connection to a switch

In Figure 8-10 our storage device has lost one of four connections. One connection to the Red Fabric is not functional and therefore all servers using the same storage port will now see three working paths out of of the four possible. All server zoned to the failed storage port will be impacted.

Figure 8-11 on page 185 shows the storage device losing access to the Red Fabric.

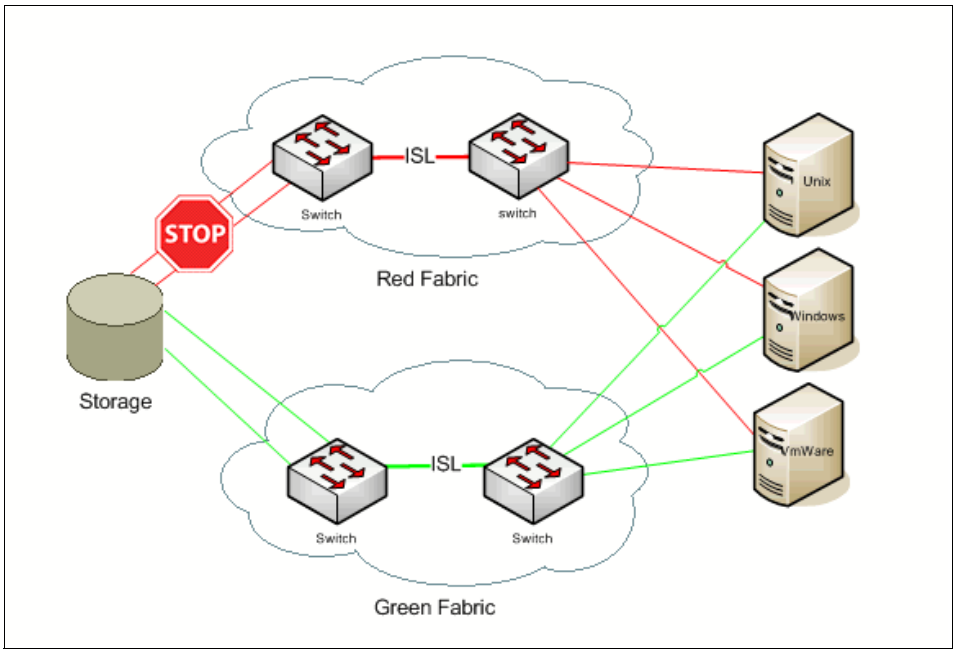

Figure 8-11 Storage device that has lost 2 out of 4 connections to the SAN

In Figure 8-11 our storage device has lost access to the Red Fabric. All devices in the Red Fabric are running normally, it is only these two specific storage ports that have failed. This will leave our servers with two working paths through the Green Fabric. This impacts all servers zoned to these storage ports

Figure 8-12 on page 186 shows the storage device offline.

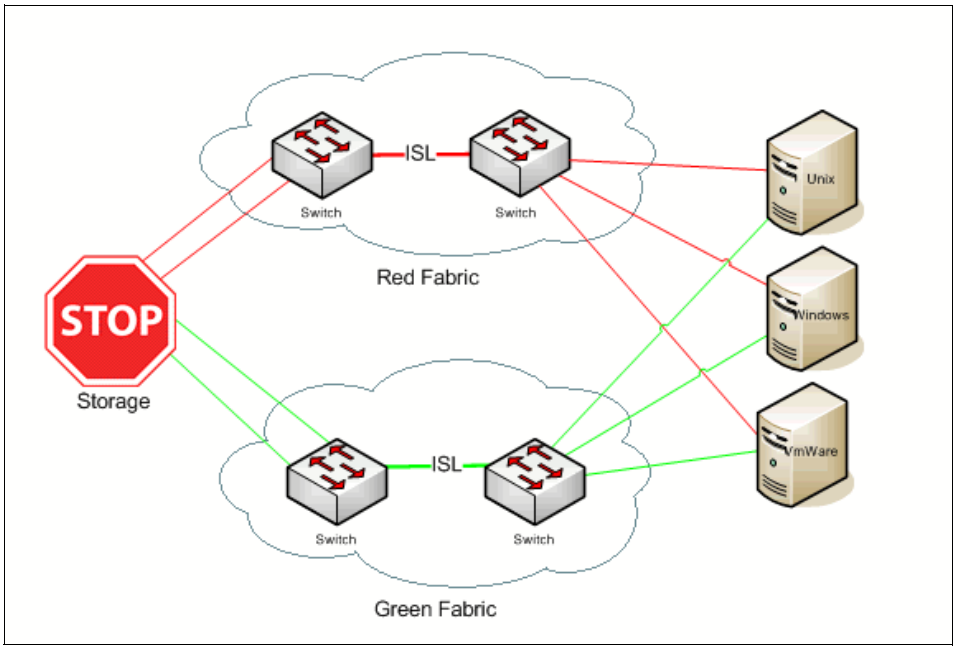

Figure 8-12 The storage device is offline, no working connections

In Figure 8-12 we have lost our storage device. No paths to any volumes on this device are available. No data is accessible. All servers zoned to this storage device are severely impacted and have no access to this storage device.

If a correctly installed and supported version of a multipath driver is installed on the servers we would have “survived” all scenarios expect the last one with the minimum impact.

9

Security

In this chapter we provide an overview of the need for security, the different techniques available, and some of the key points to be aware of.

9.1 Security in the SAN

Security has always been a major concern for networked systems administrators and users. Even for specialized networked infrastructures, such as SANs, special care has to be taken so that information does not get corrupted, either accidentally or deliberately, or fall into the wrong hands. And, we also need to make sure that at a fabric level the correct security is in place, for example, to make sure that a user does not inadvertently change the configuration incorrectly.

Now that SANs have “broken” the traditional direct-attached storage paradigm of servers being cabled directly to servers, the inherent security that this provided has been lost. The SAN and its resources may be shared by many users and many departments. The SAN may be shared by different operating systems that have differing ideas as to who owns what storage. To protect the privacy and safeguard the storage, SAN vendors came up with a segmentation feature to overcome this. This feature is called zoning.

The fabric itself would enforce the separation of data so that only those users intended to have access could communicate with the data they were supposed to.

Zoning, however, does not provide security in that sense, it only implements the means of segregation/isolation. The real security issue is the vulnerability when the data itself has to travel outside of the data center, and over long distances. This will often involve transmission over networks that are owned by different carriers.

We need to look at security from two different angles; for data in flight as explained in 9.4.2, “Data-in-flight” on page 194 and for data at rest explained in 9.4.3, “Data-at-rest” on page 195.

More often than not data is not encrypted when sent from source to target and any information is readable with the correct tools, even though it is a little bit more complicated than simply eavesdropping a telephone line. Since all data is sent at a block level with the Fibre Channel protocol (meaning that all data sent is squeezed into the FC frame before sending) “sniffing” a frame or two could give you 2112 bytes of data. As an example of the difficulty, this would be similar to 1/333.000 of a normal CD or 13 milliseconds of a CD spanning 74 minutes. Obviously this will not give you much information without putting it in the right context.

There is more concern if the whole Fibre Channel port or disk volumes/arrays are mirrored, or tapes containing information end up in the wrong hands. However, to tamper with information from a SAN isn’t something that just happens, as it is something that takes a concerted effort.

The storage architect and administrators need to understand that in a SAN environment, often with a combination of diverse operating systems and vendor storage devices, that some combination of technologies will be required to ensure that the SAN is secure from unauthorized systems and users, whether accidental or deliberate.

In the discussions that follow we briefly explore some of the technologies and their associated methodologies that can be used to ensure data integrity, and to protect and manage the fabric. Each technology has advantages and disadvantages; and each must be considered based on a well thought out SAN security strategy, developed during the SAN design phase.

9.2 Security principles

It is a well-known fact that “a chain is only as strong as its weakest link” and when talking about computer security, the same concept applies: there is no point in locking all the doors and then leaving a window open. A secure, networked infrastructure must protect information at many levels or layers, and have no single point of failure.

The levels of defense need to be complementary, and work in conjunction with each other. If you have a SAN, or any other network for that matter, that crumbles after a single penetration, then this is not a recipe for success.

There are a number of unique entities that need to be given consideration in any environment. We discuss some of the most important ones in the topics that follow.

9.2.1 Access control

Access control can be performed both by means of *authentication* and *authorization* techniques:

Authentication Means that the secure system has to challenge the user (usually by means of a password) so that he or she identifies himself.

Authorization Having identified a user, the system will be able to “know” what this user is allowed to do and what they are not.

As true as it is in any IT environment, it is also true in a SAN environment that access to information, and to the configuration or management tools, must be restricted to only those people that are need to have access, and authorized to make changes. Any configuration or management software is typically protected with several levels of security, usually starting with a user ID and password that must be assigned appropriately to personnel based on their skill level and responsibility.

9.2.2 Auditing and accounting

It is essential that an audit trail is maintained for auditing and troubleshooting purposes, specially when doing Root Cause Analyze (RCA) after a incident has happened. Logs should be inspected on a regular basis and archived.

9.2.3 Data security

Whether at we talk about data at rest or in-flight data, the data security comprises of both data *confidentiality* and *integrity*:

Data confidentiality the system has to guarantee that the information cannot be accessed by unauthorized people, remaining confidential for them but available for only authorized personnel. As shown in the next section, this is usually accomplished by using data *encryption*.

Data integrity the system has to guarantee that the data is stored or processed within its boundaries, that is not altered or tampered with in any way.

This is a security and integrity requirement aiming to guarantee that data from one application or system does not become overlaid, corrupted, or otherwise destroyed, whether intentionally or by accident, by other applications or systems. This may involve some form of authorization, and/or the ability to fence off one system’s data from another systems.

This has to be balanced with the requirement for the expansion of SANs to enterprise-wide environments, with a particular emphasis on multi-platform connectivity. True cross-platform data sharing solutions, as opposed to data partitioning solutions, are also a requirement. Security and access control also need to be improved to guarantee data integrity.

In the topics that follow, we overview some of the common approaches to securing data that are encountered in the SAN environment. The list is not meant to be an in-depth discussion, but merely an attempt to acquaint the reader with the technology and terminology likely to be encountered when a discussion on SAN security takes place.

9.2.4 Securing a fabric

In this section some of the current methods for securing a SAN fabric are presented.

Fibre Channel Authentication Protocol

The Switch Link Authentication Protocol (SLAP/FC-SW-3) establishes a region of trust between switches. For an end-to-end solution to be effective, this region of trust must extend throughout the SAN, which requires the participation of fabric-connected devices, such as HBAs. The joint initiative between Brocade and Emulex establishes Fibre Channel Authentication Protocol (FCAP) as the next-generation implementation of SLAP. Customers gain the assurance that a region of trust extends over the entire domain.

FCAP has been incorporated into its fabric switch architecture and has proposed the specification as a standard to ANSI T11 (as part of FC-SP). FCAP is a Public Key Infrastructure (PKI)-based cryptographic authentication mechanism for establishing a common region of trust among the various entities (such as switches and HBAs) in a SAN. A central, trusted third party serves as a guarantor to establish this trust. With FCAP, certificate exchange takes place among the switches and edge devices in the fabric to create a region of trust consisting of switches and HBAs.

The fabric authorization database is a list of the WWNs and associated information like domain IDs of the switches that are authorized to join the fabric.

The fabric authentication database is a list of the set of parameters that allows the authentication of a switch within a fabric. An entry of the authentication database holds at least the switch WWN, authentication mechanism Identifier, and a list of appropriate authentication parameters.

Zoning

Initially SAN's did not have any zoning, it was an any-to-any kind of communication, and there was no real access control mechanism to protect storage used by one host from being accessed by another host. When SAN's grew, this became a security risk as SANs became more complex and where running more vital part of the business. To mitigate the risk of unwanted cross communication zoning was invented to isolate communication to devices within the same zone.

Persistent binding

Server-level access control is called persistent binding. Persistent binding uses configuration information stored on the server, and is implemented through the server's HBA driver. The process binds a server device name to a specific Fibre Channel storage volume or logical unit number (LUN), through a specific HBA and storage port WWN. Or, put in more technical terms, it is a host-centric way to direct an operating system to assign certain SCSI target IDs and LUNs.

LUN masking

One approach to securing storage devices from hosts wishing to take over already assigned resources is logical unit number (LUN) masking. Every storage device offers its resources to the hosts by means of LUNs. For example, each partition in the storage server has its own LUN. If the host (server) wants to access the storage, it needs to request access to the LUN in the storage device. The purpose of LUN masking is to control access to the LUNs. The storage device itself accepts or rejects access requests from different hosts. The user defines which hosts can access which LUN by means of the storage device control program. Whenever the host accesses a particular LUN, the storage device will check its access list for that LUN, and it will allow or disallow access to the LUN.

Port binding

To provide a higher level of security, you can also use port binding to bind a particular device (as represented by a WWN) to a given port that will not allow any other device to plug into the port, and subsequently assume the role of the device that was there. The reason for this is that the “rogue” device that was inserted will have a different WWN that the port was bound to.

RBAC or role based access control

A role based access control feature is available in most SAN devices today. By using RBAC you can control user access and user authority in a simple way by giving users only access or permission to execute task that are within the users skill set or job role.

Normally there are three definitions for RBAC:

- Role assignment
- Role authorization
- Permission authorization

Usually each role can contain multiple users and each user can be part of multiple roles. For example, if role1 users are only allowed access to configuration commands, and role2 users are only allowed access to debug commands, then if John belongs to both role1 and role2, he can access configuration as well as debug commands.

These predefined roles in a SAN environment are very important to ensure correct login and access is defined for each user.

9.2.5 Zoning, masking and binding

Although neither of these can be classed as security products or mechanisms, as combining all their functionality together can make the SAN more secure than it would be without them.

9.3 Data security

These standards propose to secure FC traffic between all FC ports and the domain controller.

These are some of the methods that will be used:

- FCPAP refers to Secure Remote Password Protocol (SRP), RFC 2945.
- DH-CHAP refers to Challenge Handshake Authentication Protocol (CHAP), RFC 1994.
- FCSec refers to IP Security (IPsec), RFC 2406.

The FCSec aim is to provide authentication of these entities:

- Node-to-node
- Node-to-switch
- Switch-to-switch

An additional function that may be possible to implement is frame level encryption.

The ability to perform switch-to-switch authentication in FC-SP enables a new concept in Fibre Channel: The secure fabric. Only switches that are authorized and properly authenticated are allowed to join the fabric.

Whereas, authentication in the secure fabric is twofold: The fabric wants to verify the identity of each new switch before joining the fabric, and the switch that is wanting to join the fabric wants to verify that it is connected to the right fabric. Each switch needs a list of the WWNs of the switches authorized to join the fabric, and a set of parameters that will be used to verify the identity of the other switches belonging to the fabric.

Manual configuration of such information within all the switches of the fabric is certainly possible, but not advisable in larger fabrics. And there is the need of a mechanism to manage and distribute information about authorization and authentication across the fabric.

9.4 SAN encryption

What is data encryption? Or symmetric and asymmetric encryption? In-flight data or data at rest? In the topics that follow we will try to explain this terminology and help the reader to understand the fundamentals in encryption and key management.

9.4.1 Basic encryption definition

One of the first questions to answer is do I need encryption? In this section we describe basic encryption, cryptographic terms, and ideas on how you can protect your data.

Encryption is one of the simple ways to secure your data. If the data is stolen, lost, or acquired in any way, it cannot be read without the correct encryption key.

Encryption has been used to exchange information in a secure and confidential way for many centuries. Encryption transforms data that is unprotected (plain or clear text) into encrypted data, or ciphertext, by using a key. It is difficult to “break” ciphertext to change it back to clear text without the associated encryption key.

There are two main types of encryption: symmetric encryption and asymmetric encryption (also called public-key encryption).

Symmetric When the same secret password, or key, is used to encrypt a message and decrypt the corresponding cipher text

Asymmetric When one key is used to encrypt a message and another to decrypt the corresponding cipher text.

A symmetric crypto-system follows a fairly straightforward philosophy: two parties can securely communicate as long as both use the same *cryptographic algorithm* and possess the same secret key to encrypt and decrypt messages. This is the simplest and most efficient way of implementing secure communication, as long as the participating parties are able to securely exchange secret keys (or passwords).

Figure 9-1illustrates symmetric encryption.

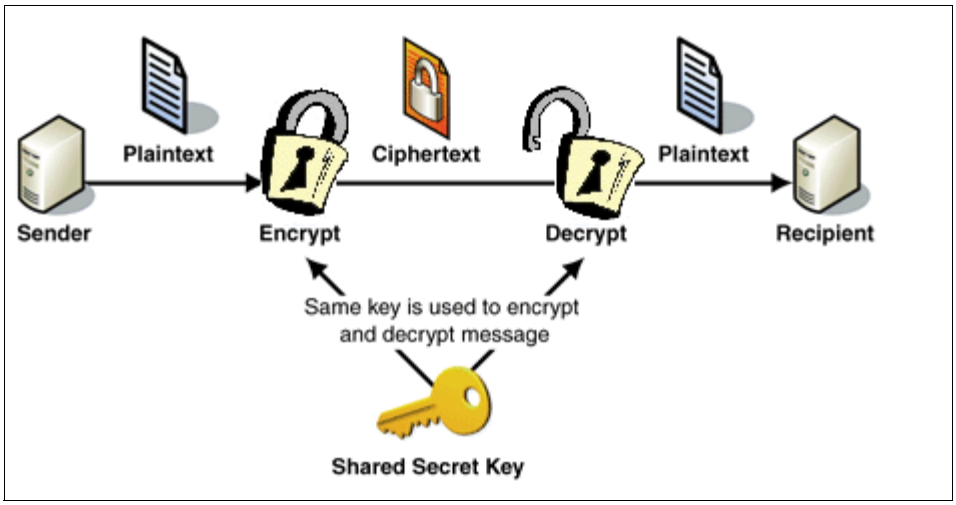

Figure 9-1 Symmetric cryptography

An asymmetric (or public-key) crypto-system is a cryptographic system that uses a pair of unique keys, usually referred to as public and private keys. Each individual is assigned a pair of these keys to encrypt and decrypt information. A message encrypted by one of these keys can only be decrypted by the other key and vice-versa:

- One of these keys is called a “public key” because it is made available to others for use when encrypting information that will be sent to an individual. For example, people can use a person's public key to encrypt information they want to send to that person. Similarly, people can use the user's public key to decrypt information sent by that person.
- The other key is called “private key” because it is accessible only to its owner. The individual can use the private key to decrypt any messages encrypted with the public key. Similarly, the individual can use the private key to encrypt messages, so that the messages can only be decrypted with the corresponding public key.

This means that exchanging keys is not a security concern. An analogy to public-key encryption is that of a locked mailbox with a mail slot. The mail slot is exposed and accessible to the public; its location (the street address) is in essence the public key. Anyone knowing the street address can go to the door and drop a written message through the slot; however, only the person who possesses the key can open the mailbox and read the message.

Figure 9-2 on page 194 illustrates this process.

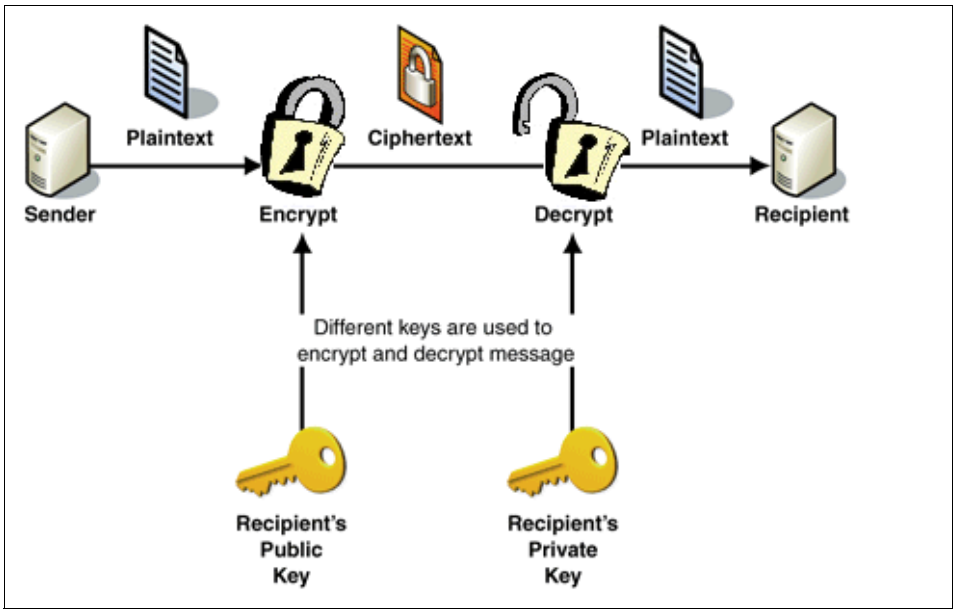

Figure 9-2 Asymmetric cryptography

The main disadvantage of public-key encryption when compared to symmetric encryption is that it demands much higher computing power to be performed as efficiently. For this reason, most of the current security systems use public-key mechanisms as a way to securely exchange symmetric encryption keys between parties, that then use symmetric encryption for their communication. In this case, the exchanged symmetric secret key (or password) is called *session key*.

However, there is still an issue when talking about public-key crypto-systems: when you initially receive someone's public key for the first time, how do you know that this individual is really who he or she claims to be? If “spoofing” someone's identity is so easy, how do you knowingly exchange public keys? The answer is to use a *digital certificate*. A digital certificate is a digital document issue by a trusted institution that vouches for the identity and key ownership of an individual—it guarantees authenticity and integrity.

In the next sections, some of the most common encryption algorithms and tools are presented along with terminology explanations.

9.4.2 Data-in-flight

Also known as data-in-motion, this term generically refers to protecting information any time the data leaves its primary location. For example, when data is transmitted from the source across any type of network to a target. To secure this transmission we technologies such as Secure Sockets Layer (SSL), Virtual Private Networks (VPN) and IP Security (IPSec) to assure data confidentiality. Then we use other technologies such as digital certificates, message authentication codes, and keyed hashes, to ensure data integrity. Data-in-flight is also information (data) that leaves the data center through, for example, open network or leased dark fiber.

All of these areas can be addressed with encryption-based technologies.

9.4.3 Data-at-rest

Protecting data as it resides on the storage media, disk or tape, is typically referred to as data-at-rest.

If encryption is used as part of the strategy for the protection of data-at-rest, this also indirectly addresses the issue of exposed tape media as, even if tapes fall into the wrong hands, the data stored on them is unreadable without the correct key. Of course, all of this assumes that you have enacted the appropriate key management techniques.

To gain the needed security level you will be building layers of security on your SAN. You first increase the level of difficulty for an unauthorized user to even gain access to the data, and then compound that with the fact that private data is not stored in human-readable form.

9.4.4 Digital certificates

If you are using one of these encryption methods, you must also be certain that the person or machine you are sending to is the correct one. When you initially receive someone's public key for the first time, how do you know that this individual is really who the person they claim to be? If “spoofing” someone's identity is so easy, how do you knowingly exchange public keys? The answer is to use a digital certificate. A digital certificate is a digital document issued by a trusted institution that vouches for the identity and key ownership of an individual: it guarantees authenticity and integrity.

There are trusted institutions all over the world that generate trusted certificates. We will use this kind of mechanism also for the first time using a certificate generated by our switch. For more details see 9.4.6, “Key management considerations and security standards”.

9.4.5 Encryption algorithm

After you have decided that encryption is a must, you must also be aware that there are several encryption schemes to choose from. The most popular encryption algorithms in use today include the following:

- 3DES
- DES
- AES
- RSA
- ECC
- Diffie-Hellmann
- DSA
- SHA

To get more information about the details of IBM System Storage Data Encryption refer to *IBM System Storage Data Encryption*, SG24-7797, and for an example of how IBM implements encryption on the IBM System Storage SAN Volume Controller, refer to *Implementing the Storwize V7000 and the IBM System Storage SAN32B-E4 Encryption Switch*, SG24-7977.

If we look at the security aspect on its own then we have been focusing on establishing a perimeter of defense around system assets. While securing access to our environments continues to be an important part of security, the typical business cannot afford to lock down its entire enterprise.

Open networks are now commonly used to connect customers, partners, employees, suppliers, and their data. While this offers significant advantages, it raises concerns about how a business protect its information assets and complies with industry and legislative requirements for data privacy and accountability. By using data encryption as a part of the solution a lot of this can be mitigated as explained in next section.

9.4.6 Key management considerations and security standards

An encryption algorithm requires a key to transform the data. All cryptographic algorithms, at least the reputable ones, are in the public domain. Therefore, it is the key that controls access to the data. We cannot emphasize enough that you must safeguard the key to protect the data. A good tool for that purpose is the Tivoli Key Lifecycle Management (TKLM) which we briefly describe in the next section.

Tivoli Key Lifecycle Management

Due to the nature, security, and accessibility of encryption, data that is encrypted is dependent on the security of, and accessibility to, the decryption key. The disclosure of a decryption key to an unauthorized agent (individual person or system component) creates a security exposure in that the unauthorized agent would also have access to the ciphertext that is generated with the associated encryption key.

Furthermore, if all copies of the decryption key are lost (whether intentionally or accidentally), no feasible way exists to decrypt the associated ciphertext, and the data contained in the ciphertext is said to have been cryptographically erased. If the only copies of certain data that exists is cryptographically erased, then access to that data has been permanently lost for all practical purposes.

That is why the security and accessibility characteristics of encrypted data can create considerations for you that do not exist with storage devices that do not contain encrypted data.

The primary reason for using encryption is that data is kept secure from disclosure, or available to others that do not have sufficient authority. At the same time, it must be accessible to any agent that has both the authority and the requirement to gain access.

Two considerations are important in this context:

- Key security

  To preserve the security of encryption keys, the implementation must ensure that no one individual (system or person) has access to all the information required to determine the encryption key.

- Key availability

  To preserve the access to encryption keys, redundancy can be provided by having multiple independent key servers that have redundant communication paths to encrypting devices. This ensures that the backup of each key server's data is maintained. Failure of any one key server or any one network will not prevent devices from obtaining access to the data keys needed to provide access to the data.

The sensitivity of possessing and maintaining encryption keys, and the complexity of managing the number of encryption keys in a typical environment, results in a client requirement for a key server. A key server is integrated with encrypting products to resolve most of the security and usability issues associated with key management for encrypted devices. However, you must still be sufficiently aware of how these products interact in order to provide appropriate management of the computer environment

Note: Be aware that even with a key server, generally at least one encryption key, normally called the master key (MK), must be maintained manually. For example, this is the key that manages access to all other encryption keys: a key that encrypts the data used by the key server to exchange keys.

Fundamentally, TKLM works by allowing administrators to connect with storage devices and then create and manage keystores—secure repositories of keys and certificate information used to encrypt and decrypt data—or leverage existing keystores already in place. Over the course of key lifecycle, all management functions, including creation, importation, distribution, backup and archiving are easily accomplished using TKLM's graphic interface, which can be accessed using any standard browser on the network. TKLM thus serves as a central point of control, unifying key management even when different classes of storage devices are involved. More information about TKLM can be found at:

http://www-01.ibm.com/software/tivoli/beat/10212008.html

There are two security standards that are very important to ensuring the integrity of encryption products: FIPS 140 and Common Criteria. The official title for the standard Federal Information Processing Standard 140 (FIPS-140) is Security Requirements for Cryptographic Modules. FIPS 140-2 stands for the second revision of the standard and was released in 2001. Common Criteria has seven Evaluation Assurance Levels (EAL), which were defined in 1999. Together, these standards support a small industry for certifying security products and ensuring the integrity of encryption systems.

9.4.7 b-type encryption methods

IBMs b-type encryption devices are used to encrypt data at rest on a storage media and, starting with FOS 7.0, with 16 Gbps E_Ports in-flight encryption is available. When we talk about storage media that could be either disk or tape.

In-flight encryption

The in-flight encryption and compression feature of Fabric OS allows frames to be encrypted or compressed at the egress point of an ISL between two b-type switches, and then to be decrypted or decompressed at the ingress point of the ISL. This feature uses port-based encryption and compression. It is supported on 16 Gbps E_Ports only.

The purpose of encryption is to provide security for frames while they are in flight between two switches. The purpose of compression is for better bandwidth use on the ISLs, especially over long distance. An average compression ratio of 2:1 is provided. Frames are never left in an encrypted or compressed state when delivered to an end device and both ends of the ISL must terminate at 16 Gbps ports.

Figure 9-3 shows the b-type in-flight encryption architecture.

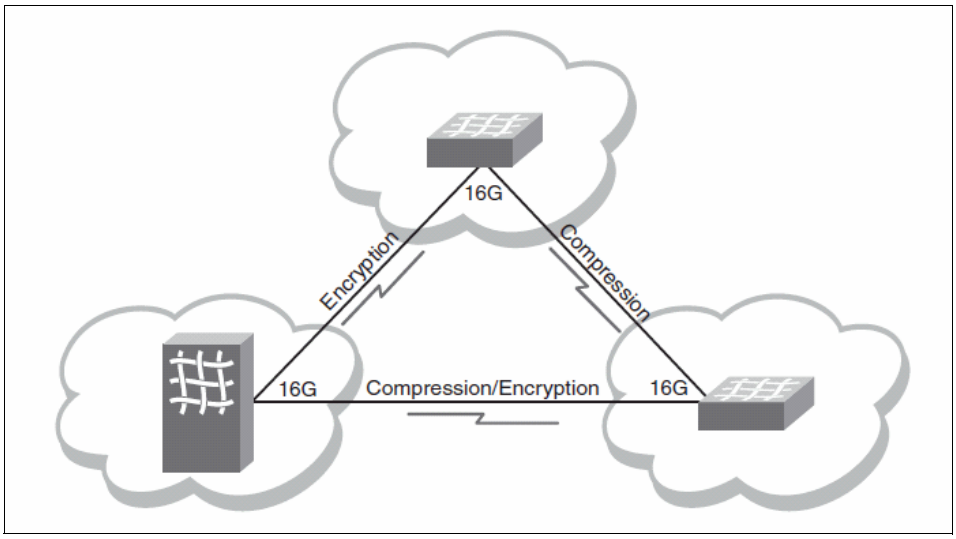

Figure 9-3 In-flight architecture

Encryption at rest

b-type fabric-based encryption solutions work transparently with heterogeneous servers, tape libraries and storage subsystems. While host-based encryption works only for a given operating system and storage-based encryption works only for a given vendor, b-type products are deployed in the core of the fabric to encrypt Fibre Channel-based traffic. Users deploy b-type encryption solutions via either the FS8-18 Encryption Blade or the 2U, rack-mounted IBM SAN32B-E4 Encryption Switch.

The Device Encryption Key or the DEK is very important, since it is needed to encrypt and decrypt the data; it needs to be random and 256 bits in length. b-type encryption devices use a True Random Number Generator (TRNG) to generate each DEK. For encrypting data destined for a disk drive, one DEK is associated with one Logical Unit (LUN). The Institute of Electrical and Electronic Engineers 1619 (IEEE 1619) standard on encryption algorithms for disk drives is known as AES256-XTS. The encrypted data from the AES256-XTS algorithm is the same length as the un-encrypted data, so that the b-type encryption device can encrypt the data block by block without expanding the size of the data. The key management is done using external software such as TKLM.

Figure 9-4 shows a simple b-type encryption setup.

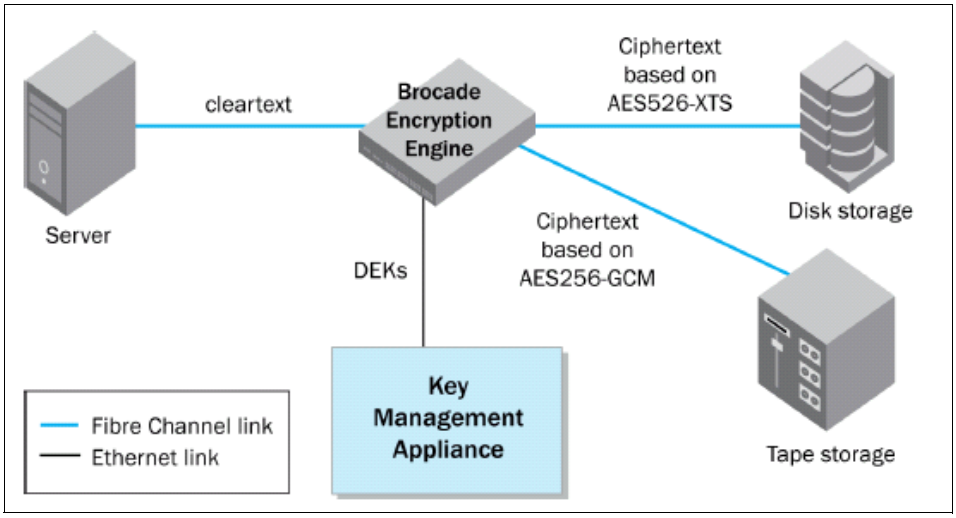

Figure 9-4 b-type encryption and key management

9.4.8 Cisco encryption methods

Cisco has two methods of encrypting SAN information: in-flight encryption and storage media encryption. Both methods are briefly explained and further information is to be found at:

http://www.cisco.com/en/US/prod/collateral/ps4159/ps6409/ps5990/white_paper_c11-545124.html

and:

http://www.cisco.com/en/US/prod/collateral/ps4159/ps6409/ps6028/ps8502/product_data_sheet0900aecd8068ed59.html

In-flight encryption.

Cisco TrustSec Fibre Channel Link Encryption is an extension of the FC-SP standard and uses the existing FC-SP architecture. Fibre Channel data traveling between E_Ports on 8 Gbps modules is encrypted. Cisco uses the 128-bit Advanced Encryption Standard (AES) encryption algorithm and enables either AES-Galois/Counter Mode (AES-GCM) or AES-Galois Message Authentication Code (AES-GMAC). AES-GCM encrypts and authenticates frames, and AES-GMAC authenticates only the frames that are being passed between the two peers. Encryption is performed at line rate by encapsulating frames at egress with encryption using the GCM authentication mode with 128-bit AES encryption. At ingress, frames are decrypted and authenticated with integrity checks.

There are two primary use cases for Cisco TrustSec Fibre Channel Link Encryption. In the first use case, customers are communicating outside the data center over native Fibre Channel (for example, dark fiber, Coarse Wavelength-Division Multiplexing [CWDM] or Dense Wavelength-Division Multiplexing [DWDM]). In the second use case, encryption is performed within the data center for security-focused customers such as defense and intelligence services.

Figure 9-5 shows Cisco TrustSec Fibre Channel Link Encryption

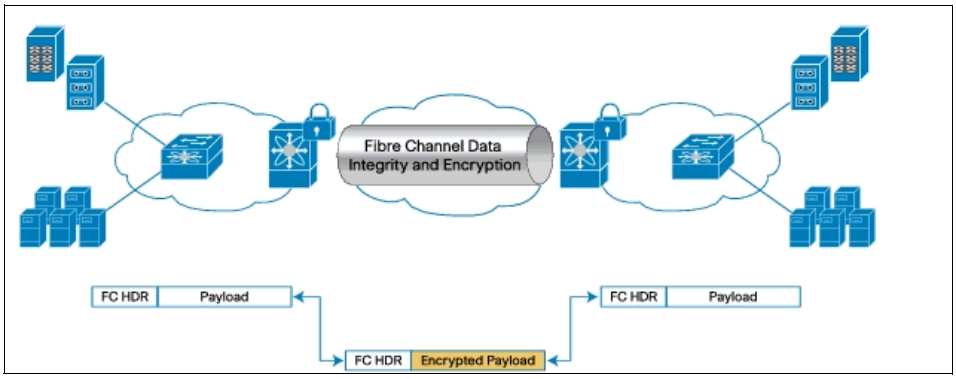

Figure 9-5 Cisco TrustSec encryption

Encryption at rest

Cisco uses Storage Media Encryption (SME) which protects data at rest on heterogeneous tape drives, virtual tape libraries (VTLs), and disk arrays in a SAN environment using highly secure IEEE Advanced Encryption Standard (AES) algorithms.

Encryption is performed as a transparent Fibre Channel fabric service, which greatly simplifies the deployment and management of sensitive data on SAN-attached storage devices. Storage in any virtual SAN (VSAN) can make full use of Cisco SME. Secure lifecycle key management is included, with essential features such as key archival, shredding, automatic key replication across data centers, high-availability deployments, and export and import for single- and multiple-site environments. Provisioning and key management for Cisco SME are both integrated into Cisco Fabric Manager/ DCNM; no additional software is required for key management.

Figure 9-6 on page 200 shows the SME architecture.

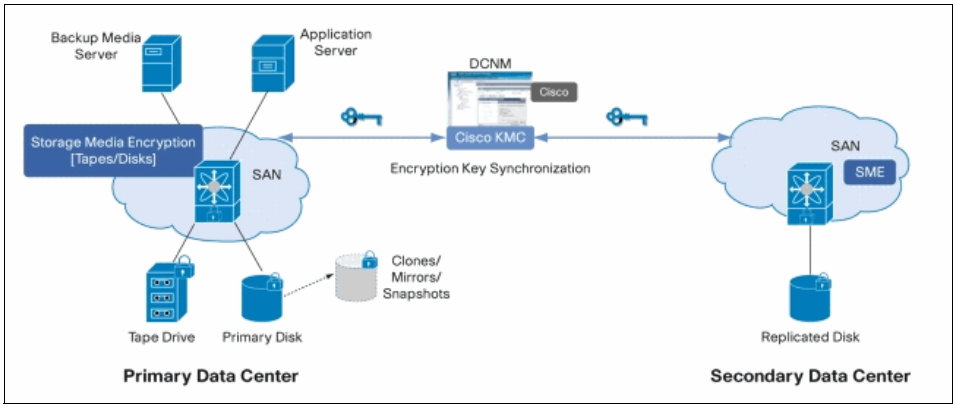

Figure 9-6 SME architecture

9.5 Encryption standards and algorithms

Some of the most popular encryption algorithms in use today are:

AES Advanced Encryption Standard (AES) is a symmetric 128-bit block data encryption technique developed by Belgian cryptographers Joan Daemen and Vincent Rijmen. The U.S. government adopted the algorithm as its encryption technique in October 2000, replacing the DES encryption it used. AES works at multiple network layers

simultaneously. The National Institute of Standards and Technology (NIST) of the U.S. Department of Commerce selected the algorithm, called Rijndael (pronounced Rhine Dahl or Rain Doll), out of a group of five algorithms under consideration. AES is the first publicly accessible and open cipher approved by the National Security Agency (NSA) for top secret information.

RSA The RSA algorithm involves three steps: key generation, encryption and decryption.This algorithm was created by Ron Rivest, Adi Shamir and Len Adleman at MIT; the letters RSA are the initials of their surnames in 1977. It was the first algorithm known to be suitable for digital signing as well as data encryption, and one of the first great advances in public key cryptography. RSA is still widely used in electronic commerce protocols, and is believed to be secure given sufficiently long keys and the use of up to date implementations.

ECC Elliptic curve cryptography is an approach to public-key cryptography based on the mathematics of elliptic curves over finite fields. The use of elliptic curves in cryptography was suggested independently by Neal Koblitz and Victor S. Miller in 1985. Elliptic curves are also used in several integer factorization algorithms that have applications in cryptography, such as, for instance, Lenstra elliptic curve factorization, but this use of elliptic curves is not usually referred to as “elliptic curve cryptography.”

Diffie-Hellman Diffie-Hellman (D-H) key exchange is a cryptographic protocol which allows two parties that have no prior knowledge of each other to jointly establish a shared secret key over an insecure communications channel. This key can then be used to encrypt subsequent communications using a symmetric key cipher.

DSA The Digital Signature Algorithm (DSA) is a United States Federal Government standard for digital signatures. It was proposed by the National Institute of Standards and Technology (NIST) in August 1991 for use in their Digital Signature Standard (DSS), specified in FIPS 186, adopted in 1993. A minor revision was issued in 1996 as FIPS 186-1, and the standard was expanded further in 2000 as FIPS 186-2 and again in 2009 as FIPS 186-3. DSA is covered by U.S. Patent 5,231,668, filed July 26, 1991, and attributed to David W. Kravitz, a former NSA employee.

SHA The Secure Hash Algorithm (SHA) family is a set of related cryptographic hash functions. The most commonly used function in the family, SHA-1, is employed in a large variety of popular security applications and protocols, including TLS, SSL, PGP, SSH, S/MIME, and IPSec. The algorithm has also been used on Nintendo's Wii gaming console for signature verification when booting.

9.6 Security common practices

As we said before, you may have the most sophisticated security system installed in your house — but it is not worth anything if you leave the window open. Some of the security best practices at a high level, that you would expect to see at the absolute minimum, are:

- Default configurations and passwords should be changed.

- Configuration changes should be checked and double checked to ensure that only the data that is supposed to be accessed can be accessed.
- Management of devices usually takes a “telnet” form—with encrypted management protocols being used.
- Remote access often relies on unsecured networks. Make sure that the network is secure and that some form or protection is in place to guarantee only those with the correct authority are allowed to connect.
- Make sure that the operating systems that are connected are as secure as they ought to be, and if the operating systems are connected to an internal and external LAN, that this cannot be exploited. Access may be gotten by exploiting loose configurations.
- Assign the correct roles to administrators.
- Ensure the devices are in physically secure locations.
- Make sure the passwords are changed if the administrator leaves. Also ensure they are changed on a regular basis.

Finally, the SAN security strategy in its entirety must be periodically addressed as the SAN infrastructure develops, and as new technologies emerge and are introduced into the environment.

These will not absolutely guarantee that your information is 100 percent secure, but they will go some way to ensuring that all but the most ardent “thieves” are kept out.

10

Solutions

The added value of a SAN lies in the exploitation of its technology to provide tangible and desirable benefits to the business using fast, secure, reliable and highly available networking solutions. Benefits range from increased availability and flexibility to additional functionality that can reduce application downtime.

In this chapter we provide a description of general SAN applications, and the types of components required to implement them. There is far more complexity than is presented here. For instance, this text will not cover how to choose one switch over another, or how many ISLs are necessary for a given SAN design. These strategic decisions must be always taken by experienced IT architects, and that is beyond the intended scope of this book. We introduce the basic principles and key considerations that must be taken to choose an optimal solution for SAN deployments.

10.1 Introduction

During last few years and with the continued development of the communication and computing technologies and products, Storage Area Networks have evolved since their inception and are getting much more complex. Nowadays we are not talking only about simple fiber-optic connection between SAN devices such as SAN switches, routers, tape drives, disk device subsystems, and target host systems using standard Fibre Channel Host Bus Adapters (HBA). It has moved way beyond that and will continue to do so.

Today, we are looking for solutions that enable us to increase the data transfer rate within the most complex datacenters, provide high availability of managed applications and systems, implement data security, and storage efficiency, and all while reducing the associated costs and power consumption.

We have to find a smooth, effective, and cost-efficient way to migrate our current or legacy SAN infrastructure to the less complex, but more powerful and flexible datacenter of the next generation.

There are many categories in which SAN solutions can be classified. We have chosen to classify ours as: infrastructure simplification, business continuity and information lifecycle management. In the topics that follow we discuss the use of basic SAN design patterns to build solutions for different requirements, ranging from simple data movement techniques that are frequently employed as a way to improve business continuity, up to sophisticated storage pooling techniques that are used to simplify complex infrastructures.

Before we do that we will present some basic principles to be considered when planning a SAN implementation, or upgrade.

10.2 Basic solution principles

A number of important decisions need to be made by the system architect either when a new SAN is being designed or when an existing SAN is being expanded; such decisions usually refer to the choice of the connectivity technology, the best practices for adding capacity to a SAN or the more suitable technology for achieving data integration. This section discusses some of these aspects.

10.2.1 Connectivity

Connecting servers to storage devices through a SAN fabric is often the first step taken in a phased SAN implementation. Fibre Channel attachments have the following benefits:

- Running SCSI over Fibre Channel for improved performance
- Extended connection distances (sometimes called remote storage)
- Enhanced addressability

Many implementations of Fibre Channel technology are simple configurations that remove some of the restrictions of existing storage environments, and allow you to build one common physical infrastructure. The SAN uses common cabling to the storage and the other peripheral devices. The handling of separate sets of cables, such as OEMI, ESCON, SCSI single-ended, SCSI differential, SCSI LVD, and others have caused the IT organization management much trauma as it attempted to treat each of these differently. One of the biggest problems is the special handling that is needed to circumvent the various distance limitations.

Installations without SANs commonly use SCSI cables to attach to their storage. SCSI has many restrictions, such as limited speed, a very small number of devices that can be attached, and severe distance limitations. Running SCSI over Fibre Channel helps to alleviate these restrictions. SCSI over Fibre Channel helps improve performance and enables more flexible addressability and much greater attachment distances compared to normal SCSI attachment.

A key requirement of this type of increased connectivity is providing consistent management interfaces for configuration, monitoring, and management of these SAN components. This type of connectivity allows companies to begin to reap the benefits of Fibre Channel technology, while also protecting their current storage investments.

The SAN infrastructure flexibility and simplification can be dramatically improved by utilizing Fibre Channel over Ethernet (FCoE), that evolved over the last few years. This enablement can easily replace dedicated switching solutions for LAN and SAN by a single device that is able to transfer both types of data — IP packets and storage data. These deployments we call Converged Networks and in the following topics we will briefly present the basic migration steps to convergency.

10.2.2 Adding capacity

The addition of storage capacity to one or more servers may be facilitated while the device is connected to a SAN. Depending on the SAN configuration and the server operating system, it may be possible to add or remove devices without stopping and restarting the server.

If new storage devices are attached to a section of a SAN with loop topology (mainly tape drives), the Loop Initialization Protocol (LIP) may impact the operation of other devices on the loop. This may be overcome by quiescing operating system activity to all the devices on that particular loop before attaching the new device. This is far less of a problem with the latest generation of loop-capable switches. If storage devices are attached to a SAN by a switch, using the switch and management software it is possible to make the devices available to any system connected to the SAN.

10.2.3 Data movement and copy

Data movement solutions require that data be moved between similar or dissimilar storage devices. Today, data movement or replication is performed by the server or multiple servers. The server reads data from the source device, perhaps transmitting the data across a LAN or WAN to another server, and then the data is written to the destination device. This task ties up server processor cycles and causes the data to travel twice over the SAN - once from source device to a server, and then a second time from a server to a destination device.

The objective of SAN data movement solutions is to avoid copying data through the server, and across a LAN or WAN, thus freeing up server processor cycles and LAN or WAN bandwidth. Today, this data replication can be accomplished in a SAN through the use of an intelligent tools and utilities and between datacenters using for example FCoE protocol on WAN.

The following sections list some of the available copy services functions.

Data migration

One of the critical tasks for SAN administrator is to move data between two independent SAN infrastructures or just from an old storage system, that is being sunset, to the new enterprise and highly performing disk system. There are basically two scenarios - when SANs are

independent and cannot be interconnected together even if they reside in the same datacenter, and when the disk systems can be cross-connected through SAN switches.

Data replication over SAN

In this scenario we are able to interconnect both storage devices (both SANs) together and migrate data directly from an old to the new storage box without any interruption of service or performance impact on the application or host server. This kind of migration we consider as block-level data copy, storage systems do not analyze the data on disks, just split it into blocks and copy those changed or modified. Many storage vendors, including IBM, offer replication services for their disk storage systems as an optional feature of service delivery, usually as a part of backup/recovery solution. Copy services can be even further extended to long distances through WAN to fulfil disaster recovery requirements or just to make application services highly available across geographies.

Figure 10-1 demonstrates how this data (LUN) migration works in principle.

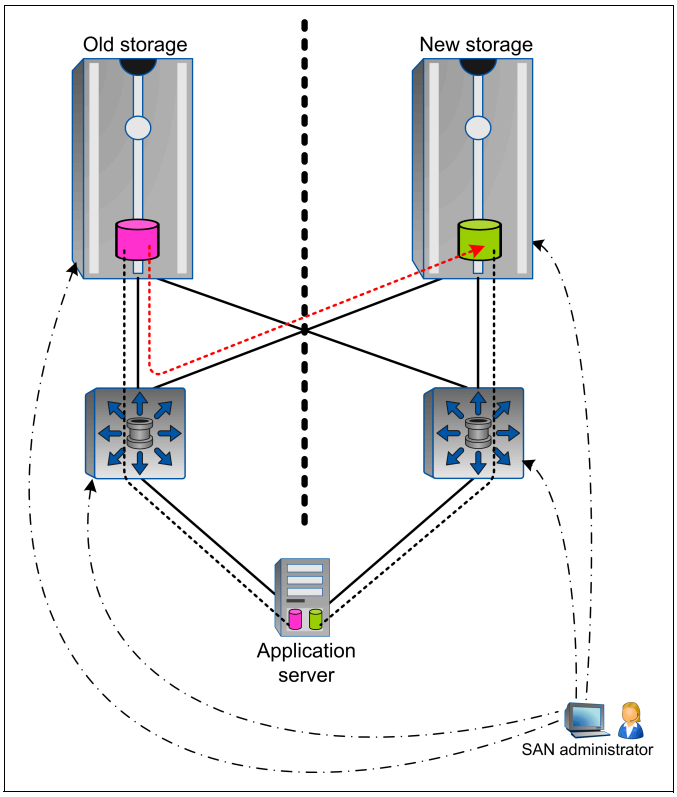

Figure 10-1 SAN based replication of LUNs

In the example the storage administrator is challenged to migrate data to the newly deployed, highly performing disk storage system without interruption to the most critical client's SAP application. Luckily, we have the ability to manage both source and target storage systems and these are configured to communicate together through SAN switches. Disk copy services are able to replicate specific LUNs from the old to the new storage devices and, most importantly, without any performance impact to the SAP application.

In addition, this procedure is often used to prepare a standby application server connected to the replicated disk LUNs, or just to replace the old server hardware where the SAP application is running, all with the minimum outage necessary to switch the application over to the prepared server.

Host-based data migration

Host-based migration of storage data is the option used when the storage administrator is not able to establish connection between the source and target disk storage system. It usually happens in datacenters with two independent SANs, in most cases each managed by different team of administrators or even by different vendors.

The principle of the migration is shown in Figure 10-2.

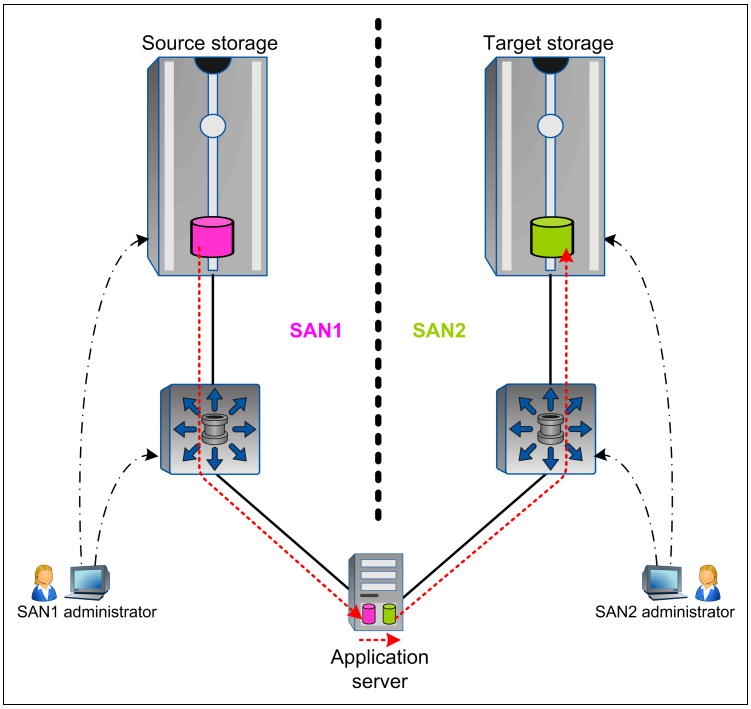

Figure 10-2 Host-based migration of data

The application server is connected to both SANs using independent Host Bus Adapters (HBAs). Application owner and SAN2 administrator analyze the current disk structure assigned from source storage system and the same disk capacity is to be assigned by the SAN2 administrator to the application server. The application or system owner then migrates the data from the source disk to the target disk manually using operating system functions (the application should be offline) or disk mirroring needs to be enabled. Once the data is synchronized between source and target disks, the mirror can be broken, source disks unassigned, and the source storage system disconnected from the application server. The disadvantage of this solution is a significant I/O operation on the source and target LUNs that could potentially impact the performance of critical applications.

Remote data copy and migration

Remote copy or data migration is a business requirement used to protect data from disasters, or to migrate data from one location to avoid application downtime for planned outages such as hardware or software maintenance. Another challenge of remote copy services is to provide a highly available or fault-tolerant infrastructure for business critical systems and applications across datacenters, typically over long distances, sometimes even continents.

Remote copy solutions are either synchronous or asynchronous, and they require different levels of automation in order to guarantee data consistency across disks and disk subsystems. Remote copy solutions are implemented only for disks at a physical or logical volume data block level. There are complete solutions from various vendors to support data

migration projects to optimally schedule and utilize clients' network resources and to eliminate impact on critical production environments. Products such as these help clients efficiently and effectively migrate the whole SAN data volumes from small remote datacenters to the central one across WAN without interruption to the service.

In the future, with more advanced storage management techniques such as outboard hierarchical storage management and file pooling, remote copy solutions would need to be implemented at the file level. This implies more data to be copied, and requires more advanced technologies to guarantee data consistency across files, disks, and tape in multi-server heterogeneous environments. Datacenter networking infrastructure is required to support various data transfer protocols to support these requirements, such as Fibre Channel over Ethernet (FCoE), Converged Enhanced Ethernet (CEE), or just simple iSCSI.

Realtime snapshot copy

Another outboard copy service enabled by Fibre Channel technology is realtime snapshot (also known as T0 or time=zero) copy. This is the process of taking an online snapshot, or freezing the data (databases, files, or volumes) at a certain time, and then allowing applications to update the original data while the frozen copy is duplicated. With the flexibility and extensibility that Fibre Channel brings, these snapshot copies can be made to either local or remote storage devices. The requirement for this type of function is driven by the need for 24x7 availability of key database systems. This solution is optimal in homogenous infrastructures consisting of the devices from a single vendor.

10.2.4 Upgrading to faster speeds

One of the other considerations of any SAN environment is how newer, faster technology is to be introduced. Both 8 Gbps FC and 10 GbE Ethernet products already have a big footprint in the market and participate in datacenter networking. We are now seeing vendors move forward with even faster technologies and products such as 16 Gbps FC ports and Host Bus Adapters. For most applications this will not mean that they can immediately benefit. Applications that have random or "bursty" I/O will not necessarily gain any advantage. Only those applications and systems that stream large amounts of data are likely to see the most immediate benefits. One place that makes sense for 16 Gbps to be utilized is the inter-switch link. This has two advantages: firstly, the increased speed between switches is the obvious one, and secondly the advantage is that it may be possible to have fewer ISLs with the increased bandwidth. This means that it may be possible to reassign ISLs and use them to attach hosts or storage.

Another consideration that must be taken into account is the cost factor. IT architects and investors must evaluate their current SAN solutions in datacenters and take strategic decisions if it is beneficial to continue with the upgrade to a dedicated Fibre Channel solution running 16 Gbps devices, or if it is the right time to think about migration to converged networks and utilize, for example, Fibre Channel protocols over Ethernet (FCoE). There are many products available on the market that support such transformation and transition and protect the clients' investments for the future.

10.3 Infrastructure simplification

High on the list of critical business requirements is the need for IT infrastructures to better support business integration and transformation efforts. At the heart of these datacenter integration and transformation efforts is often the simplification and streamlining of core storage provisioning services and storage networking.

Viewed in the broadest sense, infrastructure simplification represents an optimized view and evolutionary approach (or the next logical step beyond basic server consolidation and virtualization) for companies on the road to becoming true on-demand businesses that are highly competitive in the market.

If your IT infrastructure has become a complex set of disparate, server specific, silo'ed applications operating across an endless sea of servers (i.e., transaction processing servers, database servers, tiered application servers, data gateways, human resource servers, accounting servers, manufacturing servers, engineering servers, e-mail servers, Web servers, etc.), then you need to be able to answer questions such as: “Where can we deploy the next application?” or “Where can we physically put the next server?” or “How can we extend our storage resources?” or “How can we connect more and more virtual or physical servers?” or just “Is there a simpler way to manage this?” We try to answer all these questions in the following topics.

10.3.1 Where does the complexity come from?

A SAN, in theory, is a simple thing and that is a path from a server to a common storage resource. So, where did all the complexity creep in from?

Limited budgets and short-sighted strategic thinking have pushed IT organizations into looking for short term solutions to pain points. When a new application or project appears, the easy, inexpensive option is to add another low cost server. While this “server sprawl” or proliferation of UNIX and Windows Intel servers is an attractive short term solution, the infrastructure costs to support these inexpensive servers often exceeds the purchase price of the server.

Now we also add storage systems to the sprawl. Every server has two or four Host Bus Adapters (HBAs) and a share of the consolidated storage. As we add more servers, we run out of SAN ports, so we add another switch, and then another, and finally another. Now we have “SAN sprawl” with a complex interlinked fabric that is difficult to maintain or change.

To make things more difficult, the servers were probably purchased from multiple vendors, with decisions made on cost, suitability to a specific application, or merely someone’s personal preference. Different vendors’ servers were tested on very specific SAN configurations. Every server producer has its own interoperability matrix or list of SAN configurations that the vendor has tested, and that particular vendor supports. It could be difficult for a SAN administrator to find the appropriate devices and configurations that work together smoothly.

10.3.2 Storage pooling

Before SANs, the concept of the physical pooling of devices in a common area of the computing center was often just not possible, and when it was possible, it required expensive and unique extension technology. By introducing a network between the servers and the storage resources, this problem is minimized. Hardware interconnections become common across all servers and devices. For example, common trunk cables can be used for all servers, storage, and switches.

This section briefly describes the two main types of storage device pooling: *disk pooling* and *tape pooling*.

Disk pooling

Disk pooling allows multiple servers to utilize a common pool of SAN-attached disk storage devices. Disk storage resources are pooled within a disk subsystem or across multiple IBM and non-IBM disk subsystems, and capacity is assigned to independent file systems supported by the operating systems on servers. The servers are potentially a heterogeneous mix of UNIX, Windows, and even z/OS.

Storage can be dynamically added to the disk pool and assigned to any SAN-attached server when and where it is needed. This provides efficient access to shared disk resources without a level of indirection associated with a separate file server, since storage is effectively *directly attached* to all the servers, and efficiencies of scalability result from consolidation of storage capacity.

When storage is added, zoning can be used to restrict access to the added capacity. As many devices (or LUNs) can be attached to a single port, access can be further restricted using LUN-masking, that is, specifying who can access a specific device or LUN.

Attaching and detaching storage devices can be done under the control of a common administrative interface. Storage capacity can be added without stopping the server, and can be immediately made available to the applications.

Figure 10-3 shows an example of disk storage pooling across two servers.

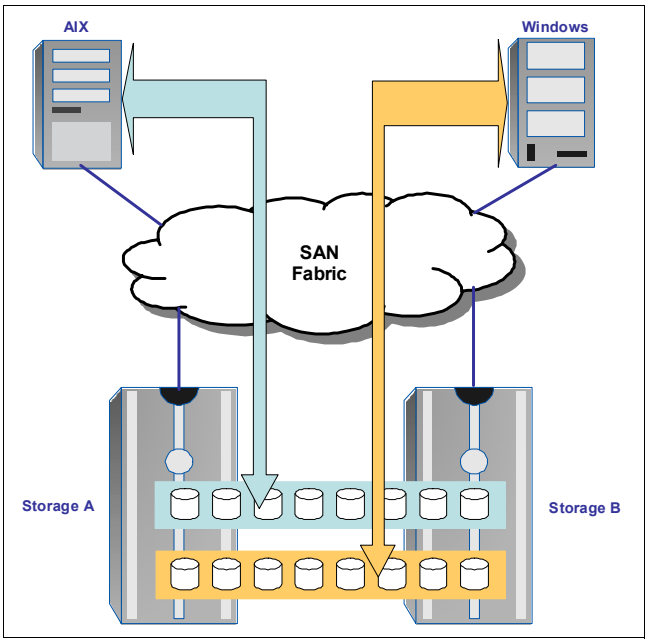

Figure 10-3 Disk pooling concept

One server is assigned a pool of disks formatted to the requirements of the file system, and the second server is assigned another pool of disks, possibly in another format. The third pool may be the space not yet allocated or can be a pre-formatted disk for future use. Again, all the changes in the disk structure can be done dynamically, without any interruption to the service.

Tape pooling

Tape pooling addresses the problem faced today in an open systems environment in which multiple servers are unable to share tape resources across multiple hosts. Older methods of sharing a device between hosts consist of either manually switching the tape device from one

host to the other, or writing applications that communicate with connected servers through distributed programming.

Tape pooling allows applications on one or more servers to share tape drives, libraries, and cartridges in a SAN environment in an automated, secure manner. With a SAN infrastructure, each host can directly address the tape device as though it were connected to all of the hosts.

Tape drives, libraries, and cartridges are owned by either a central manager (tape library manager) or a peer-to-peer management implementation, and are dynamically allocated and reallocated to systems (tape library clients) as required, based on demand. Tape pooling allows for resource sharing, automation, improved tape management, and added security for tape media.

Software is required to manage the assignment and locking of the tape devices in order to serialize tape access. Tape pooling is a very efficient and cost effective way of sharing expensive tape resources, such as automated tape libraries. At any particular instant in time, a tape drive can be owned by one system only.

This concept of tape resource sharing and pooling is proven in medium to enterprise backup and archive solutions using for example IBM Tivoli Storage Manager with SAN attached IBM tape libraries.

Logical volume partitioning

At first sight one can ask: “How will logical volume partitioning make my infrastructure simpler? It looks to me like we are creating more and more pieces to manage in my storage!”. Conceptually this is correct but the benefit of logical volume partitioning is to address the need for maximum volume capacity and to effectively utilize it within target systems. It is essentially a way of dividing the capacity of a single storage server into multiple pieces. The storage subsystems are connected to multiple servers, and storage capacity is partitioned among the various subsystems.

Logical disk volumes are defined within the storage subsystem and assigned to servers. The logical disk is addressable from the server. A logical disk may be a subset or superset of disks only addressable by the subsystem itself. A logical disk volume can also be defined as subsets of several physical disks (striping). The capacity of a disk volume is set when defined. For example, two logical disks, with different capacities (for example, 50 GB and 150 GB) may be created from a single 300 GB hardware addressable disk, with each being assigned to a different server, leaving 100 GB of unassigned capacity. A single 2000 GB logical disk may also be created from multiple real disks that exist in different storage subsystems. The underlying storage controller must have the necessary logic to manage the volume grouping, and guarantee access securely to the data.

The function of storage controller can be further exploited by some of the storage virtualization engines, such as the IBM SAN Volume Controller (SVC), that offers even greater and more scalability and virtualization of storage resources with less management effort and with clearer visibility to the target host systems.

Figure 10-4 shows multiple servers accessing logical volumes created using the different alternatives mentioned above. (The logical volume *Unallocated volume* is not assigned to any server.).

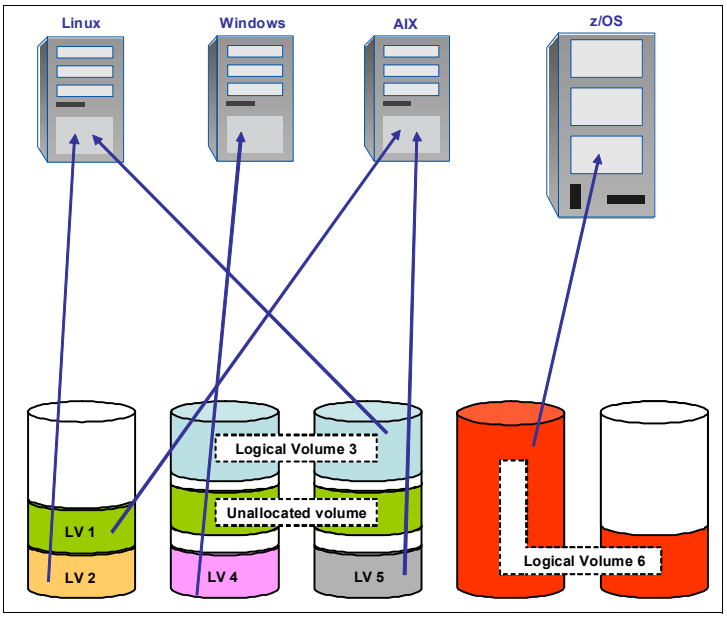

Figure 10-4 Conceptual model of logical volume partitioning.

10.3.3 Consolidation

We can improve scalability, security, and manageability by enabling devices in separate SAN fabrics to communicate without merging fabrics into a single, large SAN fabric. This capability enables clients to initially deploy separate SAN solutions at the departmental and data center levels and then to consolidate them into large enterprise SAN solutions as their experience and requirements grow and change. This kind of solution is also known as Data Center Bridging.

Clients have deployed multiple SAN islands for different applications with different fabric switch solutions. Growing availability of iSCSI server capabilities has created the opportunity for low-cost iSCSI server integration and storage consolidation. Additionally, depending on the choice of router, they will provide FCIP or iFCP capability.

The available multiprotocol SAN routers provide an iSCSI Gateway Service to integrate low-cost Ethernet-connected servers to existing SAN infrastructures. It also provides Fibre Channel, FC-FC Routing Service to interconnect multiple SAN islands without requiring the fabrics to merge into a single large SAN.

In Figure 10-5 we show an example of using a multiprotocol router and converged core switch to extend SAN capabilities across the long distances or just over the metropolitan areas.

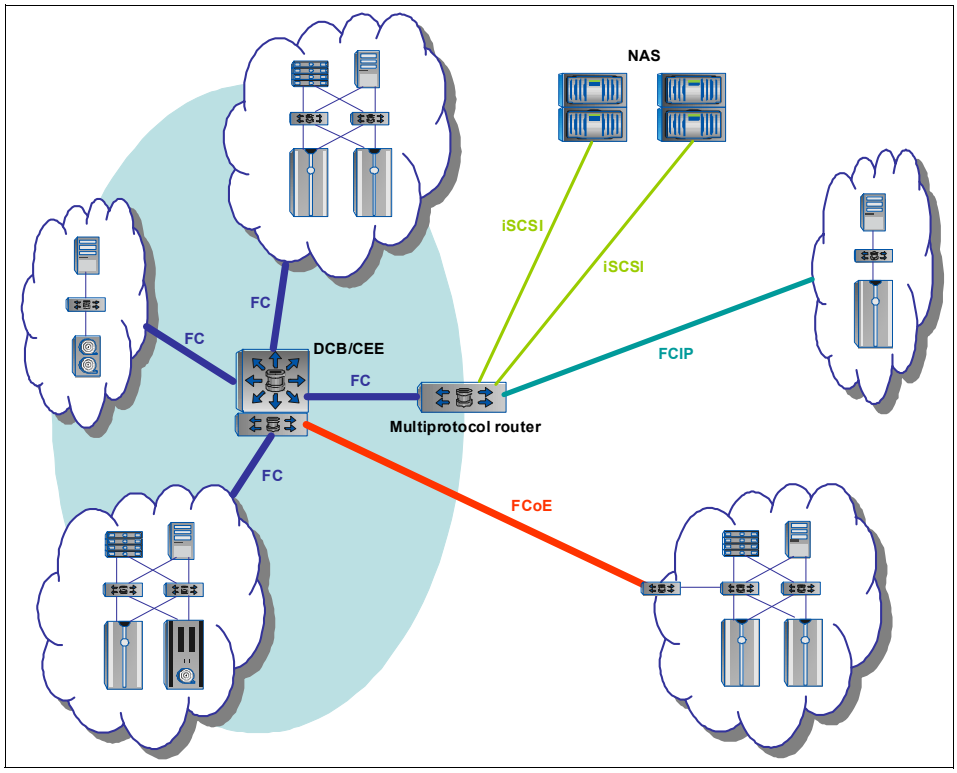

Figure 10-5 The concept of SAN consolidation

A multiprotocol capable router solution brings a number of benefits to the marketplace. In our example, there are discrete SAN islands, and number of different protocols involved. To merge these SAN fabrics, it would involve a number of disruptive and potentially expensive actions such as:

- Downtime
- Purchase of additional switches/ports
- Purchase of HBAs
- Migration costs
- Configuration costs
- Purchase of additional licenses
- Ongoing maintenance

However, by installing a multiprotocol router or core FCoE enabled switch or director, the advantages are:

- Least disruptive method
- No need to purchase extra HBAs
- Minimum number of ports to connect to the router
- No expensive downtime
- No expensive migration costs
- No ongoing maintenance costs other than router
- Support of other protocols
- Increases Return of Investment (ROI) by consolidating resources
- Can be used to isolate the SAN environment to be more secure

There are more benefits that the router and core switch can provide. In this example, an FC-FC routing service that negates the need for a costly SAN fabric merge exercise, the advantages are apparent and real. The router can also be used to provide:

- Device connectivity across multiple SANs for infrastructure simplification
- Tape backup consolidation for information lifecycle management
- Long distance SAN extension for business continuity
- Low-cost server connectivity to SAN resources

10.3.4 Migration to a converged network

Medium and enterprise datacenters usually run multiple separate networks, including an Ethernet network for client to server and server to server communications and a Fibre Channel SAN for the same type of connections. To support various types of networks, datacenters use separate redundant interface modules for each network — Ethernet network interface cards (NICs) and Fibre Channel interfaces (HBAs) in their servers, and redundant pairs of switches at each layer in the network architecture. Use of parallel infrastructures increases capital costs, makes data center management more difficult, and diminishes business flexibility.

The principle of consolidation of both independent networks to share a single, integrated networking infrastructure relies on utilization of Fibre Channel over Ethernet (FCoE) and helps addresses these challenges efficiently and effectively. In the following topics we briefly describe how to migrate your current infrastructure to a converged network in three principal steps. The prerequisite of the converged network is lossless 10 Gbps (GoE) Ethernet, inline with the Data Center Bridging standards (DCB). Figure 10-6 on page 214 presents the concept of the migration to convergency.

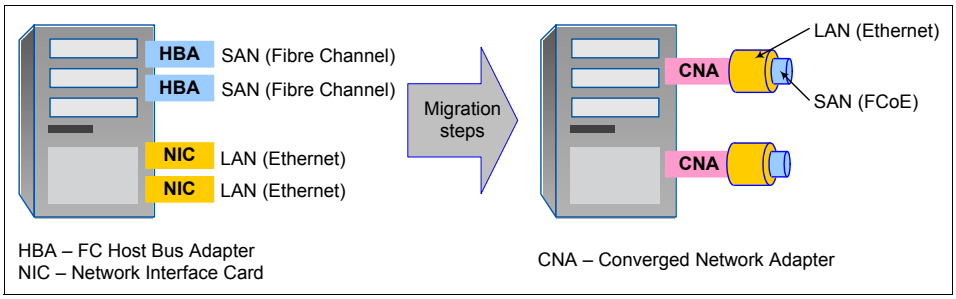

Figure 10-6 Conceptual model of migration to Converged network

The key benefits of migration to converged networks we can summarize as:

- Reduced capital expenditures by 15-30% depending on current infrastructure
- Reduced operational and management cost by 30-50%
- Improved network and storage efficiencies
- Increased assets and storage utilization
- Improved flexibility and business agility
- Reduced power consumption by 15-20%

Step 1 - Access layer convergence

Let's say we have separate adapters for 1 or 10 Gbps ethernet communication (from 2 to 8 per server) and FC HBAs for storage networking (from 2 to 6 adapters, usually dual-port). In this step we are going to replace these combinations by using Converged Network Adapters (CNA) - see the Figure 10-7 on page 215.

Note: For illustration purposes, in all figures we present only a single fabric datacenter solution, in real datacenters dual-fabric deployment is essential.

Additionally we need to install switch that supports both protocols, IP and FCoE, usually as a top-of-rack (TOR) device. So, the TOR, that supports Data Center Bridging (DCB) standards and multiple protocols can continue to work with the existing environment and separate the network traffic from the data storage traffic and direct each of them to the correct part of the datacenter networking infrastructure. All the segmentation of the traffic is done at the access layer and this is the first step of the overall process.

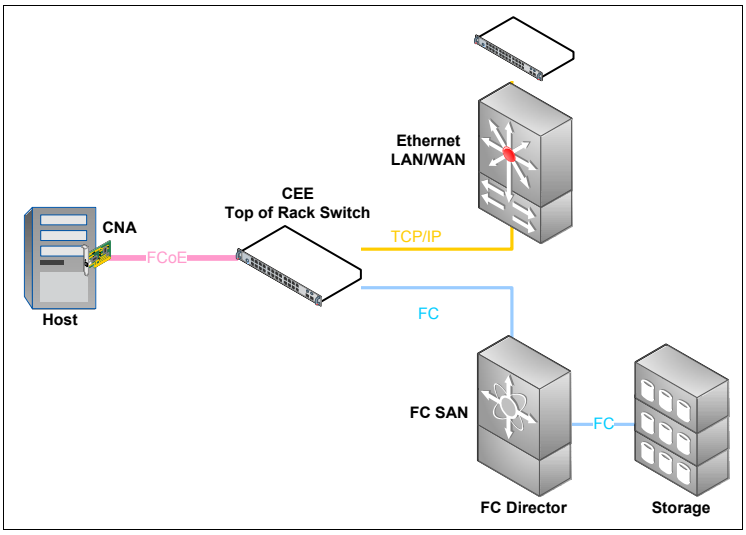

Figure 10-7 Access layer convergence

Step 2 - Fabric convergence

The second step is to implement more core type of switches that support datacenter bridging and converged network protocols. So, rather than implementing converged network on TOR switches or blades, we will move this function to the core directors. There is a second stage of the development of DCB standards that introduces Multi-Hop bridging, as there are different solutions from each of the SAN networking products' vendors. See Figure 10-8 for details.

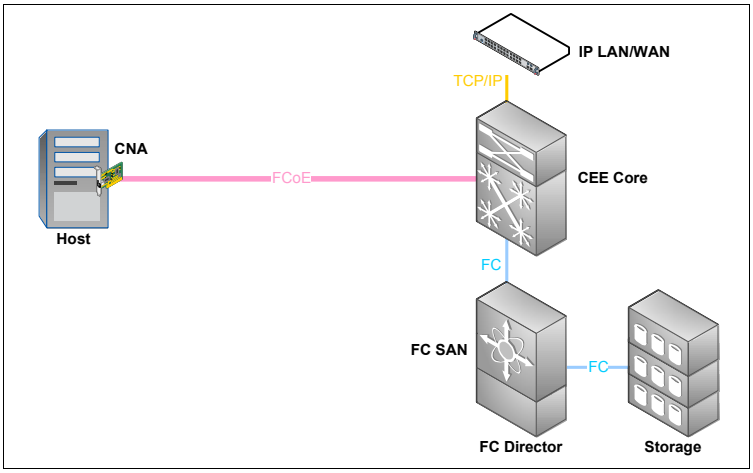

Figure 10-8 Fabric convergence

Step 3 - Storage convergence

For the final step of the migration we implement native FCoE enabled storage devices. At present, there are various vendors with midrange to enterprise disk storage systems that already offer Fibre Channel over Ethernet. This step will enable clients to migrate the current FC-attached storage data to the FCoE enabled storage system and disconnect the original

FC core and edge switches. This dramatically reduces the requirements for operation and management of the infrastructure, reduces the power consumption, and simplifies the complexity of the network (rack space, cabling). Figure 10-9 shows the final status of the converged network infrastructure.

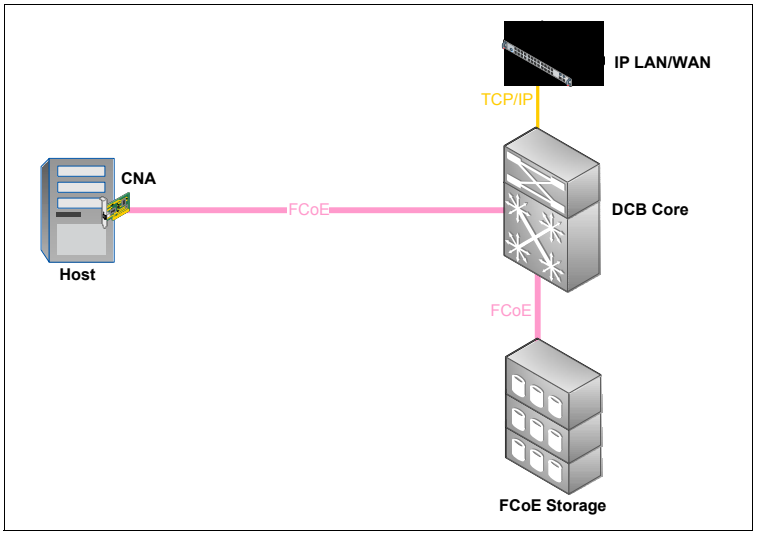

Figure 10-9 Storage convergence

FCoE in the converged network offers several benefits and advantages over existing approaches to I/O consolidation:

- Compatibility with existing Fibre Channel SANs by preserving well-known FC concepts including virtual SANs (VSANs), World Wide Names (WWNs), FC IDs (FCIDs), multipathing, and zoning to servers and storage arrays
- A high level of performance, comparable to the performance of current Ethernet and FC networks, achieved by using a hardware-based Ethernet network infrastructure that is not limited by the overhead of higher-layer TCP/IP protocols
- The exceptional scalability of Ethernet at the highest available speeds (1, 10, and 40 GoE and eventually 100 GoE)
- Simplified operations and management (no change to the management infrastructure currently deployed in SANs)

10.4 Business continuity and Disaster Recovery

On demand businesses rely on their IT systems to conduct business. Everything must be working all the time! Failure truly is not an option these days. A sound and comprehensive business continuity strategy that encompasses high availability, near continuous operations, fault-tolerant systems, and disaster recovery is essential.

Today, data protection of multiple network or SAN attached servers is performed according to one of two backup and recovery paradigms: local backup and recovery, or network backup and recovery.

The local backup and recovery solution has the advantage of speed, because the data does not travel over the network. However, with a local backup and recovery approach, there are costs for overhead (because local devices must be acquired for each server, and are thus

difficult to utilize efficiently), and management overhead (because of the need to support multiple tape drives, libraries, and mount operations).

The network backup and recovery approach using shared tape libraries and tape drives, the best over SAN, is cost-effective, because it allows centralization of storage devices using one or more network-attached devices. This centralization shortens the return on investment as the installed devices are utilized efficiently. One tape library can be shared across many servers. Management of a network backup and recovery environment is often simpler than the local backup and recovery environment, because it eliminates the potential need to perform manual tape mount operations on multiple servers.

SANs combine the best of both approaches. This is accomplished by central management of backup and recovery, assigning one or more tape devices to each server, and using FC protocols to transfer data directly from the disk device to the tape device, or vice versa, over the SAN.

Another hot-topic in this category is instant business continuity in case of a device failure. Clients running the most critical applications cannot afford to wait until their data is restored to fixed or standby server or devices from backups. Server or application clusters allow clients to continue their business with minimal outage or even without any disruption. We are talking about highly available or fault-tolerant systems.

In the following sections we discuss these approaches in more detail.

10.4.1 Clustering and high availability

SAN architecture naturally allows multiple systems (target hosts) to access the same disk resources in your medium or enterprise disk storage systems, even concurrently. This feature enables specific applications (AIX HACMP™, PowerHA®, Windows Cluster Services, etc.) running on the hosts to introduce highly available or fault tolerant application systems. They assure that in a case of a single host failure, the application is automatically, without any manual administrator intervention, moved over to the backup cluster host (High Availability - short outage) or failed host is isolated from application processing and workload is balanced between other working cluster nodes (Fault-tolerant - no disruption to service).

The conceptual scheme of a highly available cluster solution is shown in Figure 10-10 on page 218.

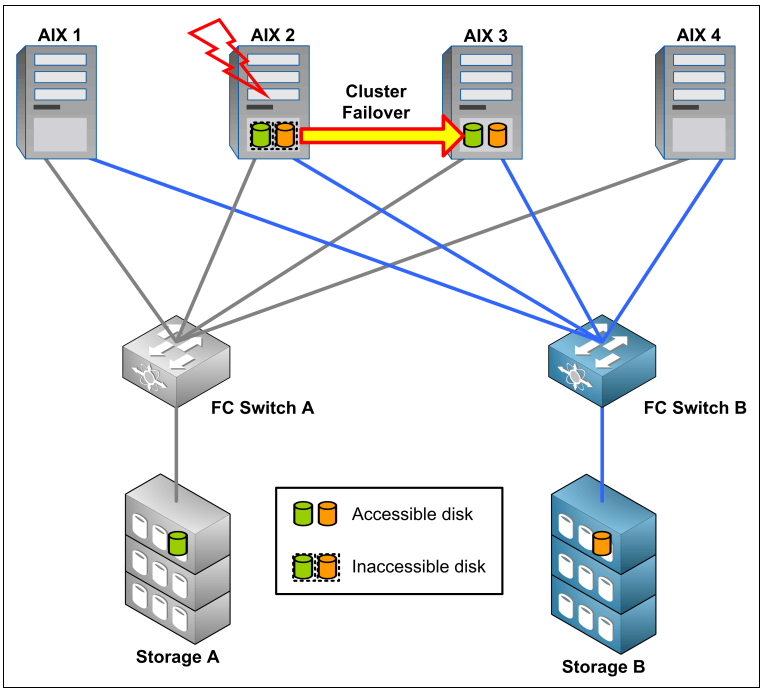

Figure 10-10 Highly available cluster system

In this example we have an application running on system AIX2 and it is managing mirrored disks from both storage systems (green and orange). SAN zoning allows both cluster nodes (AIX2 and AIX3) to operate the same set of disks and cluster has one primary cluster node active (an Active-Passive cluster). At the time of AIX2 failure, cluster services running on AIX3 recognize the failure and automatically move all the application resources to AIX3, activate the disk sets there and start the application in the correct sequence.

Figure 10-11 on page 219 explains the configuration of fault-tolerant clustered environment.

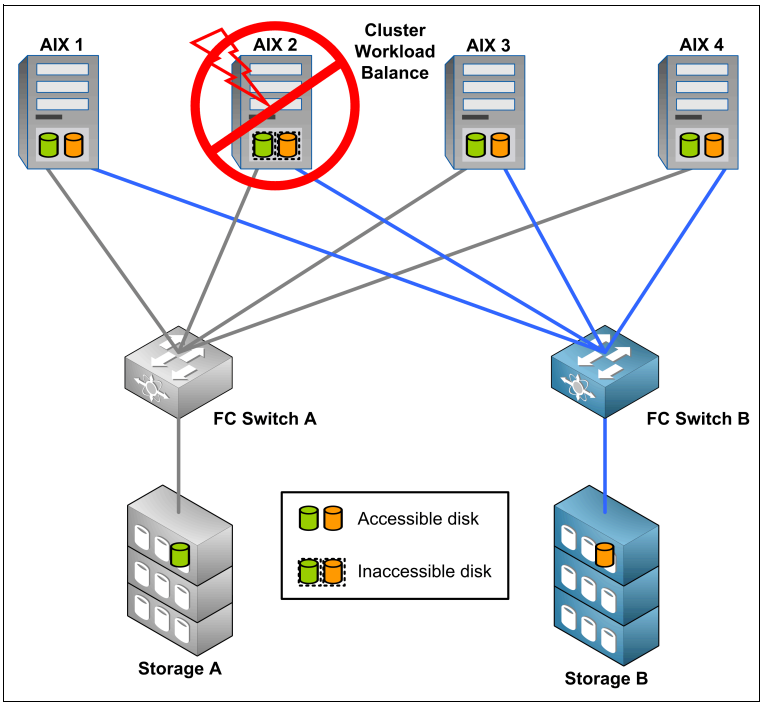

Figure 10-11 Fault-tolerant cluster system

SAN zoned disks are available to all four AIX host systems, the clusters are active and the master application works concurrently on them with workload balancing. In the case of an AIX2 system failure, the cluster application automatically inactivates assigned disks and redistributes the workload between the remaining active cluster nodes. There is no interruption to the business. A configuration such as this is costly and usually is only employed for business critical applications such as banking systems, air traffic control systems, and so on.

10.4.2 LAN-free data movement

The network backup and recovery paradigm implies that data flows from the backup and recovery client (usually a file or database server) to the centralized backup and recovery server, or between backup and recovery servers, over the ethernet network. The same applies for the archive or storage management applications. Often the network connection is the bottleneck for data throughput especially in case of large database systems. This is due to the network connection bandwidth limitations. The SAN offers the advantage to offload the backup data out of LAN.

Tape drive and tape library sharing

A basic requirement for LAN-free backup and recovery is the ability to share tape drives and tape libraries between central backup tape repository and backup and recovery clients with large database files. Systems with higher amount of small files will still use the network for data transportation as they cannot benefit from LAN-free.

In the tape drive and tape library sharing approach, the backup and recovery server or client that requests a backup copy to be copied to or from tape will read or write the data directly to the tape device using SCSI commands. This approach bypasses the network transport's latency and network protocol path length; therefore, it can offer improved backup and recovery speeds in cases where the network is the constraining factor. The data is read from

the source device and written directly to the destination device. Central backup and recovery server only controls the tape mount operations and store the references (metadata) into its embedded database system.

Figure 10-12 shows an example of tape library sharing and LAN-free backups.

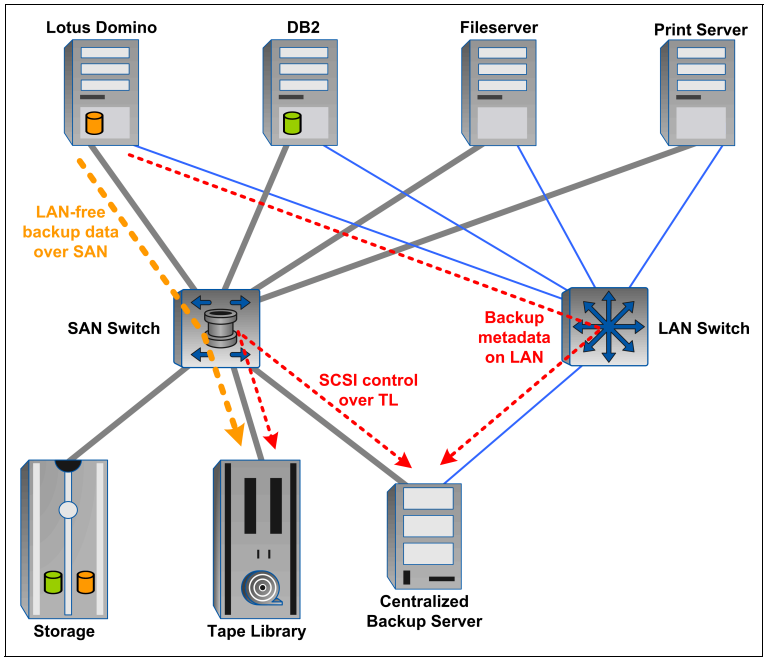

Figure 10-12 LAN-free and LAN-based backups

While IBM Lotus® Domino and IBM DB2 database systems benefit from greater performance of backups over Fibre Channel directly to tapes, small servers with higher amount of files still continue to perform backups on LAN or WAN.

IBM offers enterprise, centralized backup/recovery solution that supports various platforms and database systems. IBM Tivoli Storage Manager (ITSM) and its component ITSM for Storage Area Networks enables clients to perform online backups and archives of large application systems directly to tape over SAN, and without significant impact on performance.

10.4.3 Disaster backup and recovery

SAN can facilitate disaster backup solutions because of the greater flexibility allowed in connecting storage devices to servers, and also the greater distances that are supported when compared to SCSI restrictions. It is now possible to perform extended distance backups for disaster recovery within a campus or city, as shown in Figure 10-13.

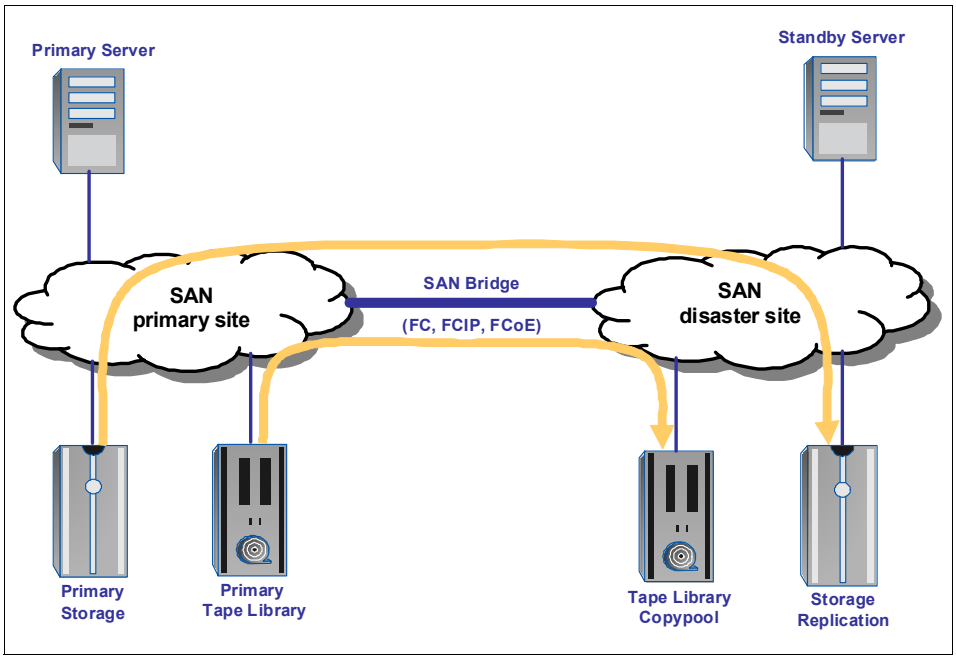

Figure 10-13 Disaster backup at remote site using SAN bridging

When longer distances are required, SANs must be connected using gateways and WANs, one of the solutions is Fibre Channel over Ethernet (FCoE).

Depending on business requirements, disaster protection deployments may make use of copy services implemented in disk subsystems and tape libraries (that might be achieved by utilization of SAN services), SAN copy services, and most likely a combination of both.

10.5 Information Lifecycle Management

Information Lifecycle Management (ILM) is a process for managing information through its lifecycle, from conception until disposal, in a manner that optimizes storage and access at the lowest cost.

ILM is not just hardware or software, it includes processes and policies to manage the information. It is designed upon the recognition that different types of information can have different values at different points in their lifecycle. Predicting storage needs and controlling costs can be especially challenging as the business grows.

The overall objectives of managing information with ILM are to help reduce the total cost of ownership (TCO) and help implement data retention and compliance policies. In order to effectively implement ILM, owners of the data need to determine how information is created, how it ages, how it is modified, and if/when it can safely be deleted. ILM segments data according to value, which can help create an economical balance and sustainable strategy to align storage costs with businesses objectives and information value.

10.5.1 ILM elements

To manage the data lifecycle and make your business ready for On Demand, there are four main elements that can address your business in an ILM structured environment. They are:

- Tiered storage management
- Long-term data retention and archiving
- Data lifecycle management
- Policy-based archive management

In the next sections we describe each of these elements in more detail.

10.5.2 Tiered storage management

Most organizations today seek a storage solution that can help them manage data more efficiently. They want to reduce the costs of storing large and growing amounts of data and files and maintain business continuity. Through tiered storage, you can reduce overall disk-storage costs, by providing benefits like:

- Reducing overall disk-storage costs by allocating the most recent and most critical business data to higher performance disk storage, while moving older and less critical business data to lower cost disk storage.
- Speeding business processes by providing high-performance access to most recent and most frequently accessed data.
- Reducing administrative tasks and human errors. Older data can be moved to lower cost disk storage automatically and transparently.

Typical storage environment

Storage environments typically have multiple tiers of *data value*, such as application data that is needed daily, and archive data that is accessed infrequently. However, typical storage configurations offer only a single tier of storage, as shown in Figure 10-14, which limits the ability to optimize cost and performance.

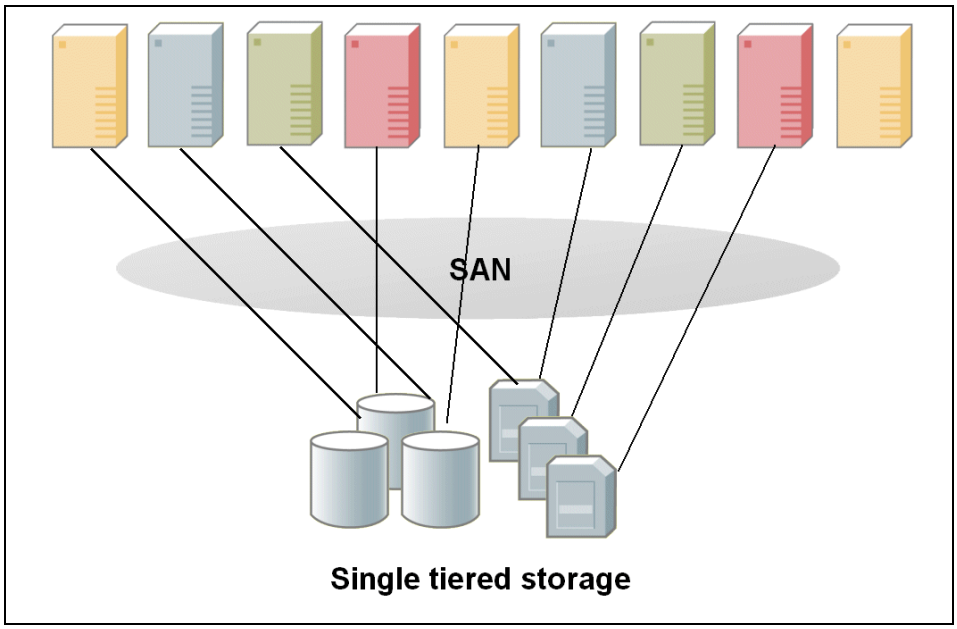

Figure 10-14 Traditional non-tiered storage environment

Multi-tiered storage environment

A tiered storage environment that utilizes the SAN infrastructure affords the flexibility to align storage cost with the changing value of information. The tiers will be related to data value.

The most critical data is allocated to higher performance disk storage, while less critical business data is allocated to lower cost disk storage.

Each storage tier will provide different performance metrics and disaster recovery capabilities. Creating classes and storage device groups is an important step to configure a tiered storage ILM environment.

Figure 10-15 shows a multi-tiered storage environment.

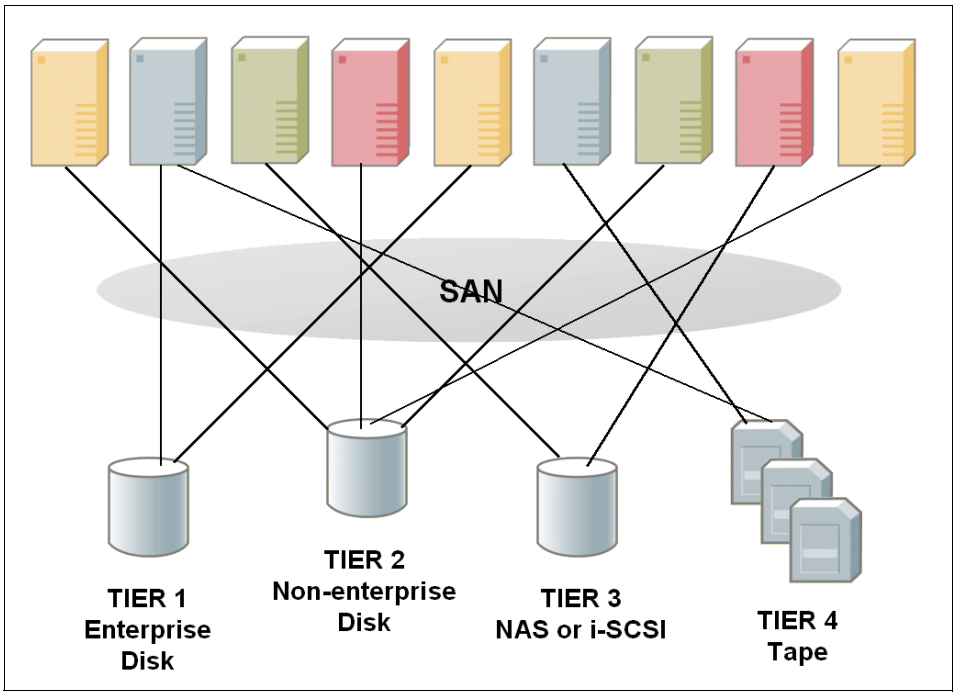

Figure 10-15 ILM tiered storage environment

An IBM ILM solution in a tiered storage environment is designed to:

- Reduce the total cost of ownership (TCO) of managing information. It can help optimize data costs and management, freeing expensive disk storage for the most valuable information.
- Segment data according to value. This can help create an economical balance and sustainable strategy to align storage costs with business objectives and information value.
- Help make decisions about moving, retaining, and deleting data, because ILM solutions are closely tied to applications.
- Manage information and determine how it should be managed based on content, rather than migrating data based on technical specifications. This approach can help result in more responsive management, and offers you the ability to retain or delete information in accordance with business rules.
- Provide the framework for a comprehensive enterprise content management strategy.

10.5.3 Long-term data retention

There is a rapidly growing class of data that is best described by the way in which it is managed rather than the arrangement of its bits. The most important attribute of this kind of data is its retention period, hence it is called *retention managed data*, and it is typically kept in an archive or a repository. In the past it has been variously known as archive data, fixed content data, reference data, unstructured data, and other terms implying its read-only

nature. It is often measured in terabytes and is kept for long periods of time, sometimes forever.

In addition to the sheer growth of data, laws and regulations governing the storage and secure retention of business and client information are increasingly becoming part of the business landscape, making data retention a major challenge to any institution. An example of these is the Sarbanes-Oxley Act in the US, of 2002.

Businesses must comply with these laws and regulations. Regulated information can include e-mail, instant messages, business transactions, accounting records, contracts, or insurance claims processing, all of which can have different retention periods, for example, for 2 years, for 7 years, or retained forever. Data is an asset when it needs to be kept; however, data kept past its mandated retention period could also become a liability. Furthermore, the retention period can change due to factors such as litigation. All these factors mandate tight coordination and the need for ILM.

Not only are there numerous state and governmental regulations that must be met for data storage, but there are also industry-specific and company-specific ones. And of course these regulations are constantly being updated and amended. Organizations need to develop a strategy to ensure that the correct information is kept for the correct period of time, and is readily accessible when it needs to be retrieved at the request of regulators or auditors.

It is easy to envisage the exponential growth in data storage that will result from these regulations and the accompanying requirement for a means of managing this data. Overall, the management and control of retention managed data is a significant challenge for the IT industry when taking into account factors such as cost, latency, bandwidth, integration, security, and privacy.

10.5.4 Data lifecycle management

At its core, the process of ILM moves data up and down a path of tiered storage resources, including high-performance, high-capacity disk arrays, lower-cost disk arrays such as serial ATA (SATA), tape libraries, and permanent archival media where appropriate. Yet ILM involves more than just data movement; it encompasses scheduled deletion and regulatory compliance as well. Because decisions about moving, retaining, and deleting data are closely tied to application use of data, ILM solutions are usually closely tied to applications.

ILM has the potential to provide the framework for a comprehensive information-management strategy, and helps ensure that information is stored on the most cost-effective media. This helps enable administrators to make use of tiered and virtual storage, as well as process automation. By migrating unused data off of more costly, high-performance disks, ILM is designed to help:

- Reduce costs to manage and retain data.
- Improve application performance.
- Reduce backup windows and ease system upgrades.
- Streamline data management.
- Allow the enterprise to respond to demand—in real-time.
- Support a sustainable storage management strategy.
- Scale as the business grows.

ILM is designed to recognize that different types of information can have different value at different points in their lifecycle. As shown in Figure 10-16 on page 225, data can be allocated to a specific storage level aligned to its cost, with policies defining when and where data will be moved.

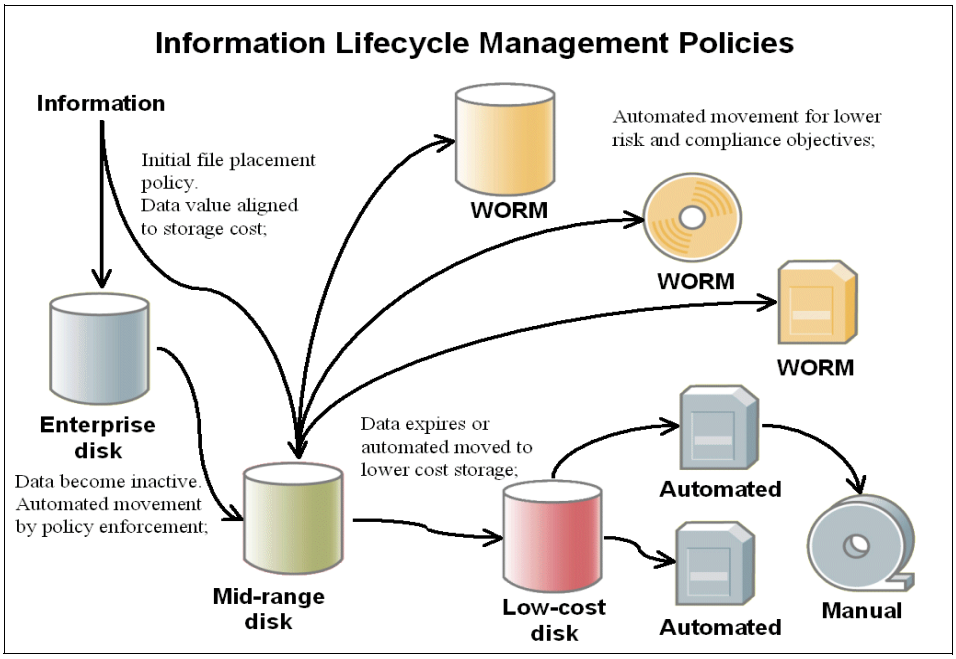

Figure 10-16 ILM policies

10.5.5 Policy-based archive management

As businesses of all sizes migrate to e-business solutions and a new way of doing business, they already have mountains of data and content that have been captured, stored, and distributed across the enterprise. This wealth of information provides a unique opportunity. By incorporating these assets into e-business solutions, and at the same time delivering newly generated information media to their employees and clients, a business can reduce costs and information redundancy and leverage the potential profit-making aspects of their information assets.

Growth of information in corporate databases such as Enterprise Resource Planning (ERP) systems and e-mail systems makes organizations think about moving unused data off the high-cost disks. They need to:

- Identify database data that is no longer being regularly accessed and move it to an archive where it remains available.
- Define and manage what to archive, when to archive, and how to archive from the mail system or database system to the back-end archive management system.

Database archive solutions can help improve performance for online databases, reduce backup times, and improve application upgrade times.

E-mail archiving solutions are designed to reduce the size of corporate e-mail systems by moving e-mail attachments and/or messages to an archive from which they can easily be recovered if needed. This action helps reduce the need for end-user management of e-mail, improves the performance of e-mail systems, and supports the retention and deletion of e-mail.

The way to do this is to migrate and store all information assets into an e-business enabled content manager. ERP databases and e-mail solutions generate large volumes of information and data objects that can be stored in content management archives. An archive solution allows you to free system resources, while maintaining access to the stored objects for later reference. Allowing it to manage and migrate data objects gives a solution the ability to have

ready access to newly created information that carries a higher value, while at the same time still being able to retrieve data that has been archived on less expensive media.

More information is available in *ILM Library: Information Lifecycle Management Best Practices Guide*, SG24-7251.

11

SAN and Green Datacenters

System storage networking products and their deployment in large enterprise datacenters significantly participate in the rapid growth of floorspace, power, and cooling resources.

In this chapter we briefly introduce the concepts of a green datacenter strategy and how SAN and IBM System Networking align with the green goal. In addition, we also describe the IBM Smarter Datacenter that facilitates the evolution of energy efficient operations.

11.1 Datacenter constraints

Many datacenters are running out of power and space. They cannot add more servers because they have reached either their power or space limits, or perhaps they have reached the limit of their cooling capacity.

In addition, environmental concerns are becoming priorities because they can also impede a company's ability to grow. Clients all over the world prefer to purchase products and services from companies that have a sustainable approach to the environment, and are able to meet any targets that may be imposed on them, whether from inside their company or from outside in the form of government legislation.

As environmental sustainability is a business imperative nowadays, many datacenter clients are looking at ways to save energy and cut costs so that the company can continue to grow. However, it is also a time to consider transformation in spending, not simply cutting costs. Only smarter investments in technology and perhaps a different way of thinking is needed to achieve green efficiency in datacenters.

Datacenters need to provide flexibility to respond quickly to future unknowns in business requirements, technology, and computing models. They need to adapt to be more cost-effective, for both capital expenses (CapEx) and operating expenses (OpEx). Additionally, they require active monitoring and management capabilities to provide the operational insights to meet the required availability, resiliency, capacity planning, and energy efficiency.

The IT architects have to consider four key factors that drive the efficiency of datacenter operations:

- **Energy cost** - the cost of a kilowatt of electricity has only risen slightly in recent years, but the cost of operating servers has increased significantly. The context around this paradox is that the energy consumption of the servers is increasing exponentially faster than the utility cost. Rising demand has accelerated the adoption of virtualization technologies and increased virtual image densities, driving total server energy consumption higher, while the amortized cost of operation per workload has been decreasing.
- **Power capacity** - some companies cannot deploy more servers because additional electrical power is not available. Many suppliers, especially those in crowded urban areas, are telling clients that power feeds are at capacity limits and that they simply have no more power to sell. New server, storage, and networking products give better performance at lower prices but can also be power hungry. The effort to overcome a power supply threshold is a huge investment.
- **Cooling capacity** - many datacenters are now 10 to 20 years old, and the cooling facilities are not adapted to the present needs. Traditional cooling methods allowed for 2-3 kW of cooling per rack. Today's requirements are 20-30 kW per rack. Heat density is many times past the design point of the data center.
- **Space limitation** - each time a new project or application comes online, new images, servers or storage subsystems are added. Therefore, the space utilization is growing exponentially because of business requirements. When images, servers and storage cannot be added, except by building another data center, next growth becomes significantly expensive.

11.1.1 Energy flow in datacenter

To understand how to reduce energy consumption, we need to understand where and how the energy is used. We can study energy use in a datacenter by taking three different views:

- How energy is distributed between IT equipment (servers, storage, network devices) and supporting facilities (power, cooling and lighting)
- How energy is distributed between the different components of the IT equipment (processor, memory, disk, and so forth)
- How the energy allocated to IT resources is utilized to produce business results (Are idle resources powered on, spending energy without any effect?)

Figure 11-1shows how energy is used by several components of a typical non-optimized data center. Each component is divided into two portions - IT equipment (servers, storage, network) and the infrastructure around, that supports the IT equipment (chillers, humidifiers, air condition. power distribution units, UPS, lights, and so on).

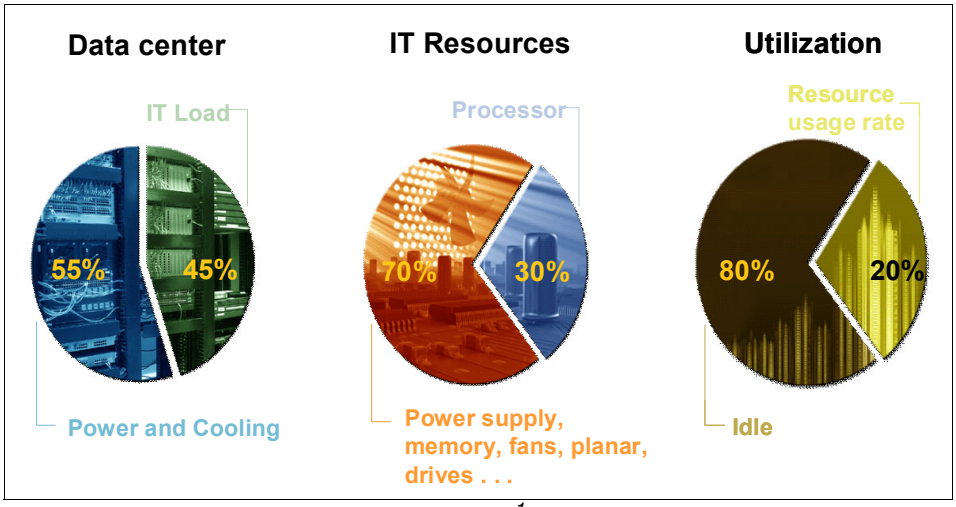

Figure 11-1 Energy usage in typical datacenter[1]

In typical datacenters, the IT equipment does not consume 55% of the overall energy that is brought into the datacenter, so this portion of the energy is not producing calculations, data storage, and so forth. The concept of green datacenter is able to eliminate this waste and reduce such inefficiency.

Remember: Basic laws of thermodynamics state that energy cannot be created or destroyed, it only changes form, and with efficiency of the conversion less than 100% (in a real world significantly less than 100%).

Solution designers and IT architects must also consider the energy consumption of the components at the IT equipment level. For example, in a typical server, the processor uses only 30% of the energy and the remainder of the system uses 70%. Therefore, efficient hardware design is crucial. Features like virtualization of physical servers can help to change this ratio to a more reasonable value.

Finally, companies should consider the use of IT resources in the data center. A typical server utilization rate is around 20%. Underutilized systems can be a big issue because a lot of energy is expended on non-business purposes, thus wasting a major investment. Again, server virtualization, consolidation, and addressed provisioning of IT resources help utilize the entire capacity of your IT equipment.

Datacenters must become immensely more efficient in order to meet the need while keeping costs in check as the demand for and price of resources continue to rise. But the realization of

[1] Data source: Creating Energy-Efficient Data Centers, U.S. Department of Energy

this efficiency requires a deep and pervasive transformation in how datacenters are designed, managed, operated, populated, and billed, mandating a unified and coordinated effort across organizational and functional boundaries toward a common set of goals.

In the following text we introduce the concept of green datacenters and how IBM supports the migration to the concept of the datacenters of next generation, that works effective, cost-efficient, and environment friendly.

11.2 Datacenter optimization

To enable your datacenter to become more effective, less power consuming, and cost-efficient in terms of infrastructure management and operation, the IT architects have to consider two components of migration strategy:

- **Optimization of the site and facilities** that include datacenter cooling, heating, ventilation, and air conditioning (HVAC), uninterruptible power supply (UPS), power distribution to the site and within the datacenter, standby power supply or alternative power sources
- **Optimization of the IT equipment** in the datacenter that generates business value to the clients, such as servers, (physical and virtual), disk and tape storage devices, networking products.

Applying innovative technologies within the datacenter can yield more computing power per kilowatt. The IT equipment continues to become more energy efficient. With technology evolution and innovation outpacing the life expectancy of datacenter equipment, many companies are finding that replacing older IT equipment with newer models can significantly reduce overall power and cooling requirements and free up valuable floor space. For example, IBM studies have demonstrated that blade servers reduce power and cooling requirements by 25-40% over 1U technologies. While replacing equipment before it is fully depreciated might seem unwise, the advantages that new models can offer (lower energy consumption and two to three times more computing power than older models) combined with potential space, power, and cooling recoveries, are usually enough to offset any lost asset value.

11.2.1 Strategic considerations

The strategy of moving towards having a green data center and the overall cost effective IT infrastructure consists of four major recommended steps:

- Centralization
  - Consolidate many small remote centers into fewer
  - Reduce infrastructure complexity
  - Improve facility management
  - Reduce staffing requirements
  - Improve management cost
- Physical consolidation
  - Consolidate many servers into fewer on physical resource boundaries
  - Reduce system management complexity
  - Reduce servers physical footprint in datacenters
- Virtualization
  - Remove physical resource boundaries
  - Increase hardware utilization

  - Allocate less than physical boundary
  - Reduce software license costs
- Application integration
  - Migrate many applications to fewer, more powerful server images
  - Simplify IT environment
  - Reduce operational resources
  - Improve application specific tuning and monitoring

11.3 Green storage

Computer systems are not the only candidates for energy savings; as the amount of managed data grows over the years exponentially, one of the top candidates are also storage system within datacenters. Each component of the storage system has power and cooling requirements.

Published studies show, that the proportion of energy used by storage disk systems and storage networking products varies between 15-25% of the overall energy consumption of the typical datacenter. This number significantly increases as they are continuously growing requirements for storage space.

However, no matter what efficiency improvements are made, active (spinning) disk drives still use energy as long as they are powered on. Consequently, the most energy-intensive strategy for data storage is to keep all the organization's data on active disks. While this provides the best access performance, it is not the most environmentally friendly approach, nor is it normally an absolute requirement.

Green storage technologies occupy less raw storage capacity to store the same amount of native valuable clients' data, therefore the energy consumption per gigabyte of raw capacity falls accordingly.

The storage strategy for green datacenters includes the following elements:

- Information Lifecycle Management (ILM)
- Consolidation and virtualization
- On demand storage provisioning
- Hierarchical storage and storage tiering
- Compression and deduplication

In the following sections we briefly discuss each of them.

11.3.1 Information Lifecycle Management

Information Lifecycle Management (ILM) is a process for managing information through its lifecycle, from conception until disposal, in a manner that optimizes storage and access at the lowest cost.

ILM is not just hardware or software, it includes processes and policies to manage the information. It is designed upon the recognition that different types of information can have different values at different points in their lifecycle. Predicting storage needs and controlling costs can be especially challenging as the business grows. Although the total value of stored information has increased overall, historically, not all data is created equal, and the value of that data to business operations fluctuates over time.

This trend is shown in Figure 11-2, and is commonly referred to as the *data lifecycle*. The existence of the data lifecycle means that all data cannot be treated the same.

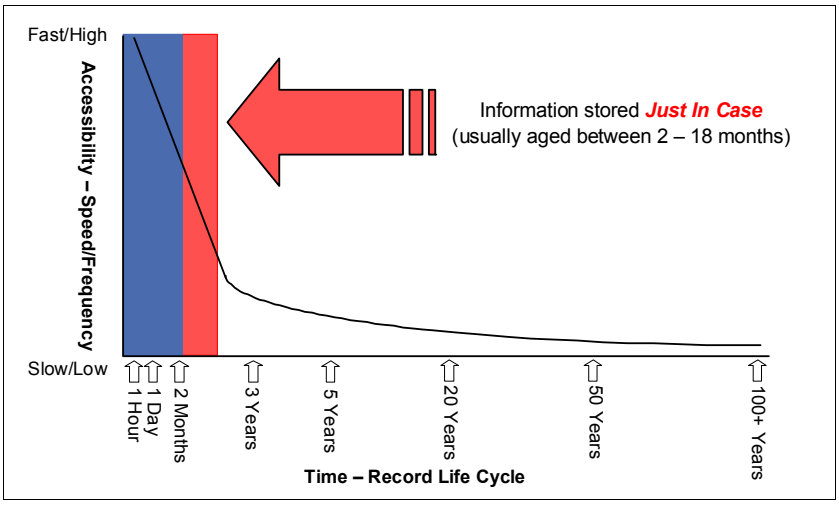

Figure 11-2 Data lifecycle

However, infrequently accessed or inactive data can become suddenly valuable again as events occur, or as new business initiatives or projects are taken on. Historically, the requirement to retain information has resulted in a "buy more storage" mentality. However, this approach has only served to increase overall operational costs and complexity, and has increased the demand for hard-to-find qualified personnel.

Typically, only around 20% of the information is active and frequently accessed by users. The remaining 80% are either inactive or even obsolete. See the Figure 11-3 for details.

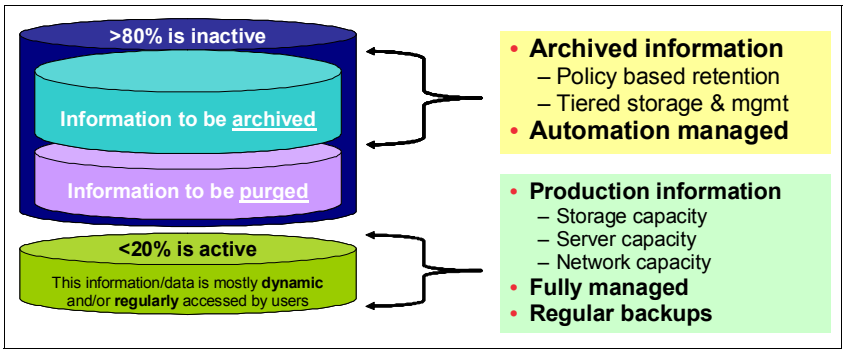

Figure 11-3 Usage of data

The automated identification of the storage resources in an infrastructure and analysis of how effectively those resources are being used is the crucial part of the ILM. File-system and file-level evaluation uncovers categories of files that, if deleted or archived, can potentially represent significant reductions in the amount of data that must be stored, backed up, and managed. The key position in the ILM process has the automated control through policies that are customizable with actions that can include centralized alerting, distributed responsibility and fully automated response. This includes also deletion of data.

See more details in *ILM Library: Information Lifecycle Management Best Practices Guide*, SG24-7251.

11.3.2 Storage consolidation and virtualization

As the need for data storage continues to spiral upward, traditional physical approaches to storage management become increasingly problematic. Physically expanding the storage environment can be costly, time-consuming, and disruptive; especially when it has to be done again and again in response to ever-growing storage demands. Yet manually improving storage utilization to control growth can be challenging. Physical infrastructures can also be inflexible at a time when businesses need to be able to make ever more rapid changes in order to stay competitive.

The alternative is a centralized, consolidated storage pool of disk devices (Figure 11-4), that is easy to manage, and transparent to be provisioned to the target host systems. Going further, the consolidated or centralized storage can be virtualized, where storage virtualization software presents a view of storage resources to servers that is different from the actual physical hardware in use.

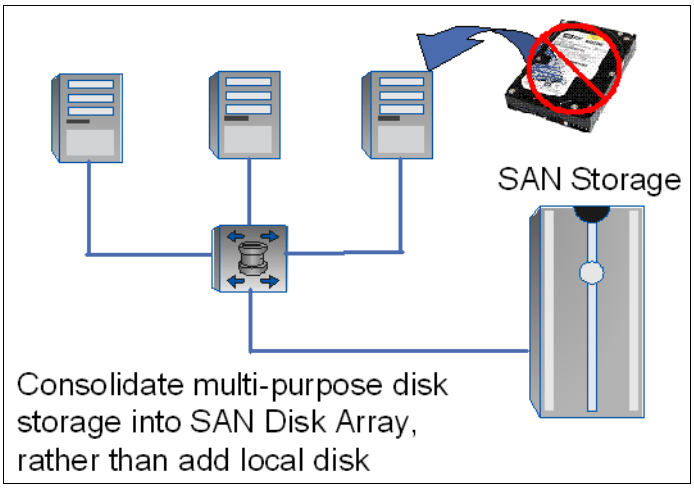

Figure 11-4 Storage consolidation

This logical view can hide undesirable characteristics of storage while presenting storage in a more convenient manner for applications. For example, storage virtualization may present storage capacity as a consolidated whole, hiding the actual physical boxes that contain the storage.

In this way, storage becomes a logical pool of resources that exists virtually, regardless of where the actual physical storage resources are located in the larger information infrastructure. These software-defined virtual resources are easier and less disruptive to change and manage than hardware-based physical storage devices, since they do not involve moving equipment or making physical connections. As a result, they can respond more flexibly and dynamically to changing business needs. Similarly, the flexibility afforded by virtual resources makes it easier to match storage to business requirements.

Virtualization offers significant business and IT advantages over traditional approaches to storage. Storage virtualization can help organizations to:

- Reduce data center complexity and improve IT productivity by managing multiple physical resources as fewer virtual resources.
- Flexibly meet rapidly changing demands by dynamically adjusting storage resources across the information infrastructure.
- Reduce capital and facility costs by creating virtual resources instead of adding more physical devices.

- Improve utilization of storage resources by sharing available capacity and deploying storage on demand only as it is needed.
- Deploy tiers of different storage types to help optimize storage capability while controlling cost and power and cooling requirements.

The value of a virtualized infrastructure is in the increased flexibility created by having pools of system resources on which to draw; in the improved access to information afforded by a shared infrastructure; and the lower total cost of ownership that comes with decreased management costs, increased asset utilization, and the ability to link infrastructure performance to specific business goals.

To learn more about how the IBM Storage Virtualization solutions can help your organization meet your storage challenges, study *IBM Information Infrastructure Solutions Handbook*, SG24-7814, or visit:

http://www.ibm.com/systems/storage/virtualization/

11.3.3 On demand storage provisioning

The provisioning of SAN attached storage capacity to a server can be a time consuming and cumbersome process. The task requires skilled storage administrators, and the complexity of the task can restrict an IT department's ability to respond quickly to requests to provision new capacity. However, there is a solution to this through automation. On demand storage provisioning solution monitors the current disk usage of specified target host systems and applications and allocates additional disk capacity for the period of business need.

End-to-end storage provisioning is the term applied to the whole set of steps required to provision usable storage capacity to a server. It covers the configuration of all the elements in the chain from carving out a new volume on a storage subsystem, through creating a file system at the host and making it available to the users or applications.

Typically, this involves a storage administrator using several different tools each focused on the specific task at hand, or the tasks are spread across several people. This spread of tasks and tools creates many inefficiencies in the provisioning process, which impacts the ability of IT departments to respond quickly to changing business demands. The resulting complexity and high degree of coordination required can also lead to errors and possible impact to the systems and application availability.

Automation of the end-to-end storage provisioning process through the use of workflow automation can significantly simplify this task of provisioning storage capacity. Each individual step is automated and best practice rules around zoning, device configuration and path selection can be applied automatically. The benefits are increased responsiveness to business requirements, lower administration costs and higher application availability.

11.3.4 Hierarchical storage and tiering

Companies continue to deploy storage systems that deliver many different classes of storage ranging from high performance/high cost to high capacity/low cost. Through the deployment of SANs, many of these storage assets are now physically connected to servers that run many different types of applications and create many kinds of information. Finally, with the arrival of network-resident storage services for distributed management of volume, files, and data replication, IT managers can intelligently provision, reallocate, and protect storage assets to meet the needs of many different applications across the network instead of device by device.

In a tiered storage environment, data is classified and assigned dynamically to different tiers. For example we could use expensive fast performing storage components to store often accessed, mission critical files in contrast to use cheaper storage for less used non-critical data. As a conclusion it improves efficiency and saves costs. We can identify the following typical storage tiers, categorized based on performance and cost per gigabyte:

- High-performing SAN attached disk system (SSD, SAS)
- Medium performing SAN attached disks (SAS, SATA)
- Network attached storage systems (NAS)
- Tape storage and other media with sequential access

Each level of storage tier can be assigned manually by storage administrator or data can be moved automatically between tiers, based on migration policies. The conceptual model of storage tiering is shown in Figure 11-5.

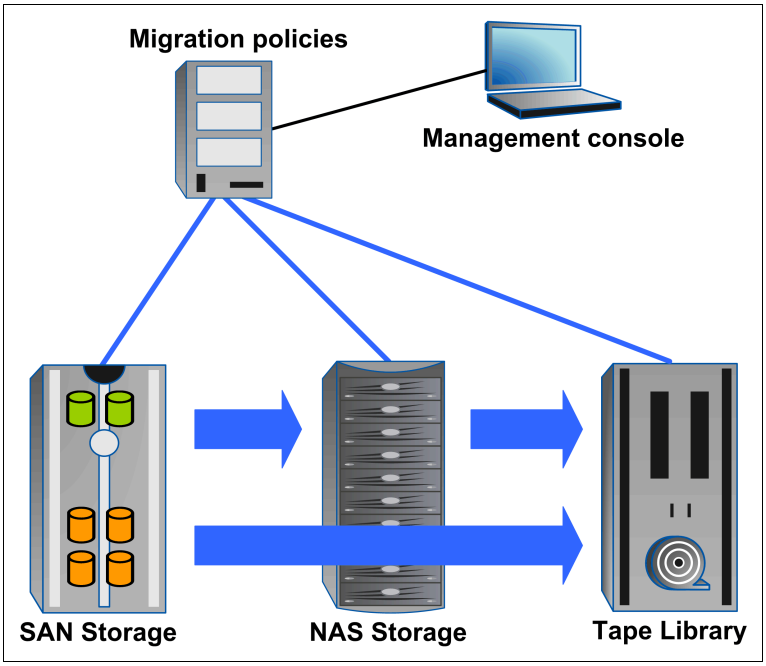

Figure 11-5 Principal of tiered storage

IBM offers variety of tools and utilities for storage tiering and hierarchical management for different scenarios. Starting with IBM Easy Tier in the enterprise disk storage systems, throughout the way to the IBM Tivoli Storage Manager (ITSM) for Hierarchical Storage Management for Windows and ITSM for Space Management for AIX/Linux platform. Specific position in tiered management has the Global Parallel File System (GPFS™), that allows data migration between different level of storage sources as well.

11.3.5 Data compression and deduplication

Business data growth rates will continue to increase rapidly in the coming years. Likewise, retention and retrieval requirements for new and existing data will expand, driving still more data to disk storage. As the amount of disk-based data continues to grow, there is an ever-increasing focus on improving data storage efficiencies across the information infrastructure.

Data reduction is a tactic which can decrease the disk storage and network bandwidth required, lower Total Cost of Ownership (TCO) for storage infrastructures, and optimize use of

existing storage assets and improve data recovery infrastructure efficiency. Compression and deduplication and other forms of data reduction are features that can exist within multiple components of the information infrastructure.

The *compression* immediately reduces the amount of required physical storage across all storage tiers: A solution that supports online compression of existing data allows storage administrators to gain back free disk space in the existing storage system without the need to change any administrative processes or enforcing users to clean up or archive data. The benefits to the business are immediate because the capital expense of upgrading the storage system is delayed. As data is stored in compressed format at the primary storage system, all other storage tiers and the transports in between them observe the same benefits. Replicas, backup images, and replication links all require less expenditure after implementing compression at the source.

After compression is applied to the stored data, the required power and cooling per unit of storage are reduced because more logical data is stored on the same amount of physical storage. In addition, within a particular storage system, more data can be stored, therefore the overall rack unit requirements are lowered. Figure on page 237 shows the typical compression rates that can be achieved with specific IBM products.

The exact compression ratio depends on the nature of the data. IBM has seen compression ratios as high as 90 percent in certain Oracle database configurations and in the neighborhood of 50 percent with *pdf* files. As always, compression ratios vary by data type and how the data is used.

In contrast with compression, the *data deduplication* mechanism identifies identical chunks of data within a storage container and keeps only one copy of each chunk, while all the other logically identical chunks will be pointed to this chunk. There are various implementations of this method. One option is in-line deduplication and the other one is post-processing. Each chunk of data must be identified in a way that is easily comparable. Chunks are processed using a parity calculation or cryptographic hash function. This processing gives the chunks shorter identifiers known as a hash values, digital signatures, or fingerprints. These fingerprints can be stored in an index or catalog where they can be compared quickly with other fingerprints to find matching chunks.

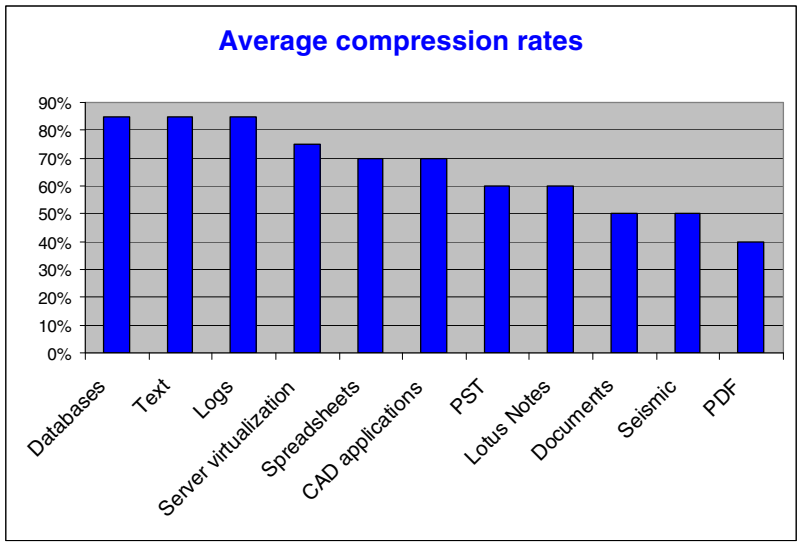

Figure 11-6 The average compression rates

Data deduplication processing can occur on the client, on an infrastructure server, or on the storage system. Each option has factors to consider:

- **Client-based deduplication** reduces the amount of data being transferred over the network to the storage system, but there is are often requirements for extra CPU and disk I/O processing on the client side.
- **Server-based deduplication** allows you to deduplicate multiple client's data at a scheduled time, but requires extra CPU and disk I/O processing on the infrastructure server (for example IBM Tivoli Storage Manager server).
- **Storage-based deduplication** occurs at the disk storage device level, where the data is stored. It is generally transparent to the clients and servers. It uses CPU and disk I/O on the storage system.

Additional information about data compression and deduplication, and concrete solutions from IBM can be found in *Introduction to IBM Real-time Compression Appliances*, SG24-7953 and *Implementing IBM Storage Data Deduplication Solutions*, SG24-7888.

12

The IBM product portfolio

This chapter guides you through the IBM System Storage SAN components that IBM offers using its marketing channels, either as OEM products or as an authorized reseller.

In addition to the typical Fibre Channel attached products, we also provide brief descriptions of storage and virtualization devices, even though they are usually not classified as SAN devices.

12.1 Classification of IBM SAN products

To stay competitive in the global marketplace, the right people need to have the right access to the right information at the right time in order to be effective, creative and highly innovative. IBM offers a comprehensive portfolio of SAN switches, storage, software, services, and solutions to reliably bring information to people in a cost effective way. IBM provides flexible, scalable and open standards-based business-class and global enterprise-class storage networking solutions for the on demand world.

IBM helps you to align your storage investment with the value of the information using a wide range of tiered storage options, policy-based automation and intelligent information management solutions. The IBM System Storage portfolio offers the industry's broadest range of storage solutions (including disk, tape, SAN, software, financial, and services offerings) enabling companies to create long-term solutions that can be tailored to their business needs. IBM System Storage tiered disk, tape and SAN switch solutions provide a wide variety of choices to align and move data to cost-optimized storage based on policies, matching the storage solution with the service level requirements and the value of the data in the growing environments.

IBM enables their clients to confidently protect strategic information assets and to efficiently comply with regulatory and security requirements with the unrivaled breadth of storage solutions from IBM. IBM SAN directors and routers provide metro and global connectivity between sites over Internet Protocol, IP networks. IBM SAN extension solutions include data compression and encryption services to help protect your data in flight between secure data centers.

IBM solutions are optimized for the unique needs of midsize organizations, large enterprises, cloud computing providers, and others. Clients can get just what they need, saving time and money. A key benefit of selecting IBM for your next information infrastructure project is access to a broad portfolio of outstanding products and services. IBM offers highly rated, patented technology that delivers unique value.

In this chapter we do not deep-dive into great technical details of each product as the intention of this chapter and book is to introduce the principles and basic components of the SAN environments in a reasonable extent, in a way easy to understand and follow.

We also do not cover the SAN networking products for IBM BladeCenter® technologies, as those are covered in *IBM BladeCenter Products and Technology*, SG24-7523.

We can identify the following categories of SAN products that IBM offers:

- SAN Fibre Channel networking
  - Entry level SAN switches
  - Midrange SAN switches
  - Enterprise SAN directors
  - Multiprotocol routers
- Storage device subsystems
  - Entry level disk systems
  - Midrange disk systems
  - Enterprise disk systems
- Tape storage systems
  - Fibre Channel tape drives
  - Autoloaders and entry level tape libraries
  - Midrange tape libraries

  - Enterprise tape libraries
- Storage virtualization
  - Disk storage virtualization
  - Tape virtualization
  - SAN products for Cloud computing
- IP-based datacenter networking for SAN environments
  - Hardware products for converged networks
  - Software solutions for virtual fabrics

For more technical details and additional information about other IBM storage products, refer to the *IBM System Storage Solutions Handbook*, SG24-5250; and for a comprehensive description of each product and its market position, visit the IBM storage web page:

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

12.2 SAN Fibre Channel networking

This section provides the brief information about IBM products for Fibre Channel based (optical) datacenter networking solutions, starting from entry level SAN switches, to midrange switches, and up to enterprise SAN directors and multiprotocol routers.

For the latest IBM SAN products refer to:

http://www-03.ibm.com/systems/networking/switches/san/index.html

12.2.1 Entry SAN switches

Entry level SAN switches represent easy-to-use preconfigured datacenter networking solutions for small and medium business (SMB) environments. IBM offers two products that fall into this category:

- IBM System Storage SAN24B-4 Express
- Cisco MDS 9124 Express for IBM System Storage

Here we summarize both of them.

IBM System Storage SAN24B-4 Express

Provides high-performance, scalable and simple-to-use fabric switching with 8, 16 or 24 ports operating at 8, 4, 2, or 1 Gigabits per second (depending on which optical transceiver is used) for servers running Microsoft Windows, IBM AIX, UNIX, and Linux operating systems, server clustering, infrastructure simplification and business continuity solutions. The SAN24B-4 Express includes EZSwitchSetup wizard, which is an embedded setup tool designed to guide novice users through switch setup, often in less than five minutes. Figure 12-1 shows the front view of the SAN24B-4 switch.

Figure 12-1 Front view of IBM System Storage SAN24B-4 Express switch

A single SAN24B-4 Express switch can serve as the cornerstone of a storage area network (SAN) for those that want to obtain the benefits of storage consolidation and are just beginning to implement FC storage systems. Such an entry level configuration can consist of one or two FC links to a disk storage array or to a Linear Tape Open (LTO) tape drive. An entry level eight port storage consolidation solution can support up to seven servers with a single path to either disk or tape. The Ports on Demand feature is designed to enable a base switch to grow to 16 and 24 ports to support more servers and more storage devices without taking the switch offline. A high-availability solution can be created with redundant switches. This capability is ideal for server clustering environments.

Such a configuration can support from 6 to 22 servers, each with dual FC adapters cross-connected to redundant SAN24B-4 Express switches, which are cross-connected to a dual controller storage system. While the focus of the SAN24B-4 Express is as the foundation of small to medium sized SANs, it can be configured to participate as a full member in an extended fabric configuration with other members of the IBM System Storage and former TotalStorage SAN b-type and m-type families. This capability helps provide investment

protection as SAN requirements evolve and grow over time. IBM System Storage SAN24B-4 Express switch provides the following features and characteristics:

- Efficient 1U design with 8, 16 or 24 ports configuration on demand
- Auto-sensing 8, 4, 2, 1 Gbps ports enabling high performance and improved utilization while proving easy installation and management
- Hot swappable SFPs
- Inter-switch link (ISL) trunking to up to 8 ports provides total bandwidth of 128 Gbps
- Provides Fibre Channel interfaces E_port, F_port, FL_port, and M_port
- Advanced zoning (hardware-enforced) helps protect against non-secure, unauthorized and unauthenticated network and management access and World Wide Name spoofing
- Hot firmware activation enables fast firmware upgrades that eliminates disruption to the existing fabric
- Backward compatibility with IBM b-type and m-type
- Optional as-needed licensed features include:
  - Adaptive Networking
  - Advance Performance Monitor
  - Extended Fabric
  - Fabric Watch
  - Trunking activation
  - Server Application Optimization (SAO)

Cisco MDS 9124 Express for IBM System Storage

Provides high-performance, scalable and simple-to-use fabric switching with 8, 16 or 24 ports operating at 1, 2 and 4 Gbps for servers running Microsoft Windows, UNIX, Linux, Novell NetWare and IBM OS/400® operating systems, server clustering, infrastructure simplification and business continuity solutions. The switch includes replaceable power supply, Virtual SAN, Cisco Fabric Manager and redundant power supply feature designed to simplify setup and ongoing maintenance for Cisco MDS 9000 users. The front view of the multilayer switch is shown in Figure 12-2 on page 243.

Figure 12-2 Front view of Cisco MDS 9124 Multilayer fabric switch

A single Cisco MDS 9124 switch can serve as an initial building block for a Storage Area Network for those who want to obtain the benefits of storage consolidation and are just beginning to implement Fibre Channel storage systems. An entry level configuration, for example, could consist of one or two Fibre Channel links to a disk storage array, or to an LTO tape drive. An entry level, eight port storage consolidation solution could support up to seven servers with a single path to either disk or tape. The On Demand Port Activation feature is designed to enable a base switch to grow from 8 to 24 ports, in 8 port increments, to support more servers and more storage devices without taking the switch offline.

Higher availability solutions can be created using multiple Cisco MDS 9124 switches. Such implementations would be well-suited to server clustering environments. Such a configuration could support from six to 22 servers, each with dual Fibre Channel adapters cross-connected

to redundant 9124 switches, which are cross-connected to a dual controller storage system. The main features and available options include:

- Efficient 1U design with 8, 16 or 24 ports configuration on demand
- Auto-sensing 4, 2, 1 Gbps ports, available with Shortwave, 4km Longwave, or 10km Longwave SFPs, all hot swapable
- Provides Fibre Channel interfaces E_port, F_port, and FL_port
- Simple SAN configuration using Cisco Fabric manager
- Replacable power supply and redundant cooling
- Optional features include:
  - Hot-swap redundant power supply
  - Cisco MDS 9000 Enterprise package activation
  - Cisco MDS 9000 Fabric Manager Server package activation

Discontinued entry SAN switches

These products and devices IBM has withdrawn from the marketing, however thay are still commonly seen in many datacenters of small or medium business solution, relying on SAN networking.

- IBM System Storage SAN10Q-2
- IBM TotalStorage SAN16B-2
- IBM TotalStorage SAN16M-2

For more information about currently available entry IBM SAN switches, visit:

http://www.ibm.com/systems/storage/san/entry/index.html

12.2.2 Midrange SAN switches

The IBM Midrange SAN switches provide scalable and affordable small and medium business (SMB) and enterprise solutions for storage networking:

- Cost conscious SMB customers with limited technical skills
- Integrated, scalable, high-availability IBM virtualization family solutions
- Heterogeneous Windows, Linux, iSeries®, UNIX, and Mainframe servers
- xSeries®, iSeries, pSeries®, and zSeries Server sales channels
- Support the IBM System Storage Virtualization family of products, System Storage and former TotalStorage devices and disk subsystems, IBM Tivoli Storage Manager, SAN Manager, SRM and Multiple Device Manager
- Integrated solutions at affordable prices with worldwide IBM support and IBM TotalStorage Solution Center (TSSC) services

The category of midrange SAN switches includes the following products:

- IBM System Storage SAN40B-4
- IBM System Storage SAN80B-4
- IBM System Storage SAN48B-5
- Cisco MDS 9148 for IBM System Storage
- IBM System Storage SAN32B-E4

IBM System Storage SAN40B-4

A compact, high-performance, easy-to-install Fibre Channel (FC) SAN switch which enables multiple servers to connect to external disk and tape systems. The SAN40B-4 supports most common operating systems and connects to most common servers and external storage devices. It supports server virtualization with virtual data paths across a single FC link. Can connect to other SAN switches, routers and directors to form a multi-switch fabric for increased connectivity and scalability of clients' SAN infrastructure. The front view of the SAN40B-4 switch (model 2498-B40) is shown in Figure 12-3 on page 245.

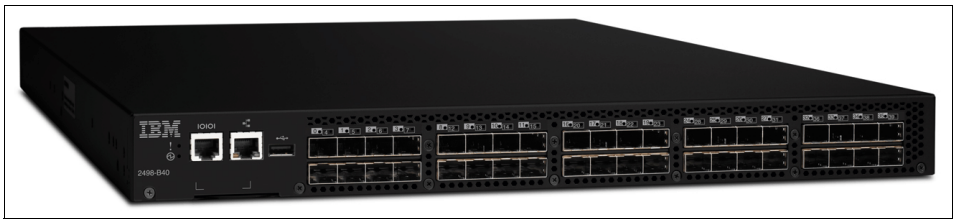

Figure 12-3 IBM System Storage SAN40B-4 switch

The IBM System Storage SAN40B-4 SAN fabric switch provides 24, 32, or 40 active ports and is designed for high performance with 8 Gbps link speeds and backward compatibility to support links running at 4, 2, and 1 Gbps link speeds. High availability features make it suitable for use as a core switch in midrange environments or as an edge-switch in enterprise environments where a wide range of storage area network (SAN) infrastructure simplification and business continuity configurations are possible. IBM Power Systems, System x, System z® and many non-IBM disk and tape devices are supported in many common operating system environments. Optional features provide specialized distance extension, dynamic routing between separate or heterogeneous fabrics, link trunking, IBM FICON, Server Application Optimization (SAO), performance monitoring and advanced security capabilities.

IBM System Storage SAN40B-4 switch provides the following features and characteristics:

- Efficient 1U design with 24, 32, and 40 ports configuration on demand
- Auto-sensing 8, 4, 2, 1 Gbps ports enabling high performance and improved utilization while proving easy installation and management
- Hot swappable SFPs and redundant power supply and cooling
- Inter-switch link (ISL) trunking to up to 8 ports provides total bandwidth of 128 Gbps
- Provides Fibre Channel interfaces E_port, F_port, FL_port, M_port, and optional EX_port
- N_port ID virtualization enables host images behind identical HBA to connect to an F_port using unique N_port ID
- Advanced zoning (hardware-enforced) helps protect against non-secure, unauthorized and unauthenticated network and management access and World Wide Name spoofing
- Hot firmware activation enables fast firmware upgrades that eliminates disruption to the existing fabric
- Backward compatibility with IBM b-type and m-type

IBM System Storage SAN80B-4

A compact, high-performance, easy-to-install Fibre Channel (FC) SAN switch that enables multiple servers to connect to external disk and tape systems. It extends the basic model SAN40B-4 by additional 40 8 Gbps ports. The SAN80B-4 (model 2498-B80) supports most common operating systems and connects to most common servers and external storage devices. Supports server virtualization with virtual data paths across a single FC link. Can

connect to other SAN switches, routers and directors to form a multi-switch fabric for increased connectivity. See Figure 12-4 for the front view of SAN80B-4.

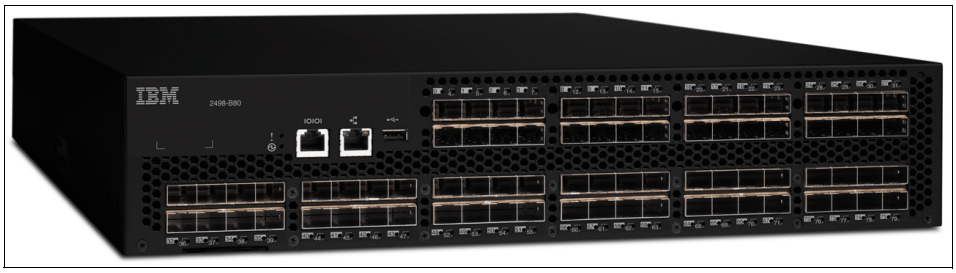

Figure 12-4 Front view of IBM System Storage SAN80B-4

The IBM System Storage SAN80B-4 SAN fabric switch provides 48, 64, or 80 active ports and is designed for high performance with 8 Gbps link speeds and backward compatibility to support links running at 4, 2, and 1 Gbps link speeds. High availability features make it suitable for use as a core switch in midrange environments or as an edge switch in enterprise environments. IBM Power Systems, System x, System z, and many non-IBM disk and tape devices are supported in many common operating system (OS) environments. Optional features provide specialized distance extension, dynamic routing between separate or heterogeneous fabrics, link trunking, IBM FICON, Server Application Optimization (SAO), performance monitoring and advanced security capabilities.

IBM System Storage SAN80B-4 demonstrates the same functions and features as the basic model of midrange switches SAN40B-4, the same operational capabilities, just with the following dissimilarities:

- Efficient 2U design with 48, 64, and 80 ports configuration on demand
- Significantly reduced power consumption by single chassis and fewer FRUs
- 16-port activation licensed feature

IBM System Storage SAN48B-5

The SAN48B-5 Switch is designed to meet the demands of hyper-scale, private cloud storage environments by delivering 16 Gbps Fibre Channel technology and capabilities that support highly virtualized environments.

The SAN48B-5 (2498-F48) delivers SAN technology within a flexible, simple, and easy-to-use solution. In addition to providing scalability, SAN48B-5 can address demanding Reliability, Availability, and Serviceability (RAS) requirements to help minimize downtime to support mission-critical cloud environments

The front view of IBM System Storage SAN48B-5 16 Gbps FC switch is shown in Figure 12-5.

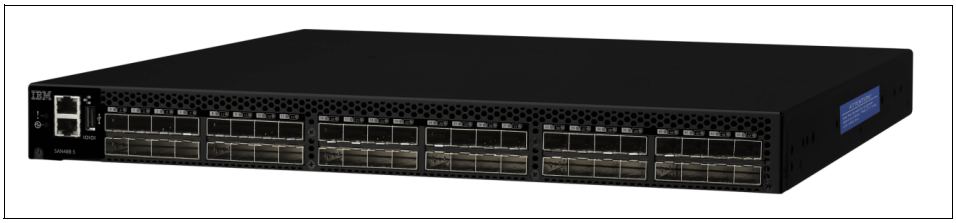

Figure 12-5 IBM System Storage SAN48B-5 fabric switch

Product highlights:

- 16 Gbps performance with up to 48 ports in an energy-efficient, 1U enclosure

- 2, 4, 8, 10, or 16 Gbps speed on all ports producing an aggregate 768 Gbps full-duplex throughput
- 128 Gbps high-performance and resilient frame-based trunking
- 10G Fibre Channel integration on the same port for DWDM metro connectivity on the same switch
- In-flight Data Compression and Encryption for efficient link utilization and security
- Redundant, hot-swap components and non-disruptive software upgrades
- Diagnostic Port (D-Port) feature for physical media diagnostic, troubleshooting and verification services
- Multi-tenancy in cloud environments through Virtual Fabrics, Integrated Routing, Quality of Service (QoS), and fabric-based zoning features

Cisco MDS 9148 for IBM System Storage

The Cisco MDS 9148 for IBM System Storage Multilayer Fabric Switch (2417-C48) is designed to provide an affordable, highly capable and scalable storage networking solution for small, midrange and large enterprise customers. It can be used as part of SAN solutions from simple single switch configurations to larger multi-switch configurations in support of fabric connectivity and advanced business continuity capabilities. As seen in Figure 12-6 on page 247 the switch is designed to offer outstanding value in a compact one-rack-unit (1U) form factor. With the ability to expand from 16 to 48 ports in eight-port increments, the Cisco MDS 9148 can be used as the foundation for small, stand-alone SANs, as a top-of-rack switch, or as an edge switch in larger core-edge SAN infrastructures.

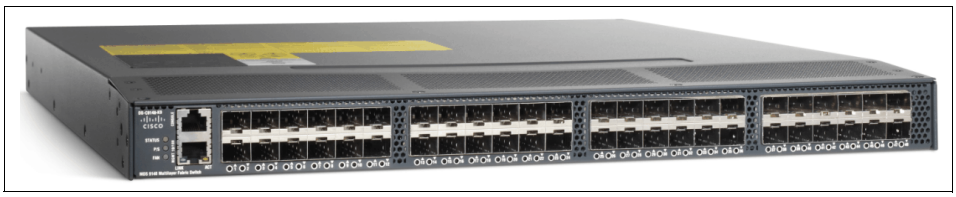

Figure 12-6 Cisco MDS 9148 for IBM System Storage

The Cisco MDS 9148 Multilayer Fabric Switch is designed to support quick configuration with zero-touch plug-and-play features and task wizards that allow it to be deployed quickly and easily in networks of any size. Powered by Cisco MDS 9000 NX-OS Software, it includes advanced storage networking features and functions and is compatible with Cisco MDS 9000 Series Multilayer Directors and Switches, providing transparent, end-to-end service delivery in core-edge deployments.

Fabric connectivity capabilities can be the basis for infrastructure simplification solutions for IBM System i, System p and System x servers and storage consolidation and high-availability server clustering with IBM System Storage disk arrays. Business continuity capabilities can help businesses protect valuable data with IBM System Storage tape subsystems and IBM Tivoli Storage Manager data protection software. The product incorporates the following features:

- Flexibility and scalability of up to 48 auto-sensing 1/2/4/8 Gbps ports in 1U standard rack-mountable or stand-alone unit
- Intelligent storage networking services powered by Cisco MDS 9000 NX-OS management software enabling Virtual SANs, Inter-VSAN routing (IVR), Port Channels, Quality of Service (QoS), and security
- Support for virtual environments with full N_port ID virtualization

- Simplified storage management and diagnostics enabled by Cisco Fabric Manager, FC ping, FC traceroute, Switched Port Analyzer (SPAN), Cisco Fabric Analyzer
- Support of 4/8 Gbps shortwave, 4 Gbps 4km longwave, or 4/8 Gbps 10km longwave SFPs
- Redundant power supplies and cooling

IBM System Storage SAN32B-E4 Encryption switch

Data is one of the most highly valued resources in a competitive business environment. Protecting that data, controlling access to it, and verifying its authenticity while maintaining its availability are priorities in our security conscious world. Increasing regulatory requirements are also helping to drive the need for the adequate security of data. Encryption is a powerful and widely used technology that helps protect data from loss and inadvertent or deliberate compromise.

The IBM System Storage SAN32B-E4 Encryption Switch (2488-E32) is a high performance standalone device designed for protecting data-at-rest in mission critical environments. In addition to helping IT organizations achieve compliance with regulatory mandates and meeting industry standards for data confidentiality, the SAN32B-E4 Encryption Switch also protects them against potential litigation and liability following a reported breach. The front view of the Encryption switch is shown in Figure 12-7.

Figure 12-7 IBM System Storage SAN32B-E4 Encryption switch

For data center fabric security, IBM provides advanced encryption services for Storage Area Networks (SAN) with the IBM System Storage SAN32B-E4 Encryption Switch. The switch is a high speed, highly reliable hardware device that delivers fabric based encryption services to protect data assets either selectively or on a comprehensive basis. The 8 Gbps SAN32B-E4 Fibre Channel Encryption Switch scales nondisruptively, providing from 48 up to 96 Gbps of encryption processing power to meet the needs of the most demanding environments with flexible, on-demand performance. It also provides compression services at speeds up to 48 Gbps for tape storage systems. Moreover, it is tightly integrated with one of the industry leading, enterprise class key management systems, the IBM Tivoli Key Lifecycle Manager (TKLM), which can scale to support key lifecycle services across distributed environments.

IBM System Storage SAN32B-E4 Encryption Switch at glance:

- 32 autosensing 8/4/2/1 Gbps active ports in base configuration
- Choice of 8 Gbps shortwave, longwave or extended distance SFPs
- Fabric based data-at-rest encryption with 48 Gbps disk and tape encryption processing. Transparent, online encryption of *cleartext* LUNs and re-keying of encrypted LUNs with no disruption
- Provides Fibre Channel interfaces E_port, F_port, FL_port, M_port, and optional EX_port
- Redundant power supplies and cooling

- Tight integration with IBM Tivoli Key Lifecycle Manager (TKLM) with support of multi-vendor disk storage subsystems

Discontinued midrange SAN switches

IBM has withdrawn the following products from the marketing, nevertheless these can be still available using the wide network of business partners, and definitely you find them in many small or medium business SAN solutions or clients' datacenters.

- IBM TotalStorage SAN32B-2
- IBM TotalStorage SAN32B-2 Express
- IBM TotalStorage SAN32M-2
- IBM TotalStorage SAN32M-2 Express
- IBM System Storage SAN64B-2
- Cisco MDS 9120 and 9140 Multilayer fabric switches
- Cisco MDS 9216i and 9216A Multilayer fabric switches
- Cisco MDS 9020 Fabric switch

Additional details about IBM midrange SAN switches are available under:

http://www.ibm.com/systems/storage/san/midrange/index.html

12.2.3 Enterprise SAN directors

The IBM Enterprise SAN directors provide to the datacenter networking infrastructure:

- The highest availability and scalability, and intelligent software to simplify management of complex, integrated enterprise SANs
- Heterogeneous Windows, Linux, iSeries, UNIX, and Mainframe servers
  - xSeries, iSeries, pSeries, and zSeries Server sales channels
  - IBM DSxxxx, FAStT, ESS, LTO, and ETS storage
- Supports the IBM System Storage Virtualization family of products and storage systems, IBM Tivoli Storage Manager, SAN Manager Storage Resource Manager, and Multiple Device Manager
- Offers customized solutions with competitive prices, worldwide IBM support, and IBM Global Services and IBM financial services

The IBM offers the following enterprise SAN directors through its marketing channels:

- IBM System Storage SAN384B-2
- IBM System Storage SAN768B-2
- Cisco MDS 9506 for IBM System Storage
- Cisco MDS 9509 for IBM System Storage
- Cisco MDS 9513 for IBM System Storage

IBM System Storage SAN384B-2 and SAN768B-2

The IBM System Storage SAN768B-2 and IBM System Storage SAN384B-2 fabric backbones are highly robust network switching platforms designed for evolving enterprise data centers. Each machine combines breakthrough performance, scalability and energy efficiency with long-term investment protection. Supporting open systems and IBM System z environments, these platforms address data growth and server virtualization challenges to:

- Enable server, SAN, and data center consolidation
- Minimize disruption and risk
- Reduce infrastructure and administrative costs.

Built for large enterprise networks, the SAN768B-2 has eight vertical blade slots to provide up to 384 16-Gbps or 512 8-Gbps FC ports. The SAN384B-2 is ideal for midsize core or edge

deployments, providing four horizontal blade slots and up to 192 16-Gbps or 256 8-Gbps FC ports. The flexible blade architecture also supports FCoE, fabric based encryption, SAN extension advanced functionality for high performance server, I/O consolidation, data protection and disaster recovery solutions.

The SAN768B-2 and SAN384B-2 are extremely efficient at reducing power consumption, cooling and the carbon footprint in data centers. While providing exceptional performance and scale, these networking backbones use less than one watt per Gbps. As members of the IBM System Storage family of b-type SAN products, the SAN768B-2 and the SAN384B-2 are designed to participate in fabrics containing other b-type and m-type devices manufactured by Brocade. This versatile hardware can serve as the backbone in a complex fabric and provide connections to other b-type and m-type directors, switches and routers.

The SAN768B-2 and SAN384B-2 backbones utilize Brocade Fabric OS (FOS), which provides several characteristic features, including Bottleneck Detection, Top Talkers (part of Advanced Performance Monitoring), and Adaptive Networking, a suite of tools that includes Ingress Rate Limiting, Traffic Isolation and Quality of Service (QoS). Managed through IBM System Storage Data Center Fabric Manager (DCFM) or the command line interface (CLI), these advanced capabilities help optimize fabric behavior and application performance.

Both products are shown in Figure 12-8.

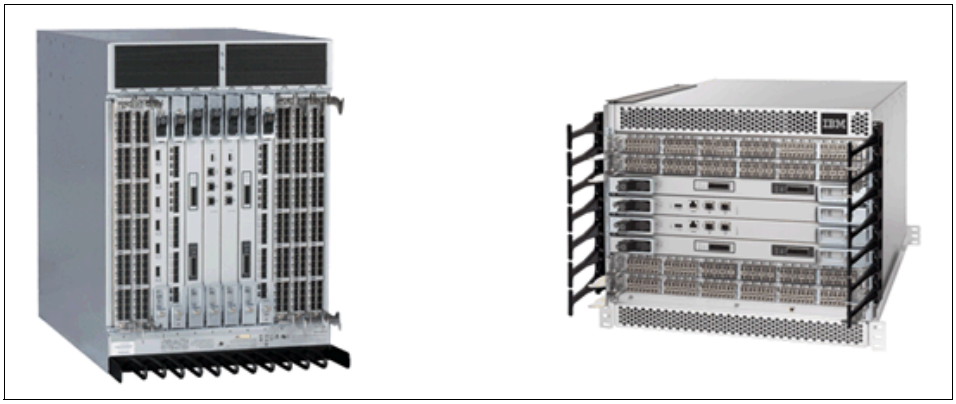

Figure 12-8 IBM System Storage SAN768B-2 (left) and SAN384B-2 (right)

Available blades for both models include:

- 16 Gbps 32-port or 48-port FC blades
- 8 Gbps 64-port FC blades
- 8 Gbps 16-port FC encryption blade
- 10 Gbps six-port FC blade
- FCoE 10 GbE 24-port blade supporting twenty four 10 Gbps CEE/FCoE ports
- 8 Gbps extension blade supporting twelve 8 Gbps FC ports and ten 1 Gbps GbE ports, or two optional 10 Gbps GbE ports

Here are the highlights of features and functions of the both products:

- Redundant control processor modules, power supplies and cooling
- Auto-sensing 16/8/4/2 Gbps E_Port, F_Port, FL_Port, EX_port, and 10 Gbps E_Port Fibre Channel interfaces and 10 Gbps converged ethernet ports
- 16/8/4/2 FICON interfaces and 1 GbE or 10 GbE FCIP ethernet interfaces
- Management protocols that include HTTP, SNMP v1/v3 (FE MIB, FC Management MIB), Telnet

- Management and operational software includes Auditing, Syslog, Web Tools, Fabric Watch, IBM System Storage Data Center Fabric Manager (DCFM), command line interface (CLI)
- Full backward compatibility with IBM System Storage and IBM TotalStorage b-type and m-type SAN directors, switches and routers; other directors, switches and routers manufactured by Brocade
- Fabric based data-at-rest encryption for disk array LUNs, heterogeneous tape drives, and virtual tape libraries to enforce data confidentiality and privacy requirements (with selected blades); support for IBM Tivoli Key Lifecycle Manager (TKLM)
- Advanced features as ISL trunking, Server Application Optimization (SAO), Virtual Fabrics, Adaptive Networking Services, Advanced Performance Monitoring, support for extended fabrics up to 10km

Cisco MDS 9506 and MDS 9509 for IBM System Storage

The Cisco MDS 9506 and MDS 9509 Multilayer Directors for IBM System Storage support 1, 2, 4, 8 and 10 Gbps Fibre Channel switch connectivity and intelligent network services to help improve the security, performance and manageability required to consolidate geographically dispersed storage devices into a large enterprise SAN. Administrators can use these directors to help address the needs for high performance and reliability in SAN environments ranging from small workgroups to very large, integrated global enterprise SANs.

Both, the Cisco MDS 9506 and MDS 9509 (Figure 12-9 on page 252) for IBM System Storage utilize two Supervisor-2 Modules designed for high availability and performance. The Supervisor-2 Module combines an intelligent control module and a high performance crossbar switch fabric in a single unit. It uses Fabric Shortest Path First (FSPF) multipath routing, which provides intelligence to load balance across a maximum of 16 equal-cost paths and to dynamically reroute traffic if a switch fails.

Each Supervisor-2 Module provides the necessary crossbar bandwidth to deliver full system performance in the MDS 9506 director with up to four Fibre Channel switching modules (8 for MDS 9509). It is designed to provide that loss or removal of a single crossbar module has no impact on system performance. Fibre Channel switching modules are designed to optimize performance, flexibility and density. The Cisco MDS 9506 Multilayer Director requires a minimum of one and allows a maximum of four switching modules, while MDS 9509 utilizes from 1 up to 16 modules. These modules are available in either a 12-, 24- and 48-port 4 Gbps configurations, allowing the Cisco MDS 9506 to support 12 to 192 Fibre Channel ports per chassis (336 in MDS 9509). Optionally, a 4-port 10 Gbps Fibre Channel module is available for high performance inter-switch link (ISL) connections over metro optical networks.

Figure 12-9 Cisco 9506 (left) and MDS 9509 (right) Multilayer Directors

Advanced traffic management capabilities are integrated into the switching modules to help simplify deployment and to optimize performance across a large fabric. The PortChannel capability allows users to aggregate up to 16 physical 2 Gbps Inter-Switch Links into a single logical bundle, providing optimized bandwidth utilization across all links. The bundle may span any port from any 16-port switching module within the chassis, providing up to 32 Gbps throughput.

Available switching modules:

- 12, 24, or 48 ports 4 Gbps Fibre Channel module
- 24 or 48 ports 8 Gbps Fibre Channel module
- 4 ports 8 Gbps with 44 ports 4 Gbps Fibre Channel module
- 4 ports 10 Gbps Fibre Channel module

Highlights include:

- Provides Fibre Channel throughput of up to 8 Gbps per port and up to 64 Gbps with each PortChannel Inter-Switch Link connection
- Offers scalability from 12 to 192 (336) Fibre Channel ports
- Offers 10 Gbps ISL ports for Inter-Data Center links over metro optical networks
- Offers Gigabit Ethernet IP, GbE ports for iSCSI or FCIP connectivity over global networks
- Includes Virtual SAN (VSAN) capability for SAN consolidation into virtual SAN islands on a single physical fabric
- Includes high-availability design with non-disruptive firmware upgrades
- Enterprise, SAN Extension over IP, Mainframe, Storage Services Enabler and Fabric Manager Server Packages provide added intelligence and value

Cisco MDS 9513 for IBM System Storage

The Cisco MDS 9513 for IBM System Storage provides 12 to 528 Fibre Channel ports, with 4 and 8 Gbps support and a high-availability design. It offers 4 to 44 10 Gbps ports for ISL connectivity across metro optical networks. It includes Virtual SAN (VSAN) capability for SAN consolidation into virtual SAN “islands” on a single physical fabric. The Cisco MDS 9513 provides network security features for large enterprise SAN deployment. The director also offers intelligent networking services to help simplify mainframe FICON and Fibre Channel

SAN management and reduce total cost of ownership (TCO). See the front view of the chassis in Figure 12-10.

Figure 12-10 Front view of Cisco MDS 9513

The Cisco MDS 9513 Multilayer Director utilizes two Supervisor-2 Modules, designed to support high availability. The Supervisor-2 Module is designed to provide industry leading scalability, intelligent SAN services, non-disruptive software upgrades, statusful process restart and failover, and redundant operation. Dual crossbar switching fabric modules provide a total internal switching bandwidth of 2.4 Tbps for inter-connection of up to eleven Fibre Channel switching modules.

Fibre Channel switching modules improve performance, flexibility and density. The Cisco MDS 9513 for IBM System Storage requires a minimum of one Fibre Channel switching module and allows a maximum of 11. These modules are available in 12-, 24- or 48-port 4 and 8 Gbps configurations, enabling the Cisco MDS 9513 to support 12 to 528 Fibre Channel ports per chassis. Optionally, a 4-port 10-Gbps Fibre Channel module is available for high-performance Inter-Switch Link (ISL) connections over metro optical networks.

Highlights include:

- Supports Fibre Channel throughput of up to 8 Gbps per port and up to 64 Gbps with each PortChannel Inter-Switch Link (ISL) connection
- Offers Gigabit Ethernet (GbE) IP ports for iSCSI or FCIP connectivity over global networks
- Offers scalability from 12 to 528 1, 2, 4, and 8 Gbps Fibre Channel ports
- High-availability design with support for non-disruptive firmware upgrades Includes Virtual SAN (VSAN) capability for SAN consolidation into virtual SAN ‘islands’ on a single physical fabric
- Enterprise, SAN Extension over IP, Mainframe, Storage Services Enabler and Fabric Manager Server Packages provide added intelligence and value

Discontinued enterprise SAN directors

IBM has withdrawn the following products from the marketing, nevertheless one can still find them in medium business or large strategic datacenters and clients' environments.

- IBM TotalStorage SAN Director M14
- IBM TotalStorage SAN140M
- IBM TotalStorage SANC40M
- IBM TotalStorage SAN256M
- IBM TotalStorage SAN256B
- IBM System Storage SAN384B
- IBM System Storage SAN768B

The comprehensive information and additional details about IBM SAN directors are available under the link:

http://www.ibm.com/systems/storage/san/enterprise/

12.2.4 Multiprotocol routers

IBM currently has two multiprotocol routers in its portfolio:

- IBM System Storage SAN06M-R
- Cisco MDS 9222i for IBM System Storage

IBM System Storage SAN06M-R multiprotocol router

An entry-level multiprotocol extension router designed to connect two Storage Area Networks (SANs) over a wide distance using the Internet as the interconnection fabric (upgrades to enterprise level functions are available). Intended to support business continuity solutions between supported servers at one site and support servers or IBM System Storage disk or tape devices at a distant location.

SAN06M-R delivers high performance with 8 Gbps FC ports and hardware assisted traffic processing for line-rate performance. It utilizes existing Internet, IP-based infrastructures for metro and global SAN extension for business continuity solutions. Up to eight virtual FCIP tunnels are available to help maximize scalability and utilization of metropolitan area network (MAN) or wide area network (WAN) resources. Router provides hardware-based compression, large window sizes and selective acknowledgement of IP packets are designed to optimize performance of SAN extension over IP networks. Front view of the device is shown in Figure 12-11.

Figure 12-11 Front view of SAN06M-R

Summary of the product and its highlights:

- 1U 19" packaging designed for rack-mount or table-top
- Designed for high-performance with up to 8 Gbps FC autosensing ports and up to 1 Gbps Ethernet (GbE) ports
- Supports either 8, 4, 2 or 4, 2, and 1 Gbps FC link speeds

- Shortwave and longwave SFPs can be intermixed in the same router
- Hardware-based compression with extensive buffering
- Optional FICON with CUP and FICON Accelerator enable support for enterprise class environments
- SAN isolation from Internet, WAN or MAN failures.

Cisco MDS 9222i for IBM System Storage

The Cisco MDS 9222i for IBM System Storage is designed to address the needs of medium sized businesses and large enterprises with a wide range of Storage Area Network (SAN) capabilities. It can be used as a cost effective high performance SAN extension over IP router switch for midrange SMB customers in support of IT simplification and business continuity solutions. It can also be used as a remote site router switch for device aggregation and SAN extension over IP to data center directors for large enterprise customers. Business continuity solutions include data protection with IBM System Storage tape libraries and devices and IBM Tivoli Storage Manager data protection software; and disaster protection with IBM System Storage disk metro and global mirroring disaster recovery solutions. The front view of the MDS 9222i is shown in Figure 12-12.

Figure 12-12 Front view of Cisco MDS 9222i

The product highlights include:

- Base switch includes eighteen 4 Gbps Fibre Channel ports with two shortwave SFPs; four GbE IP port with SAN Extension over IP
- Optional 8 Gbps FC switch ports support optical SFP+ transceivers and 10 Gbps ISL ports support optical X2 transceivers for edge switch attachment to Cisco MDS 9500 directors
- SAN extension with high performance FCIP acceleration and hardware based compression capabilities and security with hardware-based encryption
- Multiservice design for high performance business continuity solutions with Windows, UNIX, Linux, NetWare, IBM OS/400 and IBM z/OS servers
- Includes Virtual SAN (VSAN) capability for SAN consolidation into virtual SAN islands on a single physical fabric

Discontinued IBM multiprotocol routers

The following products and devices are no longer available through the IBM marketing channels:

- IBM TotalStorage04M-R
- IBM TotalStorage SAN16B-R
- IBM TotalStorage SAN16M-R
- IBM System Storage SAN18B-R

For more details and additional technical description of each of the products, visit:

http://www.ibm.com/systems/storage/san/routers/index.html

12.3 IBM System Storage Disk Systems

The IBM System Storage Disk Systems products and offerings provide compelling storage solutions with superior value for all levels of business. In the following section we briefly discuss SAN attached storage disk device subsystems as a key component of every datacenter to keep data available in an effective and cost-efficient way.

We do not discuss network-attached storage (NAS) as that is not the primary objective of this publication. SAN disk systems for storage virtualization and cloud computing are described in 12.5, “Storage virtualization and Cloud Computing” on page 271.

We will not provide detailed information about expansion units (EXP) of each of the disk subsystems, and we will refer the reader to IBM resources for more information on enclosures.

12.3.1 Entry level disk systems

Designed to deliver advanced functionality at a breakthrough price, these systems provide an exceptional solution for workgroup storage applications such as email, file, print and Web servers, as well as collaborative databases and remote boot for diskless servers. We describe the following product:

- IBM System Storage DS3500 Express

Enclosures and expansion units (not discussed in this section) include:

- IBM System Storage EXP2500 Express
- IBM System Storage EXP3000 Express
- IBM System Storage EXP3512 Expansion Enclosure
- IBM System Storage EXP3524 Expansion Enclosure

IBM System Storage DS3500 Express

The IBM System Storage DS3500 Express Storage™ Systems are the newest addition to the IBM System Storage DS3000 series family of entry disk storage systems. The DS3500 delivers affordable, entry-level configurations for small and medium businesses in compact 2U, 19-inch rack mount enclosures, with the flexibility to scale in capacity, performance, host interfaces, and advanced functions as your business grows or requirements change.

The DS3500 combines next-generation controller technology with the latest, high-performance host interface technologies to deliver new levels of performance to the DS3000 series. With the DS3500, you choose the initial system configuration that matches your performance requirements and budget:

- Single controller for an entry solution with low initial investment
- Dual controllers for higher performance
- Dual controller with the Turbo Performance option for the best results

The DS3500 is available in two models (see Figure 12-13):

- DS3512 Express with 12 3.5-inch 6 Gbps SAS attached disk
- DS3534 Express with 24 2.5-inch 6 Gbps SAS attached disk

Figure 12-13 IBM System Storage DS3512 (top) and DS3524 (bottom)

Product highlights:

- Compact 2U 19-inch rack mountable chassis
- Single or dual controllers, 6 Gbps SAS attached 3.5-inch (DS3512) or 2.5-inch (DS3524) disks
- Support for 192 disk drives through the attachment of EXP3500 expansion units (EXP3512 with 12 3.5-inch or EXP3524 with 24 2.5-inch disk bays)
- 8 Gbps FC, 6 Gbps SAS, 1 Gbps and 10 Gbps iSCSI host connectivity available
- Support for up to 128 storage partitions with RAID levels 0, 1, 3, 5, 6, 10
- Dual redundant hot-swappable power supplies and cooling fans

Discontinued entry disk systems

IBM has withdrawn the following product portfolio from the marketing:

- IBM TotalStorage DS3100
- IBM TotalStorage DS3200
- IBM TotalStorage DS3300
- IBM System Storage DS3400
- IBM TotalStorage EXP24 Expansion Unit
- IBM TotalStorage EXP420 Expansion Unit

For the additional details about each product and especially for storage expansion units that were not covered in previous text, visit:

http://www.ibm.com/systems/storage/disk/entry/index.html

12.3.2 Midrange disk systems

Throughout this section we provide a brief overview of disk storage device systems that IBM offers for small and medium business solutions. Again, we do not describe storage enclosures and expansion units in detail.

We describe:

- IBM System Storage DS5020 Express
- IBM System Storage DS5000
- IBM System Storage DS3950 Express

Enclosures that are available:

- IBM System Storage EXP5060 Expansion unit for DS5000 family
- IBM System Storage EXP520 Expansion unit for DS5020 Express
- IBM System Storage EXP395 Expansion unit for DS3950

IBM System Storage DS5020 Express

Optimized data management requires storage solutions with high data availability, strong storage management capabilities and powerful performance features. The IBM System Storage DS5020 Express is designed to provide lower total cost of ownership (TCO), high performance, robust functionality and unparalleled ease of use.

Additionally, auto-negotiating 8 Gbps Fibre Channel interfaces allow the DS5020 Express to integrate seamlessly into an existing 2 Gbps or 4 Gbps infrastructure, while offering investment protection going forward when the SAN inevitably becomes 8 Gbps.

Apart from 8 Gbps FC connections, the DS5020 Express offers optional 1 Gbps iSCSI interface for less demanding applications and lower cost implementation, up to 67.2 TB of Fibre Channel or FC-SAS physical storage capacity, 224 TB of SATA physical storage capacity, 33.6 TB of SSD physical storage capacity, and powerful system management, data management and data protection features. See Figure 12-14 for the front view.

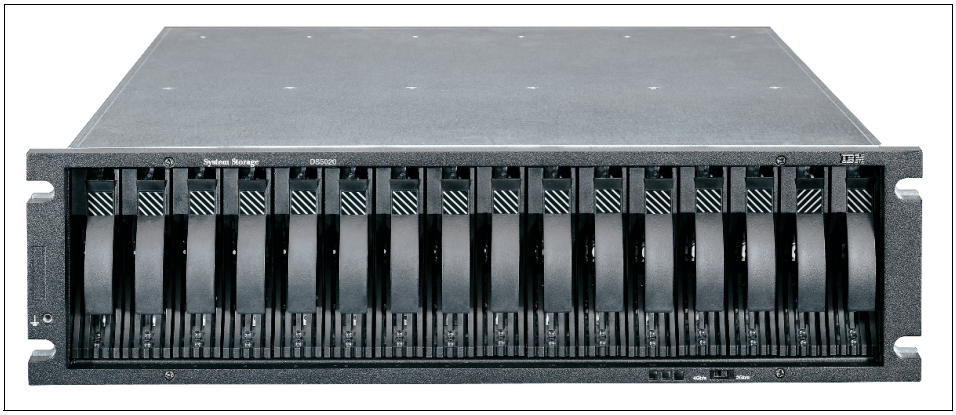

Figure 12-14 Top front view of DS5020 Express model

The product highlights:

- High-performance 8 Gbps Fibre Channel (FC) connections and 1Gbps iSCSI
- Up to 224 TB of physical storage capacity with 112 (using 6 EXP520 expansion units) 2 TB SATA disk drives in up to 128 storage partitions
- Transparent management of DS3000 and DS5000 products using DS Storage Manager
- Support for intermixing Fibre Channel/FC-SAS/SED/SATA/SSD drives enables cost-effective tiered storage
- Redundant, hot-swappable power supplies and cooling

IBM System Storage DS5000

DS5000 series storage systems (DS5100 and DS5300) are designed to meet the demanding open-systems requirements of today and tomorrow, while establishing a new standard for life cycle longevity with field-replaceable host interface cards. Seventh generation architecture delivers relentless performance, real reliability, multidimensional scalability and unprecedented investment protection.

The DS5000 storage systems are equally adept at supporting transactional applications such as databases and On Line Transaction Processing (OLTP), throughput-intensive applications such as high-performance computing (HPC) and rich media, and concurrent workloads for consolidation and virtualization. With relentless performance and superior reliability and availability, DS5000 series storage systems can support the most demanding service level agreements (SLAs) for the most common operating systems, including Microsoft Windows, UNIX and Linux. And when requirements change, you can add or replace host interfaces, grow capacity, add cache and reconfigure the system on the fly.

Figure 12-15 shows the top front view of IBM System Storage DS5100 as a member of DS5000 family of IBM storage products.

Figure 12-15 Top front view of DS5100

The product highlights include:

- Efficient, compact 4U packaging designed for 19" rack
- Easy-to-use, easy-to-configure management interface able to manage DS3000, DS4000 and DS5000 series storage systems
- Scalable up to 448 drives using the EXP5000 enclosure and up to 960 TB of high-density storage with the EXP5060 enclosure
- Support for intermixing drive types (FC, FC-SAS, SED, SATA and SSD) and host interfaces (8 Gbps FC and 1/10 Gbps iSCSI)
- Two performance levels (base: DS5100 and high: DS5300) with ability to field-upgrade performance levels
- Designed to support high availability with hot-swappable components and non-disruptive firmware upgrades

IBM System Storage DS3950 Express

As part of the DS series, the DS3950 Express offers high performance 8 Gbps capable Fibre Channel connections, optional 1 Gbps iSCSI interface for less demanding applications and lower cost implementation, up to 67.2 TB of Fibre Channel or FC-SAS physical storage capacity, up to 224 TB of SATA physical storage capacity, and powerful system management, data management and data protection features. The DS3950 Express is designed to expand from workgroup to enterprise-wide capability with up to six Fibre Channel expansion units with the EXP395 Expansion Unit.

The DS3950's design avoids over-configuration for an affordable entry-point while offering seamless "pay-as-you-grow" scalability as requirements change, that categorizes the product to the midrange category. Its efficient storage utilization lowers raw capacity requirement, and support for intermixing high performance and high capacity drives enables enclosure-based tiered storage. The IBM System Storage DS3950 Express model is shown in Figure 12-16.

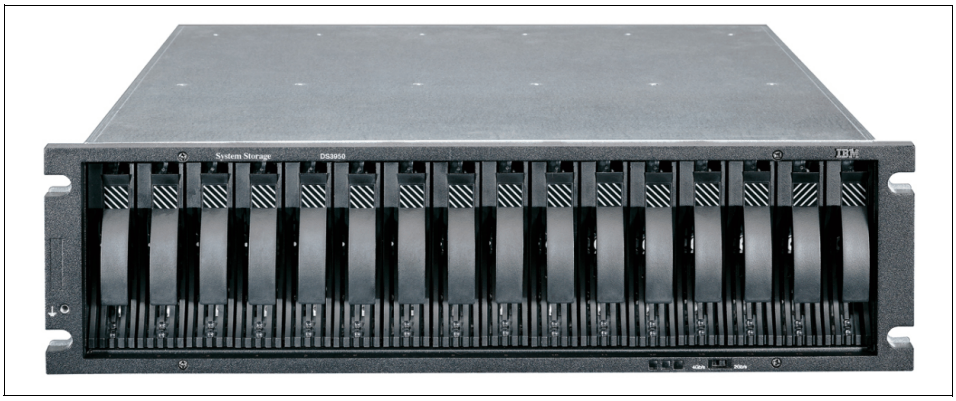

Figure 12-16 Top front view of DS3950 Express

The DS3950 Express highlights:

- Support for intermixing drive types (FC, FC-SAS, and SATA) and host interfaces (8 Gbps FC and 1 Gbps iSCSI)
- Support up to 112 disk drive modules with the attachment of six EXP395 expansion units that gives up to 224 TB of physical storage capacity, in up to 128 storage partitions
- Fully integrated replication features such as IBM FlashCopy®, Enhanced Remote Mirror, VolumeCopy
- Hot-swapable redundant power supplies and cooling fans in standard 19-inch rack-mountable unit

Discontinued midrange disk products

IBM has withdrawn the following products from the marketing, nevertheless one can still find them in medium business or large strategic datacenters and clients' environments.

- IBM TotalStorage DS4200
- IBM TotalStorage DS4700
- IBM TotalStorage DS4800
- IBM TotalStorage EXP4000 Expansion unit

For the additional details about each product and especially for storage expansion units that were not covered in previous text, follow the link:

http://www.ibm.com/systems/storage/disk/midrange/index.html

12.3.3 Enterprise disk systems

With their high capacity, scalability, broad server support, and virtualization features, the enterprise class storage systems are well suited for simplifying the storage environment by consolidating data from multiple storage systems on a single system. In this section we discuss these high-end, enterprise class IBM disk device subsystems:

- IBM System Storage DS8000 family
- IBM System Storage DSC3700

IBM System Storage DS8000

The IBM System Storage DS8000 offers high-performance, high-capacity, secure storage systems that are designed to deliver resiliency and total value for the most demanding, heterogeneous storage environments.

The DS8000 series is built on powerful and market-proven IBM POWER® microprocessors in dual two-way or dual four-way shared Symmetric Multi-Processor (SMP) complexes. The latest DS8800 model (Figure 12-17) introduces a new generation of hardware and is the fastest disk system in the IBM storage portfolio. With dual IBM POWER6® controllers, 8 Gbps host and device adapters, and 6 Gbps SAS disk drives, the DS8800 delivers a new level of performance in a design that condenses more data in a smaller footprint. Compared to the POWER5 processor in DS8700 models, the POWER6 processor offers up to over a 40 percent performance improvement. These upgrades, along with the new 2.5-inch SAS drives, deliver a dramatic boost in data throughput in a more compact footprint - a powerful combination aimed at taming the most demanding enterprise applications.

Figure 12-17 Three frames of DS8800 represent the extraordinary scalability

Despite their distinct hardware components, both the DS8800 and DS8700 models now ship with a common microcode built on over ten years of market proven reliability. This common microcode not only reinforces our commitment to superior reliability, it also enables us to deliver new functions on both the DS8800 and DS8700 hardware platforms concurrently.

Certainly, DS8700 clients will appreciate the investment protection for their deployments, and, of course, all DS8000 clients will continue to enjoy the interoperability of most remote mirror and copy functions across older DS8300, DS8100, and IBM TotalStorage Enterprise Storage Server (ESS) models.

The DS8000 series also supports a variety of major server platforms, including IBM z/OS, z/VM®, Linux on System z, IBM i, OS/400, i5/OS® and AIX operating systems, as well as Linux, HP-UX, Sun Solaris, Novell NetWare, VMware and Microsoft Windows environments, among many others. With such broad platform support, the DS8000 series can easily accommodate a wide array of applications and their distinct service levels.

IBM System Storage DS8700 and DS8800 disk device subsystems represent simply the world class performance to help optimize responsiveness in an on-demand world with minimum *five-nines* (99.999%) availability.

Here are the most important highlights and features:

- Dual Symmetric Multi-Processing (SMP) processor complexes
- 4 Gbps FC attached disk drives (DS8700) and 6 Gbps SAS (DS8800)
- Support for 2 to 32 host adapters and up to 128 FC/FICON 8 Gbps host ports
- Minimum of 8 drives and maximum of 1056 drives scales up to 2048 TB of physical storage capacity
- Up to 384 GB cache memory with innovative caching algorithms
- Storage Pool Striping to automatically avoid disk hotspots
- I/O Priority Manager aligns quality of service levels to separate application workloads in the system
- IBM FlashCopy, Metro Mirror, Global Mirror, Metro/Global Mirror and Global Copy provide flexible replication services
- IBM System Storage Easy Tier feature automatically helps optimize solid-state storage (SSD) deployments in multi-tier systems
- Full Disk Encryption drive options for advanced protection of data at-rest

IBM System Storage DSC3700

The IBM System Storage DCS3700 is designed to meet the storage needs of highly scalable, data streaming applications in high performance computing environments.

Developed with density in mind, the DCS3700 Storage System delivers up to 120 TB of physical capacity in a slim, 4U form factor. With the attachment of up to two DCS3700 Expansion Units, the physical capacity can be scaled up to 360 TB, all within a 12U space.

The DCS3700 (Figure 12-18 on page 263) features the latest technologies, including 6 Gbps SAS and 8 Gbps FC host interfaces, along with 6 Gbps SAS drives. The DCS3700 is designed to be equally adept at delivering throughput to bandwidth-intensive applications and I/O operations to transactional applications, such as databases and Microsoft Exchange. That is why the product is classified as an enterprise disk storage system, even if the dimensions are not typical for this category.

DC3700 is ideally suited for high performance streaming applications, such as rich media, financial markets, telecommunications, weather modeling and others needing rigorous bandwidth requirements

Figure 12-18 IBM System Storage DSC3700

Benefits and highlights:

- 6 Gbps SAS high density storage system delivering scalable capacity at an affordable price point
- Multilevel data protection with IBM FlashCopy, Volume Copy and Remote Mirroring across FC
- Mixed host interfaces support DAS and SAN tiering to reduce overall operation and acquisition costs
- Investment protection and cost-effective backup and recovery with remote mirror across FC host ports and compatibility with DS3500, DS5000 and DS4000
- Up to 180 drives per system with the attachment of two DCS3700 Expansion Units (60 drives per enclosure) 20 drives minimum drive quantity per enclosure

Discontinued enterprise disk systems

The following products IBM no longer offers through its marketing channels, but these are still seen in large and strategic hosting datacenters across geographies:

- IBM TotalStorage Enterprise Storage Server (ESS)
- IBM System Storage DS8100
- IBM System Storage DS8300

More details about IBM high-end, enterprise disk systems are available at:

http://www.ibm.com/systems/storage/disk/enterprise/index.html

12.4 IBM Tape Storage Systems

Tape systems have traditionally been associated with the mainframe computer market, because they represented an essential element in mainframe systems architecture since the early 1950s as a cost-effective way to store large amounts of data. In contrast, the midrange and client/server computer market has made limited use of tape technology until recently.

However, over the past few years, growth in the demand for data storage and reliable backup and archiving solutions has greatly increased the need to provide manageable and cost-effective tape library products. The value of using tape for backup purposes has only gradually become obvious and important in these environments.

In this section we briefly guide you through the IBM Fibre Channel attached tape drives, autoloaders and automated tape libraries. We do not discuss the current tape technology (mainly Linear Tape Open - LTO) as this topic is widely described in other publications such as the *IBM System Storage Tape Library Guide for Open Systems*, SG24-5946. A detailed technical guide is available here:

http://www.ibm.com/systems/storage/product/tape.html

12.4.1 Fibre Channel tape drives

Client data is vital to business operations. IBM offers entry-level and enterprise tape products that are designed to provide backup and protection of client data at an appealing cost to a company's budget. In this section we give an overview of these tape drives that include:

- IBM System Storage TS2230, TS2240, TS2250 half-height LTO tape drives
- IBM System Storage TS2340 and TS2350 full-height LTO tape drives
- IBM System Storage TS1120, TS1130, TS1140 for 3590/3592 cartridges

We only intend to describe Fibre Channel connected tape drives (i.e. 3/6 Gbps SAS attached, LVD/HVD SCSI attached) as the book is focused on SAN-related networking technology.

IBM Ultrium LTO Half-Height tape drives

IBM Ultrium Linear Tape Open (LTO) Half-Height external tape drives include the following models:

- IBM System Storage TS2230 Tape Drive Express for LTO-3 cartridges
- IBM System Storage TS2240 Tape Drive Express for LTO-4 cartridges
- IBM System Storage TS2250 Tape Drive Express for LTO-5 cartridges

TS2230 and TS2240 tape drives are suited for handling backup, save and restore, and archival data storage needs with higher capacity and higher data transfer rate than previous generation. Additionally, TS2250 tape drives continue to support tape encryption as the previous generation of LTO4. They all support Write-Once-Read-Many (WORM) tape cartridges for compliance purposes.

Figure 12-19 shows all three models of IBM Ultrium LTO Half-Height external tape drives.

Figure 12-19 IBM TS2230 (top left), TS2240 (top bottom), and TS2250 (right)

Note: IBM External Half-High tape drives utilize 6 Gbps SAS connection only. However, there is available Fibre Channel (FC) attached model, dedicated for internal installation into entry tape libraries and enclosures. These versions are not considered as external tape drives, therefore not covered in detail in this section.

Product highlights:

- External standalone or rack mountable, easy-to-install units
- Standard 6 Gbps SAS interface for external models; 8 Gbps FC (4 Gbps LTO3/LTO4) interface for internal half-height LTO5 tape drives
- Support for WORM cartridges and hardware compression, generation LTO5 offers tape partitioning on selected platforms
- 1.5 TB native capacity (3.0 TB compressed) for LTO5 tape cartridges (800 GB/1.6 TB for LTO4, 400 GB/800 GB for LTO3)
- Maximum data rate 140 MBps (LTO5), 120 MBps (LTO4), 80 MBps (LTO3)
- Transparent installation and configuration across various supported platforms utilizing dedicated IBM tape device drivers

IBM Ultrium LTO Full-Height tape drives

IBM Ultrium LTO Full-Height external tape drives incorporate industry-proved IBM Linear Tape Open (LTO) full sized tape drives of the series 3580. They are delivered in compact standalone or rack mountable enclosures easy-to-install. IBM offers two models of external tape drives in this category:

- IBM System Storage TS2340 Tape Drive Express for LTO-4 cartridges
- IBM System Storage TS2350 Tape Drive Express for LTO-5 cartridges

Similar to the half-height external tape drives, the full-height tape drives are offered by IBM as internal modules for installation in midrange or enterprise tape libraries using either 3/6 Gbps SAS or 8/4 Gbps FC interface. The both devices are presented in Figure 12-20.

Figure 12-20 Full-Height IBM External tape drives TS2340 (left) and TS2350 (right)

Brief summary of features and benefits of the both products:

- External standalone or rack mountable, easy-to-install devices
- TS2340 - 3 Gbps or 320 MBps SCSI interface, TS2350 - 6 Gbps SAS port. Internal tape drives type 3580 - 4 Gbps FC interface for LTO4, 8 Gbps FC attached LTO5 tape drives
- 1.5 TB native capacity (3.0 TB compressed) for LTO5 tape cartridges (800 GB/1.6 TB for LTO4
- Maximum data transfer rate 140 MBps (LTO5), up to 120 MBps (LTO4)
- Available WORM cartridges and hardware compression, generation LTO5 offers tape partitioning on selected platforms.

IBM System Storage 3592 tape drives

IBM System Storage 3592 series of tape drives offer a design that is focused on high capacity and performance, and high reliability for storing your mission critical data. This series includes the following available IBM tape drives:

- IBM System Storage TS1120
- IBM System Storage TS1130
- IBM System Storage TS1140

Whether configured as a standalone drive or part of an automated tape library, these models offer both fast access to data and high capacity in a single drive, helping to reduce the complexity of your tape infrastructure. They have the same form factor as its predecessors and machine type 3592. See Figure 12.5 for details.

Figure 12-21 IBM TS1120 (left), TS1130 (center) and TS1140 (right)

To help optimize drive utilization and reduce infrastructure requirements, these three models of tape drives can be shared among supported open system hosts on a Storage Area Network (SAN) or IBM FICON mainframe hosts when attached to an IBM System Storage Tape Controller for System z.

Product highlights:

- High performance with data transfer rate up to 650 MBps with compression
- Flexible media, including short and long length cartridges, re-writable and WORM formats and media partitioning on given platforms
- 4 TB capacity using JC/JY media, 1.6 TB using JB/JX media, 500 GB using JK media (TS1140)
- 1 TB capacity using JB/JX media, 640 GB using JA/JW media, 128 GB using JJ/JR media (TS1130)
- 700 GB capacity using JB/JX media, 300/500 GB using JA/JW media, 60/100 GB capacity using JJ/JR media (TS1120)
- IBM Power Systems, System i, System p, System z and System x support

12.4.2 Autoloaders and entry tape libraries

Single tape drive autoloaders and entry-class tape libraries are suited to handle backup, save and restore and for archival data storage needs in small to medium size environments. They benefit from Linear Tape Open technology as they typically incorporate IBM half-height tape drives connected over Fibre Channel or 3/6 Gbps SAS interface. Following, we only describe the model that is able to accommodate FC-attached full-height tape drives and it is:

- IBM System Storage TS3200 Tape Library Express

We do not discuss other available IBM entry tape automation products, as they do not offer an FC interface and are directly attached to the host systems:

- IBM System Storage TS2900 Tape Autoloader
- IBM System Storage TS3100 Tape Library Express

Technical details of those devices (including TS3200) can be found at:

http://www.ibm.com/systems/storage/tape/entry/index.html

IBM System Storage TS3200 Tape Library Express

The TS3200 Express Model and its storage management applications are designed to address capacity, performance, data protection, reliability, affordability and application requirements. It is designed as a functionally rich, high capacity, entry level tape storage solution incorporating LTO Ultrium tape technology. The IBM TS3200 Express model is an excellent solution for large capacity or high performance tape backup with or without random access. The TS3200 is also an excellent choice for tape automation for IBM Power Systems, IBM System x and other open systems.

IBM TS3200 Express (Figure 12-22 on page 267) is designed to support the newest generation of LTO with up to two IBM Ultrium 5 full-height tape drives or up to four IBM Ultrium 5 half-height tape drives, as well as LTO generations 3 and 4 tape drives using a 4U form factor.

Figure 12-22 Top front view of IBM TS3200 Express

IBM System Storage TS3200 Tape Library Express provides these features:

- Supports the LTO-5, LTO-4, or LTO-3 tape drives, Low Voltage Differential (LVD) 320 MBps SCSI, 8 Gbps FC and 6 Gbps SAS attachments
- Up to 2 full-height or 4 half-height IBM tape drives
- Sequential or random access mode with a standard barcode reader
- 4U rack mountable or standalone form factor with up to 48 cartridge slots provides native capacity up to 72 TB (144 TB with LTO-5 compression)
- Remote library management through a transparent, intuitive Web interface

12.4.3 Midrange tape libraries

Whether a small/medium size business is expanding operations or experiencing rapid data growth, IBM midrange tape libraries are designed to help meet the customer needs of data backup, archive and management. These tape products are designed to offer reliability, performance and flexibility for today and the future. This section introduces the IBM tape automation product:

- IBM System Storage TS3310 Tape Library.

We provide the typical features of this midrange tape library. For deep technical details, including all the available configuration scenarios, refer to *IBM System Storage Tape Library Guide for Open Systems*, SG24-5946, and the most current updates are maintained under:

http://www.ibm.com/systems/storage/tape/ts3310/index.html

The following midrange product has been withdrawn from IBM offerings:

- IBM System Storage TS3400 Tape Library.

IBM System Storage TS3310

The IBM System Storage TS3310 Tape Library is a modular, scalable tape library designed to address the tape storage needs of rapidly growing companies who find themselves space and resource constrained with tape backup and other tape applications.

Designed around a 5U high modular base library unit, the TS3310 can scale vertically with expansion for LTO tape cartridges, drives and redundant power supplies. Each expansion module (9U) contains 92 physical LTO cartridge storage cells and space for up to four LTO-5, LTO-4, and LTO-3 tape drives. Additionally, the module has space for up to two power supply modules - with one redundant. See Figure 12-23 on page 268 for the configuration of L5B base module and one expansion unit E9U.

Figure 12-23 Base unit and enclosure of TS3310

The TS3310 supports either the standard or WORM LTO data cartridge and continued support for encryption of data with LTO-4 and LTO-5 tape drives. IBM Tivoli Key Lifecycle Manager (TKLM) is required for encryption key management with Ultrium 5 drives.

Product highlights:

- Modular, scalable tape library designed to grow up to the capacity of 409 LTO cartridges and 18 LTO-5, LTO-4, and LTO3 tape drives
- Form factor from 5U (base module) up to 41U (base module plus 4 expansion units) in standard 19-inch rack mountable or standalone configuration
- Intuitive, user-friendly web-based remote management

- Hot-swap tape drives and power supplies
- Available datapath and control path failover for redundant host connectivity
- Support for a wide range of systems including IBM System p, System x, System i, AS/400®, RS/6000®, Intel, Hewlett-Packard (HP), or Sun

12.4.4 Enterprise tape libraries

The enterprise tape libraries provide storage solutions for large, unattended storage operations from today's midrange up to the enterprise (z/OS and Open Systems) environments. We discuss the following model of the enterprise tape automation product:

- IBM System Storage TS3500 Tape Library

The IBM TS3500 is available as an universal, effective tape solution for either Open Systems or System z environments, and as this is the case, IBM no longer offers:

- IBM TotalStorage 3494 Tape Library

Comprehensive information about the TS3500, including deployment in a z/OS environment, is available in *IBM TS3500 Tape Library with System z Attachment A Practical Guide to Enterprise Tape Drives and TS3500 Tape Automation*, SG24-6789, and a detailed product description is available at:

http://www.ibm.com/systems/storage/tape/ts3500/index.html

IBM System Storage TS3500

Combining reliable, automated tape handling and storage with reliable, high-performance IBM LTO Ultrium and 3592 tape drives, the TS3500 Tape Library offers outstanding retrieval performance with typical cartridge move times of less than three seconds.

The TS3500 Tape Library can be partitioned into multiple logical libraries. This feature makes it an excellent choice for consolidating tape workloads from multiple heterogeneous Open Systems servers. It also enables support for System z attachment in the same library.

In addition, the TS3500 Tape Library provides outstanding reliability and redundancy, through the provision of redundant power supplies in each frame, an optional second cartridge accessor, service bays for nondisruptive maintenance, control and data path failover, and dual grippers within each cartridge accessor. Both library and drive firmware can now be upgraded nondisruptively, that is, without interrupting the normal operations of the library.

The IBM TS3500 Tape Library (Machine Type 3584) is a modular, highly scalable tape library that consists of frames that house tape drives and cartridge storage slots. You can install a single-frame base library (Figure 12-24 on page 270) and expand it to 16 frames, tailoring the library to match your system capacity requirements.

Up to 12 IBM Ultrium tape drives can be installed in a single frame. All generations of LTO tape drives are supported by TS3500. As the enterprise tape libraries are being implemented in the most complex datacenters and client backup solutions utilizing SAN, the majority of tape drive attachment are made based on FC interface (up to 8 Gbps with Ultrium LTO-5).

Figure 12-24 Base frame of TS3500

Benefits of IBM TS3500 in enterprise SAN environments include:

- One base frame, and up to 15 expansion frames per library; up to 15 libraries interconnected per complex
- Up to 12 FC attached drives per frame (192 per library, 2,700 per complex)
- IBM 3592 JA/JJ/JB/JC and JW/JR/JX/JY (WORM) cartridges or IBM LTO Ultrium 5, 4, 3, 2, 1 cartridges
- Up to 60 PB compressed with IBM Ultrium 5 cartridges per library, up to 900 PB compressed per complex,
- Up to 180 PB compressed with 3592 extended capacity cartridges per library, up to 2.7 EB compressed per complex
- Remote management using a web browser
- Advanced Library Management System (ALMS), SNMP functionality, multipath architecture, Persistent World Wide Name

12.5 Storage virtualization and Cloud Computing

In many IT departments, increased user demand has led to haphazard storage growth, resulting in sprawling, heterogeneous storage environments. These environments make it difficult to achieve optimal utilization and to provision storage capacity for new users and applications. Storage virtualization can put an end to these problems. It enables companies to logically aggregate disk storage so capacity can be efficiently allocated across applications and users.

Virtualization solutions help take the cost and complexity out of IT infrastructures. In this section we discuss two important topics, disk and tape virtualization, and what IBM products participate in this process. Finally, we briefly explain what benefits bring IBM storage products to the Cloud computing and which of them are the key building blocks of the *IBM Smart Business Storage Cloud*.

12.5.1 Disk storage virtualization

IBM disk storage virtualization products provide simplified and centralized management of clients' medium to enterprise storage infrastructures consisting of different disk systems, even from different vendors. In this section we discuss the following IBM disk systems that enable storage virtualization:

- IBM System Storage SAN Volume Controller (SVC)
- IBM XIV Storage System
- IBM Storwize V7000 Unified

For more details about each of these products, visit:

http://www.ibm.com/systems/storage/virtualization

IBM System Storage SAN Volume Controller

SANs enable companies to share homogeneous storage resources across the enterprise. But for many clients, information resources are spread over a variety of locations and storage environments, often with products from different vendors, who supply everything from mainframes to laptops. To achieve higher utilization of resources, clients now need to share their storage resources from all their environments, regardless of the vendor. IBM System Storage SAN Volume Controller (SVC) contributes towards this goal of a solution that can help strengthen existing SANs by increasing storage capacity, efficiency, uptime, administrator productivity, and functionality.

IBM System Storage SAN Volume Controller software is delivered preinstalled on SVC Storage Engines so it is quickly ready for implementation once the engines are attached to your SAN. SVC Storage Engines are based on proven IBM System x server technology and are always deployed in redundant pairs (see Figure 12-25), which are designed to deliver high availability.

Figure 12-25 IBM SAN Volume Controller in a clustered pair

SVC is designed to take control of existing storage, retaining all your existing information. This ability helps speed and simplify implementation while helping to minimize the need for additional storage. Once the SVC is implemented, you can make changes to the configuration quickly and easily as needed.

SVC is designed to support nondisruptive data migration between storage systems. In addition, SVC helps make storage potentially available to all attached servers, greatly increasingly the flexibility for using for example VMware vMotion. Without SVC, use of vMotion could be limited by storage being dedicated to specific servers.

Because SVC appears to servers as a single type of storage, virtual server provisioning is also simplified because only a single driver type is needed in server images, which also simplifies administration of those server images. Similarly, SVC eases replacing storage or moving data from one storage type to another because these changes do not require changes to server images. Without SVC, changes of storage type could require disruptive changes to server images.

The SVC is based on IBM System x3550 with Intel Xeon 5600 2.5 GHz processor and 24 GB of memory cache. It incorporates four 8 Gbps FC ports, two 1 Gbps (optionally additional two 10 Gbps) iSCSI ports. It contains redundant power supplies in standard 19-inch rack-mountable enclosure.

Briefly summarized, the IBM System Storage SAN Volume Controller (SVC) is designed to:

- Combine storage capacity from multiple vendors for centralized management and increase storage utilization by providing more flexible access to storage assets
- Improve administrator productivity by enabling management of pooled storage from a single interface
- Reduce downtime by insulating host applications from changes to the physical storage infrastructure
- Enable a tiered storage environment to match the cost of storage to the value of data
- Support data migration among storage systems without interruption to applications
- Supports consolidated disaster recovery site servicing more than one production location

IBM XIV Storage System

IBM XIV is a proven, high-end disk storage, designed for growth with unmatched ease of use. IBM XIV eliminates the complexity of managing enterprise storage. It never compromises performance for reliability, providing consistent high performance without manual tuning for even the most demanding application workloads, while keeping TCO incredibly low. Its grid architecture delivers virtual storage that optimizes performance in virtualized environments

and integrates seamlessly with cloud technologies to provide the agility clients need to handle growth.

IBM XIV Storage System is offered in two models, both based on the same proven XIV architecture and all-inclusive pricing approach:

- **IBM XIV** features powerful storage for most application needs, with excellent price-to-performance advantages. This model offers customer-acclaimed value in handling a mix of diverse workloads at low TCO. It incorporates 1 or 2 TB SATA disk drives up to 161 TB usable capacity (15 modules).
- **IBM XIV Gen3** (Figure 12-26) features state-of-the-art hardware that can do more, and do more faster. Rely on it for your ultra-demanding performance objectives, including business intelligence, archiving, large email setups, data warehousing, and OLTP workloads, as well as ever-changing virtualized and cloud environments. 2 TB SAS attached disk drives are used.

Figure 12-26 Third generation of IBM XIV Storage System

Several architectural features contribute to the XIV system's unique performance profile:

- **Massive parallelism in a fully distributed architecture** - XIV is based on a distributed architecture of interconnected modules including multicore processors, large cache, and high-density disk drives
- **Distributed data** - the system stores data by breaking it down into 1 megabyte chunks called partitions, each mirrored for redundancy to another module.
- **Distributed bandwidth within modules** - aggressive prefetching is enabled by the large cache-to-disk bandwidth available within each module
- **Load balancing** - the system distributes the application load across all system modules uniformly
- **High performance during disk rebuild** - XIV rebuilds failed disk at unprecedented speed, due to a distributed rebuild mechanism that engages all disks in the system in the rebuild process

IBM XIV integrates easily with virtualization, email, database, analytics and data protection solutions from Microsoft, IBM, SAP, Oracle, SAS, VMware, Hyper-V, and Symantec. The XIV

Gen3 model gives applications a tremendous performance boost, helping clients meet increasing demands with fewer servers and networks. The XIV series plays a key role in IBM end-to-end dynamic infrastructure solutions, integrating seamlessly with IBM ProtecTIER, Scale Out Network Attached Storage (SONAS), SAN Volume Controller, Storwize V7000 and Tivoli products.

IBM Storwize V7000 Unified

IBM Storwize V7000 Unified is a virtualized storage system to complement virtualized server environments that provides unmatched performance, availability, advanced functions, and highly scalable capacity never seen before in midrange disk systems. Storwize V7000 Unified is a powerful midrange disk system that has been designed to be easy to use and enable rapid deployment without additional resources. Figure 12-27 shows the rack installation of both types of Storwize V7000 Unified bays - twelve 3.5-inch disks and twenty-four 2.5-inch disk drives.

Figure 12-27 IBM Storwize V7000 and its rack-mountable disk bays

Storwize V7000 Unified consolidates block and file workloads into a single storage system for simplicity of management and reduced cost. It offers greater efficiency and flexibility through built-in solid state drive (SSD) optimization and thin provisioning technologies. Storwize V7000 Unified advanced functions also enable non-disruptive migration of data from existing storage, simplifying implementation and minimizing disruption to users. In addition, it enables you to virtualize and reuse existing disk systems, supporting a greater potential return on investment (ROI). The three key functions help to provide a single point of control to support improved storage efficiency:

- **Consolidation** of storage resources by efficient scaling improves productivity and reduces cost
- **Virtualization** of storage infrastructure can optimize expenditures, resources, and capabilities. New support for VMware vStorage APIs enables Storwize V7000 to take on some storage-related tasks that were previously performed by VMware, which helps improve efficiency and frees up server resources for other more mission-critical tasks
- **Tiering** optimizes storage by enabling data to be located in away that can improve system performance, reduce costs, and simplify information management. Using IBM System Storage Easy Tier technology, Storwize V7000 can utilize SSDs confidently, effectively,

and economically by automatically and dynamically moving only the appropriate data to the SSDs in the system, based on performance monitoring.

IBM Storwize V7000 product highlights:

- Control enclosure supports up to 240 TB of physical capacity using 12 standard 2U expansion units with SAS attached disk drives.
- Two control enclosures can be clustered for high availability and optimal performance.
- Host attachments using eights 8 Gbps, four 1 Gbps or optionally 10 Gbps iSCSI ports.
- File module provides attachment to 1 Gbps and 10 Gbps network-attached storage (NAS) environments.
- Incorporated solid state drives (SSD) support business applications that need to grow dynamically with effective space utilization (thin provisioning). Easy Tier automatically migrates frequently used data to SSDs.
- Support of IBM Advanced Copy Services such as IBM FlashCopy, Metro Mirror and Global Mirror. IBM Tivoli FlashCopy Manager shortens backup and recovery times.

12.5.2 Tape storage virtualization

In order to maintain continuous business operations, address regulatory requirements and archive business records, clients need an infrastructure that enables them to manage their data from online application storage to offline, permanent archive media. Tape is a key part of both the backup and archive life cycle. Tape still provides the lowest total cost of ownership alternative for securely storing long-term archives for record keeping and disaster recovery. As data centers and data stores grow, tape operations can become more complex. This growth can lead to increased backup and restore times and higher management overhead and costs. Tape virtualization solutions can help to get over these constrains. IBM offers the following products:

- IBM System Storage TS7610 ProtecTIER Deduplication Appliance Express
- IBM System Storage TS7650 ProtecTIER Deduplication Appliance
- IBM System Storage TS7650G ProtecTIER Deduplication Gateway
- IBM System Storage TS7680 ProtecTIER Deduplication Gateway for System z
- IBM Virtualization Engine TS7700 Family

IBM has withdrawn these tape virtualization products from marketing:

- IBM Virtualization Engine TS7520
- IBM Virtualization Engine TS7530
- IBM Virtual Tape Server 3494 - B10
- IBM Virtual Tape Server 3494 - B20

Additional technical information and product details are available in *Implementing IBM Storage Data Deduplication Solutions*, SG24-7888 and *TS7680 Deduplication ProtecTIER Gateway for System z*, SG24-7796. A product summary is available at:

http://www.ibm.com/systems/storage/tape/virtualization/index.html

IBM ProtecTIER Deduplication Appliances

IBM ProtecTIER Deduplication solutions, featuring revolutionary and patented HyperFactor data deduplication technology, provide enterprise class performance, scalability, and proven enterprise level data integrity to meet the disk based data protection needs of the enterprise data center down to mid-market environments while enabling significant infrastructure cost reductions.

The IBM System Storage TS7610 ProtecTIER Deduplication Appliance Express provides fast, reliable and easy-to-deploy backup and recovery for midsize IT environments. This solution has a preconfigured repository and can be configured with either a Virtual Tape Library (VTL) or Symantec OpenStorage (OST) interface. Available in two configuration options (4 & 5.4TBs), the TS7610 provides capacity, price/performance and RAS features required by midsize customers.

The IBM System Storage TS7650 ProtecTIER Deduplication Appliance is designed to improve backup and recovery operations and is available in four preconfigured solutions ranging from 7TBs, up to 36TB in a cluster. This integrated solution makes it easy to harness the power of deduplication without making radical changes to the existing environment. Again, TS7650 has a preconfigured repository and can be deployed with either VTL or OST.

Figure 12-28 presents both models of IBM ProtecTIER Deduplication Appliances.

Figure 12-28 IBM ProtecTIER Deduplication Appliances - TS7650 and TS7610 Express

Products benefits:

- Improve backup and recovery performance with high-speed disk-based data protection, that enables more efficient, reliable storage of valuable data, while optimizing storage infrastructure and reducing TCO
- Simple to install, manage and maintain data protection for midsize IT environments
- Virtual Tape Library (VTL) or Symantec OpenStorage (OST) interface support
- Up to 100/500 MBps or more inline data deduplication performance
- Up to 25 times or more storage capacity reduction
- Emulation of up to 4/12 virtual libraries, 64/256 virtual drives and 8192/128000 virtual cartridges
- Capacity to store up to 135/900TBs or more backup data on single 5.4/36 TB appliance

IBM ProtecTIER Deduplication Gateways

While IBM ProtecTIER Deduplication Appliances offers complete preconfigured deduplication solutions with installed storage capacity for medium and enterprise datacenters, IBM ProtecTIER Deduplication Gateways require additional SAN attached storage system to hold and protect business critical and valuable backup data. IBM introduces two deduplication

products, one for Open Systems - IBM TS7650G ProtecTIER Deduplication Gateway and one dedicated for mainframe environment - IBM TS7680 ProtecTIER Deduplication Gateway for System z.

Both models of deduplication gateways offer high-performance inline data deduplication, high-availability clustering with Global Deduplication, and flexibility to support up to 1 petabyte (PB) of physical storage capacity on both IBM and non-IBM supported storage systems. The Figure 12-29 shows front panel of both, identically looking IBM ProtecTIER Deduplication Gateways.

Figure 12-29 IBM TS7650G ProtecTIER Deduplication Gateway

Benefits and highlights of ProtecTIER gateways:

- High-speed backups with up to 2000 MBps (7.2 TB/hr.) or more sustained inline deduplication backup performance and up to 2800 (10 TB/hr.) MBps or more sustained recovery performance
- Easily scalable to provide up to 25 PB backup storage capacity, up to 512 virtual tape drives and 1 million logical volumes per two node cluster
- Non-hash-based approach avoids the possibility of data loss due to a hash collision
- ProtecTIER Native Replication leverages deduplication technology in the disk repositories at both the primary and secondary sites to lower bandwidth requirements.
- Inline deduplication enables replication to occur concurrently with backup operations increasing responsiveness and ability to restore data quickly when needed.
- User friendly, intuitive GUI-based management and monitoring tools

IBM Virtualization Engine TS7700

The IBM Virtualization Engine TS7700 is a family of mainframe virtual tape solutions that are designed to optimize tape processing. With one solution, the implementation of a fully integrated tiered storage hierarchy of disk and tape takes advantage of the benefits of both technologies to help enhance performance and provide the capacity needed for today's tape processing requirements. Deploying this innovative subsystem can help reduce batch processing time, total cost of ownership and management overhead.

A TS7700 (Figure 12-30) can help improve the efficiency of mainframe tape operations by efficiently using disk storage, tape capacity, and tape speed, and by providing a large number of tape addresses. These benefits help make the TS7700 a suitable repository for local and remote backups, and archival data. Two models are available for purchase.

The TS7720 Virtualization Engine provides high capacity for workloads that are cache friendly due to their rapid recall requirements. The TS7720 features 2 TB SATA disk drives with RAID 6 to allow customers to scale their solution to meet the needs of growing workloads without affecting application availability.

The TS7740 Virtualization Engine supports attachment to and exploits the performance and capacity of the IBM System Storage TS1130 and TS1120 Tape Drives or the IBM TotalStorage 3592 Model J1A Tape Drive installed in an IBM System Storage TS3500 Tape Library or IBM TotalStorage 3494 Tape Library. Support for these tape drives can help to reduce the number of cartridges and the size of the library by allowing the storage of up to 3 TB on a single 3592 JB cartridge, assuming 3:1 compression.

Figure 12-30 IBM TS7700

12.5.3 Storage systems for Cloud Computing

IBM offers three types of cloud solutions, for storage and other services:

- Smart Business on the IBM Cloud are standardized services provided by IBM on a pay-per-use basis
- Smart Business Cloud services are private cloud services, behind clients' firewall, built and/or run by IBM
- Smart Business Systems are purpose built, integrated service delivery platforms.

IBM Smart Business Storage Cloud

IBM Information Infrastructure includes next generation virtualized storage and storage management products that can support the demands of cloud computing. Cloud computing applications are typically deployed in a virtualized environment with a strong security model. The following IBM products are key building blocks of IBM Smart Business Storage Cloud:

- **IBM System Storage SAN Volume Controller** - virtualize IBM and non-IBM storage to enable resource pooling, thin provisioning, and simplified management. See "IBM System Storage SAN Volume Controller" on page 271 for details.
- **IBM XIV Storage System** - automates and virtualize data management, and dramatically simplifies systems management to help tame clients dynamic workloads. The XIV is introduced in "IBM XIV Storage System" on page 272.
- **IBM Scale Out Network Attached Storage** - offers extreme scale-out capability for very large storage infrastructures requiring high availability. SONAS also delivers computing services that make the supporting technology almost invisible. It enables applications and

services to be uncoupled from the underlying infrastructure, enabling businesses to adjust to change quickly. As a result, SONAS can easily integrate with your organization's strategies to develop a more dynamic enterprise. The comprehensive information about SONAS is included in *IBM Scale Out Network Attached Storage: Architecture, Planning, and Implementation Basics*, SG24-7875

- **IBM Tivoli Storage Productivity Center** - manages virtualized storage and generates usage reports. Automates storage performance and capacity management. Detailed guidance through this robust family of Tivoli software products is available in the *IBM Tivoli Storage Productivity Center V4.2 Release Guide*, SG24-7894 and its benefits for SAN infrastructure in the *SAN Storage Performance Management Using TotalStorage Productivity Center*, SG24-7364.
- **Media encryption** - is essential for cloud applications because it provides a strong security model with minimal overhead, and provides a data shredding capability that costs almost nothing to invoke (simply delete the centralized encryption key). Drive-level media encryption and centralized key management is available for midrange and enterprise disk and tape. IBM Tivoli Key Lifecycle Manager is the core IBM application for simple, strong, and centralized encryption key management.

12.6 IP-based networking for SAN environments

Ethernet networking within complex datacenters, especially 10 Gbps Ethernet (10 GbE), provides significant number of benefits to the storage networks:

- Simplifies storage management
- Increases utilization of assets in the storage networks
- Improves flexibility of storage environments to adopt new solutions
- Reduces cost of infrastructure using consolidation of assets
- Improves efficiency of storage and ethernet network

Ethernet protocols such as iSCSI or Fibre Channel over Ethernet (FCoE) enable clients to start thinking about migration from typical SAN Fibre Channel networking to consolidated storage and ethernet networking on single network. IBM offers variety of products to support and enable converged networking for flexibility and efficiency of assets' utilization in complex datacenters or just to connect remote offices with all the benefits of SAN (Data Center Bridging).

In the following text we briefly discuss two types of IBM products:

- **Hardware offerings** include IBM devices that provide a consolidated networking solution utilizing Converged Network Adapters, which transport both SAN and ethernet LAN data
- **Software solutions** use IBM Virtual Fabric emulation on given Emulex network adapters and specific IBM BNT® products

We do not discuss IBM's range of Ethernet products in this book. For a full description of IBM's Ethernet approach refer to:

IBM System Networking RackSwitch

http://www.ibm.com/systems/networking/switches/rack.html

IBM System Networking BladeCenter

http://www-03.ibm.com/systems/networking/switches/bladecenter.html

IBM Distributed Virtual Switch

http://www-03.ibm.com/systems/networking/switches/virtual/index.html

12.7 Hardware solutions for network convergence

In this section we introduce products that offer converged SAN and LAN networking by utilization of Converged Network Adapters. They include:

- Cisco Nexus 5000 for IBM System Storage

The following device is withdrawn from marketing:

- IBM Converged Switch B32

Cisco Nexus 5000 for IBM System Storage

Cisco Nexus 5000 switches for IBM System Storage are designed for data center environments with technology that supports consistent low latency Ethernet solutions, with front to back cooling, and with network ports in the rear, bringing switching into close proximity with servers and making cable runs short and simple. The switch family is designed to be highly serviceable, with optional redundant, hot-pluggable power supplies and fan modules. It uses datacenter class Cisco NX-OS to support high reliability and ease of management.

The switch family, using cut-through architecture, supports line-rate 10 Gigabit Ethernet on all ports while maintaining consistent low latency independent of packet size and services enabled. The product family supports IEEE Data Center Bridging and Converged Enhanced Ethernet (CEE) capabilities that can help increase the reliability, efficiency, and scalability of Ethernet networks. These features allow the switch to support multiple traffic classes over a Ethernet fabric, thus enabling consolidation of LAN, SAN, and cluster environments. Its ability to connect Fibre Channel over Ethernet (FCoE) to native Fibre Channel protects existing storage system investments while dramatically simplifying in-rack cabling.

In addition to supporting standard 10 Gigabit Ethernet network interface cards (NICs) on servers, the Cisco Nexus 5000 switches (Figure 12-31) integrate with multifunction adapters called converged network adapters (CNAs) that combine the functions of Ethernet NICs and Fibre Channel host bus adapters (HBAs), making the transition to a single, unified network fabric consistent with existing practices, management software, and operating system drivers. The switch family is compatible with integrated transceivers and Twinax cabling solutions to help deliver cost-effective connectivity for 10 Gigabit Ethernet to the servers at the rack level, reducing or eliminating the need for expensive optical transceivers.

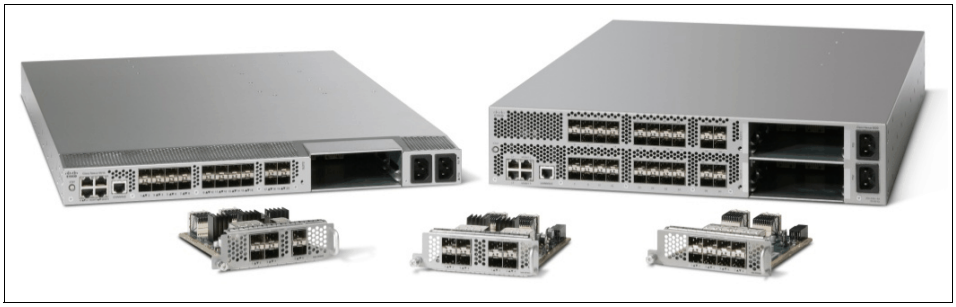

Figure 12-31 Cisco Nexus 5000 switches for IBM System Storage

Product highlights include:

- Designed as a 1U (Cisco Nexus 5010 28 port switch, Cisco Nexus 5548P and 5548UP with up to 48 ports) and as a 2U (Cisco Nexus 5020 56 port switch and Cisco Nexus 5596UP with up to 96 ports) 19-inch rack mountable or standalone enclosure
- Expansion modules include eight 1, 2, 4, 8 Gbps FC ports; four 10 GbE and four 1, 2, 4, 8 Gbps FC ports; six 10 GbE ports
- 10 GbE ports are capable of transporting both storage and LAN traffic eliminating the need for separate server SAN and LAN adapters and cables
- Consistent management is provided through consistency of both Cisco NX-OS Software and Cisco MDS 9000 SAN-OS Software management models and tools
- IEEE Data Center Bridging features for lossless transmission, priority flow control and enhanced transmission selection
- Enterprise-class availability features such as hot-swappable, field replaceable, redundant power supplies, redundant fan modules and port expansion modules

Additional technical details can be found at:

http://www.ibm.com/systems/networking/hardware/ethernet/c-type/nexus/

12.7.1 IBM Virtual Fabric solution

IBM Virtual Fabric solution for System z utilizes IBM BNT *convergence ready* products and specific Emulex Virtual Fabric Adapters. These products are not classified as typical hardware convergence solutions as we described in previous section 12.7, “Hardware solutions for network convergence” on page 280.

This new and innovative solution based on Emulex adapters and IBM BNT Rack Switch products is different than other vNIC solutions in the fact that it carves up dedicated pipes between the adapter and the switch. This solution is built on industry standards providing maximum performance in both directions while allowing for pipes to be allocated at any speed from 1 GbE to 10 GbE.

Leveraging a single 10 GbE dual-port adapter and creating virtual pipes to fewer upstream switch ports help drive out cost and complexity in their IT infrastructure by requiring up to 75% fewer adapters, cables and upstream switch ports. It is also important to note this can be leveraged across multiple application environments not just virtualization. When compared to leveraging multiple 1 GbE ports, clients could see:

- Potential of over 40% acquisition cost savings
- Up to 75% reduction in power consumption
- Significantly simpler management with less cabling and fewer components to manage
- Easy integration into existing clients setups (virtual or non-virtual)

In the following sections we briefly describe IBM BNT rack switches that participate in IBM Virtual Fabric solution. Additional details about this offering can be found at:

http://www.ibm.com/systems/x/options/networking/virtualfabric/

IBM BNT RackSwitch G8124

The IBM BNT RackSwitch G8124 is a 10 Gigabit Ethernet (GbE) switch specifically designed for the datacenters, providing a virtual, cooler and easier network solution. The G8124 offers 24 10 GbE ports in a high density, 1U footprint. Designed with top performance in mind, the RackSwitch G8124 provides line-rate, high-bandwidth switching, filtering and traffic queuing without delaying data and large data-center grade buffers to keep traffic moving.

The G8124 is virtual, providing rack-level virtualization of networking interfaces. VMready® software enables movement of virtual machines, providing matching movement of VLAN assignments, ACLs and other networking and security settings. VMready works with all leading VM providers, such as VMware, Citrix, Xen and Microsoft. The G8124 also supports Virtual Fabric, which allows for the carving up of a physical NIC into 2 - 8 virtual NICs (vNICs) and creates a virtual pipe between the adapter and the switch for improved performance, availability and security, while reducing cost and complexity.

Low latency offered by the G8124 makes it ideal for latency sensitive applications, such as high-performance computing clusters and financial applications. The G8124 also supports the newest protocols including Converged Enhanced Ethernet (CEE) and Data Center Bridging for support of Fibre Channel Over Ethernet (FCoE) and can be leveraged for NAS or iSCSI.

IBM BNT RackSwitch G8124 is shown in Figure 12-32.

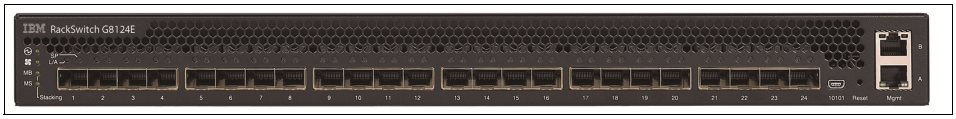

Figure 12-32 Front view of IBM BNT RackSwitch G8124

Product highlights and benefits:

- Optimal for high-performance computing and applications requiring high bandwidth and low latency
- All ports are non-blocking 10 Gigabit Ethernet with deterministic latency of 680 nanoseconds
- VMready helps reduce configuration complexity and improves security levels in virtualized environments
- Virtual Fabric capability allows for the carving up of a physical NIC into multiple virtual NICs (with Emulex adapters)
- Variable speed fans automatically adjust as needed, helping to reduce energy consumption
- Seamless, standards-based integration into existing Cisco and other networks helps reduce downtime and learning curve

IBM BNT RackSwitch G8264

IBM BNT RackSwitch G8264 is a 10 and 40 GbE switch specifically designed for the data center, providing speed, intelligence and interoperability on a proven platforms.

The RackSwitch G8264 offers up to 64x10 GbE and up to four 40 GbE ports - 1.28 Tbps in a 1U footprint. Designed with top performance in mind, the RackSwitch G8264 provides line-rate, high-bandwidth switching, filtering, and traffic queuing without delaying data. Large datacenter grade buffers keep traffic moving. Redundant power and fans along with numerous high availability features enable the RackSwitch G8264 to be available for business sensitive traffic.

The low latency offered by the G8264 (Figure 12-33) makes it ideal for latency sensitive applications such as high performance computing clusters and financial applications. The G8264 supports the newest protocols - including Data Center Bridging/Converged Enhanced Ethernet (DCB/CEE) for support of Fibre Channel over Ethernet (FCoE).

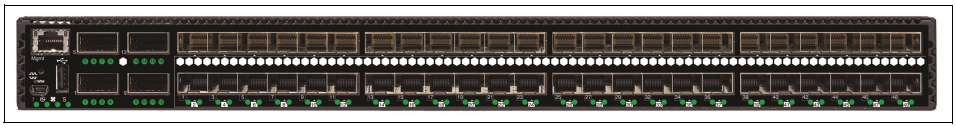

Figure 12-33 Front view of IBM BNT RackSwitch G8264

Product and business benefits:

- Optimized for High Performance Computing and other applications requiring high bandwidth and low latency
- Software is based on Internet standards for optimal interoperability with Cisco or other vendors' networks
- VMready and Virtual Fabric for virtualized networks (with dedicated Emulex adapters for IBM System x)
- Forty-eight 10 GbE SFP+ ports, four 40 GbE QSFP+ ports, up to sixty-four 10 GbE SFP+ ports with optional breakout cables
- Hot-swappable redundant power supplies and fans

12.8 IBM Flex System Networking

IBM Flex System™ offers intelligent, integrated and flexible network architecture that can fit with your existing or future environment. These high performance Ethernet offerings coupled with on demand scalability offer an easy way to scale as IT requirements grow. IBM Flex System Fabric is:

- Integrated – helps manage discrete aspects of the data center as an integrated system through the built in management appliance.
- Optimized – high performance scalable offerings with available 1Gb, 10Gb and 40Gb uplinks allow easy integration with existing network. Simple and cost effective scalability for future growth.
- Automated – automate provisioning and setup of both physical and virtual network.

To meet today's complex and ever-changing business demands, you need a solid foundation of server, storage, networking and management resources that is imple to deploy yet can quickly and automatically adapt to changing conditions. You also need access to—and the ability to take advantage of—broad expertise and proven best practices in systems management, applications, hardware maintenance and more. The IBM PureFlex™ System combines advanced IBM hardware and software along with patterns of expertise and integrate them into optimized solutions that are easy to deploy.

The network resources in a IBM PureFlex System are tightly integrated into the system to support virtualization and simple, integrated management. You can move from managing a physical network to managing a logical network in a virtualized environment—supporting business services instead of network components. With integrated management tools based on open standards, these resources are easy to provision and deploy so you can reduce the cost of managing your virtual fabric. You have fewer elements to manage, but still get port and bandwidth flexibility with highly scalable switches. With scalable components, you can buy a base product and purchase and enable additional ports without adding new hardware.

12.8.1 IBM Flex System Fabric EN4093 10Gb Scalable Switch

The IBM Flex System Fabric EN4093 10Gb Scalable Switch is a 10Gb 64-port upgradable midrange to high-end switch module, offering Layer 2/3 switching designed to install within the I/O module bays of the Enterprise Chassis. The switch has:

- Up to 42 internal 10Gb ports
- Up to 14 external 10Gb uplink ports (SFP+ connectors)
- Up to 2 external 40Gb uplink ports (QSFP+ connectors)

The switch is considered particularly suited for clients who:

- Will be building a 10Gb infrastructure
- Implementing a virtualized environment
- Are requiring investment protection for 40Gb uplinks
- Wish to reduce TCO, improve performance, while maintaining high levels of availability and security
- Wish to avoid oversubscription (Traffic from multiple internal ports attempting to pass through a lower quantity of external ports, leading to congestion & performance impact)

The EN4093 10Gb Scalable Switch is shown in Figure 12-34.

Figure 12-34 IBM Flex System Fabric EN4093 10Gb Scalable Switch

As listed in Table 12-1, the switch is initially licensed with 14 10Gb internal ports enabled and ten 10Gb external uplink ports enabled. Further ports can be enabled, including the two 40Gb external uplink ports with the Upgrade 1 and Upgrade 2 license options. Upgrade 1 must be applied before Upgrade 2 can be applied.

Table 12-1 IBM Flex System Fabric EN4093 10Gb Scalable Switch part numbers and port upgrades

Part number	Feature code[a]	Product description	Total ports enabled		
			Internal	10Gb uplink	40Gb uplink
49Y4270	A0TB / 3593	IBM Flex System Fabric EN4093 10Gb Scalable Switch ► 10x external 10Gb uplinks ► 14x internal 10Gb ports	14	10	0

Part number	Feature code[a]	Product description	Total ports enabled		
			Internal	10Gb uplink	40Gb uplink
49Y4798	A1EL / 3596	IBM Flex System Fabric EN4093 10Gb Scalable Switch (Upgrade 1) ► Adds 2x external 40Gb uplinks ► Adds 14x internal 10Gb ports	28	10	2
88Y6037	A1EM / 3597	IBM Flex System Fabric EN4093 10Gb Scalable Switch (Upgrade 2) (requires Upgrade 1): ► Adds 4x external 10Gb uplinks ► Add 14x internal 10Gb ports	42	14	2

a. The first feature code listed is for configurations ordered through System x sales channels. The second feature code is for configurations ordered through the IBM Power Systems channel.

The key components on the front of the switch are shown in Figure 12-35.

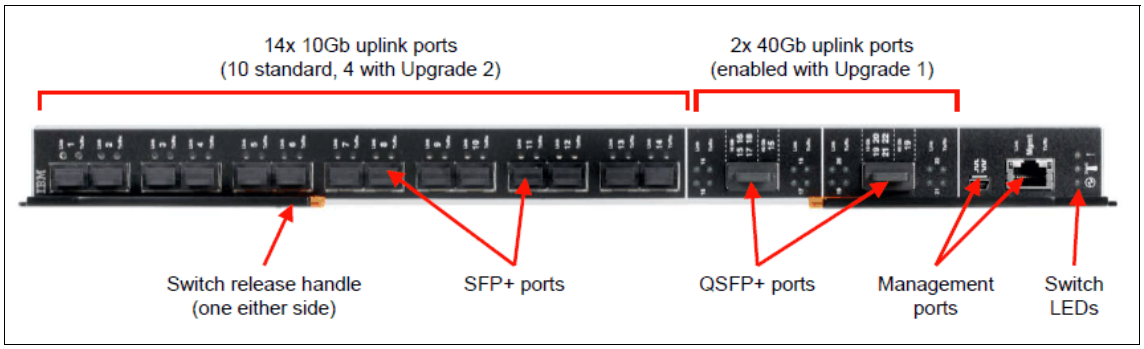

Figure 12-35 IBM Flex System Fabric EN4093 10Gb Scalable Switch

Each upgrade license enables additional internal ports. To take full advantage of those ports, each compute node will need the appropriate I/O adapter installed:

- Base switch requires a two-port Ethernet adapter (one port of the adapter goes to each of two switches)
- Upgrade 1 requires a four-port Ethernet adapter (two ports of the adapter to each switch)
- Upgrade 2 requires a six-port Ethernet adapter (three ports to each switch)

Upgrade 2: Adding Upgrade 2 enables an additional 14 internal ports, delivering the ability to have 42 internal ports, 3 ports connected to each of the 14 compute nodes in the chassis. To take full advantage of all 42 internal ports, a 6-port adapter is required, but this type of adapter is currently not available.

Upgrade 2 still provides a benefit even with a 4-port adapter, since this upgrade enables an extra four external 10Gb uplinks as well.

The rear of the switch has 14 SPF+ module ports and two QSFP+ module ports. The QSFP+ ports can be used to provide either two 40Gb uplinks or eight 10Gb ports, using one of the supported QSFP+ to 4x 10Gb SFP+ cables listed in Table 12-2. This cable splits a single 40Gb QSPFP port into 4 SFP+ 10Gb ports.

For management of the switch there is a mini USB port and also an Ethernet management port provided.

The supported SFP+ and QSFP+ modules and cables for the switch are listed in Table 12-2.

Table 12-2 Supported SFP+ modules and cables

Part number	Feature code[a]	Description
Serial console cables		
90Y9338	A2RR / None	IBM Flex System Management Serial Access Cable Kit
SFP transceivers - 1 GbE		
81Y1618	3268 / EB29	IBM SFP RJ-45 Transceiver (does not support 10/100 Mbps)
81Y1622	3269 / EB2A	IBM SFP SX Transceiver
90Y9424	A1PN / None	IBM SFP LX Transceiver
SFP+ transceivers - 10 GbE		
46C3447	5053 / None	IBM SFP+ SR Transceiver
90Y9412	A1PM / None	IBM SFP+ LR Transceiver
44W4408	4942 / 3382	10GBase-SR SFP+ (MMFiber) transceiver
SFP+ Direct-attach copper (DAC) cables - 10 GbE		
90Y9427	A1PH / ECB4	1m IBM Passive DAC SFP+
90Y9430	A1PJ / ECB5	3m IBM Passive DAC SFP+
90Y9433	A1PK / None	5m IBM Passive DAC SFP+
QSFP+ transceiver and cables - 40 GbE		
49Y7884	A1DR / EB27	IBM QSFP+ 40GBASE-SR Transceiver (Requires either cable 90Y3519 or cable 90Y3521)
90Y3519	A1MM / None	10m IBM MTP Fiber Optical Cable (requires transceiver 49Y7884)
90Y3521	A1MN / None	30m IBM MTP Fiber Optical Cable (requires transceiver 49Y7884)
QSFP+ breakout cables - 40 GbE to 4x10 GbE		
49Y7886	A1DL / EB24	1m 40Gb QSFP+ to 4 x 10Gb SFP+ Cable
49Y7887	A1DM / EB25	3m 40Gb QSFP+ to 4 x 10Gb SFP+ Cable
49Y7888	A1DN / EB26	5m 40Gb QSFP+ to 4 x 10Gb SFP+ Cable
QSFP+ Direct-attach copper (DAC) cables - 40 GbE		
49Y7890	A1DP / None	1m QSFP+ to QSFP+ DAC
49Y7891	A1DQ / None	3m QSFP+ to QSFP+ DAC

a. The first feature code listed is for configurations ordered through System x sales channels. The second feature code is for configurations ordered through the IBM Power Systems channel.

The EN4093 10Gb Scalable Switch has the following features and specifications:

- Internal ports
  - Forty-two internal full-duplex 10 Gigabit ports (Fourteen ports are enabled by default. Optional FoD licenses are required to activate the remaining 28 ports.)

  - Two internal full-duplex 1 GbE ports connected to the chassis management module
- External ports
  - Fourteen ports for 1 Gb or 10 Gb Ethernet SFP+ transceivers (support for 1000BASE-SX, 1000BASE-LX, 1000BASE-T, 10GBASE-SR, or 10GBASE-LR) or SFP+ copper direct-attach cables (DAC). Ten ports are enabled by default. An optional FoD license is required to activate the remaining four ports. SFP+ modules and DAC cables are not included and must be purchased separately.
  - Two ports for 40 Gb Ethernet QSFP+ transceivers or QSFP+ DACs (ports are disabled by default. An optional FoD license is required to activate them). QSFP+ modules and DAC cables are not included and must be purchased separately.
  - One RS-232 serial port (mini-USB connector) that provides an additional means to configure the switch module.
- Scalability and performance
  - 40 Gb Ethernet ports for extreme uplink bandwidth and performance
  - Fixed-speed external 10 Gb Ethernet ports to leverage 10 Gb core infrastructure
  - Autosensing 10/1000/1000 external Gigabit Ethernet ports for bandwidth optimization
  - Non-blocking architecture with wire-speed forwarding of traffic and aggregated throughput of 1.28 Tbps
  - Media access control (MAC) address learning: automatic update, support of up to 128,000 MAC addresses
  - Up to 128 IP interfaces per switch
  - Static and LACP (IEEE 802.3ad) link aggregation, up to 220 Gb of total uplink bandwidth per switch, up to 64 trunk groups, up to 16 ports per group
  - Support for jumbo frames (up to 9,216 bytes)
  - Broadcast/multicast storm control
  - IGMP snooping to limit flooding of IP multicast traffic
  - IGMP filtering to control multicast traffic for hosts participating in multicast groups
  - Configurable traffic distribution schemes over trunk links based on source/destination IP or MAC addresses or both
  - Fast port forwarding and fast uplink convergence for rapid STP convergence
- Availability and redundancy
  - Virtual Router Redundancy Protocol (VRRP) for Layer 3 router redundancy
  - IEEE 802.1D STP for providing L2 redundancy
  - IEEE 802.1s Multiple STP (MSTP) for topology optimization, up to 32 STP instances are supported by single switch
  - IEEE 802.1w Rapid STP (RSTP) provides rapid STP convergence for critical delay-sensitive traffic like voice or video
  - Per-VLAN Rapid STP (PVRST) enhancements
  - Layer 2 Trunk Failover to support active/standby configurations of network adapter teaming on compute nodes
  - Hot Links provides basic link redundancy with fast recovery for network topologies that require Spanning Tree to be turned off

- VLAN support
  - Up to 1024 VLANs supported per switch, with VLAN numbers ranging from 1 to 4095 (4095 is used for management module's connection only.)
  - 802.1Q VLAN tagging support on all ports
  - Private VLANs
- Security
  - VLAN-based, MAC-based, and IP-based ACLs
  - 802.1x port-based authentication
  - Multiple user IDs and passwords
  - User access control
  - Radius, TACACS+ and LDAP authentication and authorization
- Quality of Service (QoS)
  - Support for IEEE 802.1p, IP ToS/DSCP, and ACL-based (MAC/IP source and destination addresses, VLANs) traffic classification and processing
  - Traffic shaping and re-marking based on defined policies
  - Eight Weighted Round Robin (WRR) priority queues per port for processing qualified traffic
- IP v4 Layer 3 functions
  - Host management
  - IP forwarding
  - IP filtering with ACLs, up to 896 ACLs supported
  - VRRP for router redundancy
  - Support for up to 128 static routes
  - Routing protocol support (RIP v1, RIP v2, OSPF v2, BGP-4), up to 2048 entries in a routing table
  - Support for DHCP Relay
  - Support for IGMP snooping and IGMP relay
  - Support for Protocol Independent Multicast (PIM) in Sparse Mode (PIM-SM) and Dense Mode (PIM-DM).
- IP v6 Layer 3 functions
  - IPv6 host management (except default switch management IP address)
  - IPv6 forwarding
  - Up to 128 static routes
  - Support for OSPF v3 routing protocol
  - IPv6 filtering with ACLs
- Virtualization
  - Virtual Fabric with vNIC (virtual NICs)
  - 802.1Qbg Edge Virtual Bridging (EVB)
  - VMready

- Converged Enhanced Ethernet
  - Priority-Based Flow Control (PFC) (IEEE 802.1Qbb) extends 802.3x standard flow control to allow the switch to pause traffic based on the 802.1p priority value in each packet's VLAN tag.
  - Enhanced Transmission Selection (ETS) (IEEE 802.1Qaz) provides a method for allocating link bandwidth based on the 802.1p priority value in each packet's VLAN tag.
  - Data Center Bridging Capability Exchange Protocol (DCBX) (IEEE 802.1AB) allows neighboring network devices to exchange information about their capabilities.
- Manageability
  - Simple Network Management Protocol (SNMP V1, V2 and V3)
  - HTTP browser GUI
  - Telnet interface for CLI
  - SSH
  - Serial interface for CLI
  - Scriptable CLI
  - Firmware image update (TFTP and FTP)
  - Network Time Protocol (NTP) for switch clock synchronization
- Monitoring
  - Switch LEDs for external port status and switch module status indication
  - Remote Monitoring (RMON) agent to collect statistics and proactively monitor switch performance
  - Port mirroring for analyzing network traffic passing through switch
  - Change tracking and remote logging with syslog feature
  - Support for sFLOW agent for monitoring traffic in data networks (separate sFLOW analyzer required elsewhere)
  - POST diagnostics

For more information, see the IBM Redbooks Product Guide for the IBM Flex System Fabric EN4093 10Gb Scalable Switch, available from:

http://www.redbooks.ibm.com/abstracts/tips0864.html?Open

12.8.2 IBM Flex System EN4091 10Gb Ethernet Pass-thru

The EN4091 10Gb Ethernet Pass-thru module offers a 1 for 1 connection between a single node bay and an I/O module uplink. It has no management interface and can support both 1 Gb and 10 Gb dual-port adapters installed in the compute nodes. If quad-port adapters are installed in the compute nodes, only the first two ports will have access to the pass-thru module's ports.

The necessary 1 GbE or 10 GbE module (SFP, SFP+ or DAC) must also be installed in the external ports of the pass-thru, to support the desired speed (1 Gb or 10 Gb) and medium (fiber optic or copper) for adapter ports on the compute nodes.

The IBM Flex System EN4091 10Gb Ethernet Pass-thru is shown in Figure 12-36.

Figure 12-36 IBM Flex System EN4091 10Gb Ethernet Pass-thru

The ordering part number and feature codes are listed in Table 12-3.

Table 12-3 EN4091 10Gb Ethernet Pass-thru part number and feature codes

Part number	Feature code[a]	Product Name
88Y6043	A1QV / 3700	IBM Flex System EN4091 10Gb Ethernet Pass-thru

a. The first feature code listed is for configurations ordered through System x sales channels. The second feature code is for configurations ordered through the IBM Power Systems channel.

The EN4091 10Gb Ethernet Pass-thru has the following specifications

- Internal ports
  - 14 internal full-duplex Ethernet ports that can operate at 1 Gb or 10 Gb speeds
- External ports
  - Fourteen ports for 1 Gb or 10 Gb Ethernet SFP+ transceivers (support for 1000BASE-SX, 1000BASE-LX, 1000BASE-T, 10GBASE-SR, or 10GBASE-LR) or SFP+ copper direct-attach cables (DAC). SFP+ modules and DAC cables are not included and must be purchased separately.
- Unmanaged device that has no internal Ethernet management port, however, it is able to provide its vital product data (VPD) to the secure management network in the Chassis Management Module
- Supports 10Gb Ethernet signaling for CEE, FCoE and other Ethernet based transport protocols.
- Allows direct connection from the 10Gb Ethernet adapters installed in compute nodes in a chassis to an externally located top-of-rack (TOR) switch or other external device.

Four-port adapters: The EN4091 10Gb Ethernet Pass-thru has only 14 internal ports. As a result, only two ports on each compute node are enabled, one for each of two pass-thru modules installed in the chassis. If four-port adapters are installed in the compute nodes, ports 3 and 4 on those adapters are not enabled.

There are standard 3 I/O module status LEDs. Each port has link and activity LEDs.

Table 12-4 lists the supported transceivers and direct-attach copper (DAC) cables.

Table 12-4 IBM Flex System EN4091 10Gb Ethernet Pass-thru part numbers and feature codes

Part number	Feature codes[a]	Description
SFP+ transceivers - 10 GbE		
44W4408	4942 / 3282	10GbE 850 nm Fiber SFP+ Transceiver (SR)
46C3447	5053 / None	IBM SFP+ SR Transceiver
90Y9412	A1PM / None	IBM SFP+ LR Transceiver
SFP transceivers - 1 GbE		
81Y1622	3269 / EB2A	IBM SFP SX Transceiver
81Y1618	3268 / EB29	IBM SFP RJ45 Transceiver
90Y9424	A1PN / None	IBM SFP LX Transceiver
Direct-attach copper (DAC) cables		
81Y8295	A18M / EN01	1m 10GE Twinax Act Copper SFP+ DAC (active)
81Y8296	A18N / EN02	3m 10GE Twinax Act Copper SFP+ DAC (active)
81Y8297	A18P / EN03	5m 10GE Twinax Act Copper SFP+ DAC (active)
95Y0323	A25A / None	1m IBM Active DAC SFP+ Cable
95Y0326	A25B / None	3m IBM Active DAC SFP+ Cable
95Y0329	A25C / None	5m IBM Active DAC SFP+ Cable

a. The first feature code listed is for configurations ordered through System x sales channels. The second feature code is for configurations ordered through the IBM Power Systems channel.

For more information, see the IBM Redbooks Product Guide for the IBM Flex System EN4091 10Gb Ethernet Pass-thru, available from:

http://www.redbooks.ibm.com/abstracts/tips0865.html?Open

12.8.3 IBM Flex System EN2092 1Gb Ethernet Scalable Switch

The EN2092 1Gb Ethernet Switch provides support for L2/L3 switching and routing. The switch has:

- Up to 28 internal 1Gb ports
- Up to 20 external 1Gb ports (RJ45 connectors)
- Up to 4 external 10Gb uplink ports (SFP+ connectors)

The switch is shown in Figure 12-37.

Figure 12-37 IBM Flex System EN2092 1Gb Ethernet Scalable Switch

As listed in Table 12-5 on page 292, the switch comes standard with 14 internal and 10 external Gigabit Ethernet ports enabled. Further ports can be enabled, including the four external 10Gb uplink ports. Upgrade 1 and the 10Gb Uplinks upgrade can be applied in either order.

Table 12-5 IBM Flex System EN2092 1Gb Ethernet Scalable Switch part numbers and port upgrades

Part number	Feature code[a]	Product description
49Y4294	A0TF / 3598	IBM Flex System EN2092 1Gb Ethernet Scalable Switch ► 14 internal 1Gb ports ► 10 external 1Gb ports
90Y3562	A1QW / 3594	IBM Flex System EN2092 1Gb Ethernet Scalable Switch (Upgrade 1) ► Adds 14 internal 1Gb ports ► Adds 10 external 1Gb ports
49Y4298	A1EN / 3599	IBM Flex System EN2092 1Gb Ethernet Scalable Switch (10Gb Uplinks) ► Adds 4 external 10Gb uplinks

a. The first feature code listed is for configurations ordered through System x sales channels. The second feature code is for configurations ordered through the IBM Power Systems channel.

The key components on the front of the switch are shown in Figure 12-35.

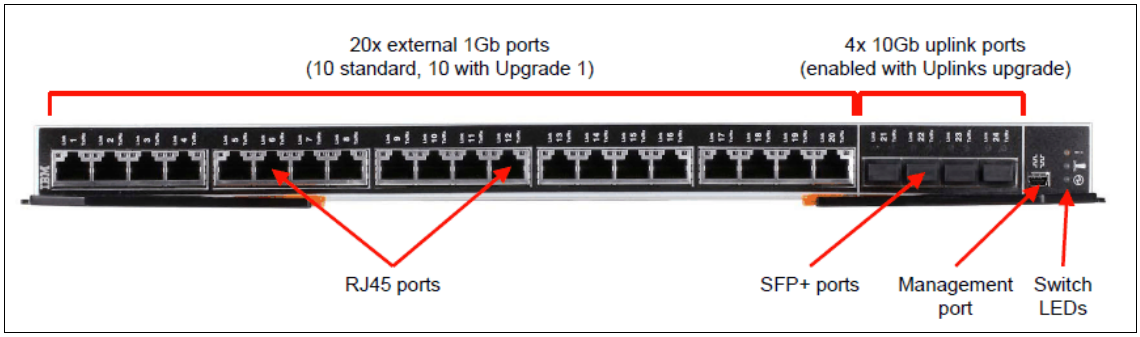

Figure 12-38 IBM Flex System EN2092 1Gb Ethernet Scalable Switch

The standard switch has 14 internal ports and the Upgrade 1 license enables 14 additional internal ports. To take full advantage of those ports, each compute node will need the appropriate I/O adapter installed:

- The base switch requires a two-port Ethernet adapter installed in each compute node (one port of the adapter goes to each of two switches)
- Upgrade 1 requires a four-port Ethernet adapter installed in each compute node (two ports of the adapter to each switch)

The standard has 10 external ports enabled. Additional external ports are enabled with license upgrades:

- Upgrade 1 enables 10 additional ports for a total of 20 ports
- Uplinks Upgrade enables the four 10Gb SFP+ ports.

These two upgrades can be installed in either order.

This switch is considered ideal for clients who are:

- Still use 1Gb as their networking infrastructure
- Deploying virtualization and require multiple 1Gb ports
- Want investment protection for 10Gb uplinks
- Looking to reduce TCO, improve performance, while maintaining high levels of availability and security
- Looking to avoid oversubscription (multiple internal ports attempting to pass through a lower quantity of external ports, leading to congestion or performance impact).

The switch has three switch status LEDs and one mini-USB serial port connector for console management.

Uplink Ports 1 to 20 are RJ45 and the 4 x 10Gb uplink ports are SFP+. The switch supports either SFP+ modules or DAC cables. The supported SFP+ modules and DAC cables for the switch are listed in Table 12-6.

Table 12-6 IBM Flex System EN2092 1Gb Ethernet Scalable Switch SFP+ and DAC cables

Part number	Feature code[a]	Description
SFP transceivers		
81Y1622	3269 / EB2A	IBM SFP SX Transceiver
81Y1618	3268 / EB29	IBM SFP RJ45 Transceiver
90Y9424	A1PN / None	IBM SFP LX Transceiver
SFP+ transceivers		
44W4408	4942 / 3282	10GbE 850 nm Fiber SFP+ Transceiver (SR)
46C3447	5053 / None	IBM SFP+ SR Transceiver
90Y9412	A1PM / None	IBM SFP+ LR Transceiver
DAC cables		
90Y9427	A1PH / None	1m IBM Passive DAC SFP+
90Y9430	A1PJ / ECB5	3m IBM Passive DAC SFP+
90Y9433	A1PK / None	5m IBM Passive DAC SFP+

a. The first feature code listed is for configurations ordered through System x sales channels. The second feature code is for configurations ordered through the IBM Power Systems channel.

The EN2092 1Gb Ethernet Scalable Switch has the following features and specifications:

- Internal ports
  - Twenty-eight internal full-duplex Gigabit ports (14 ports are enabled by default. An optional FoD license is required to activate another 14 ports.)
  - Two internal full-duplex 1 GbE ports connected to the chassis management module
- External ports
  - Four ports for 1 Gb or 10 Gb Ethernet SFP+ transceivers (support for 1000BASE-SX, 1000BASE-LX, 1000BASE-T, 10GBASE-SR, or 10GBASE-LR) or SFP+ copper direct-attach cables (DAC). These ports are disabled by default. An optional FoD license is required to activate them. SFP+ modules are not included and must be purchased separately.
  - Twenty external 10/100/1000 1000BASE-T Gigabit Ethernet ports with RJ-45 connectors (10 ports are enabled by default. An optional FoD license is required to activate another 10 ports).
  - One RS-232 serial port (mini-USB connector) that provides an additional means to configure the switch module.
- Scalability and performance
  - Fixed-speed external 10 Gb Ethernet ports for maximum uplink bandwidth
  - Autosensing 10/1000/1000 external Gigabit Ethernet ports for bandwidth optimization
  - Non-blocking architecture with wire-speed forwarding of traffic
  - Media access control (MAC) address learning: automatic update, support of up to 32,000 MAC addresses
  - Up to 128 IP interfaces per switch
  - Static and LACP (IEEE 802.3ad) link aggregation, up to 60 Gb of total uplink bandwidth per switch, up to 64 trunk groups, up to 16 ports per group
  - Support for jumbo frames (up to 9,216 bytes)
  - Broadcast/multicast storm control
  - IGMP snooping for limit flooding of IP multicast traffic
  - IGMP filtering to control multicast traffic for hosts participating in multicast groups
  - Configurable traffic distribution schemes over trunk links based on source/destination IP or MAC addresses or both
  - Fast port forwarding and fast uplink convergence for rapid STP convergence
- Availability and redundancy
  - Virtual Router Redundancy Protocol (VRRP) for Layer 3 router redundancy
  - IEEE 802.1D STP for providing L2 redundancy
  - IEEE 802.1s Multiple STP (MSTP) for topology optimization, up to 32 STP instances supported by single switch
  - IEEE 802.1w Rapid STP (RSTP) (provides rapid STP convergence for critical delay-sensitive traffic like voice or video)
  - Per-VLAN Rapid STP (PVRST) enhancements

  - Layer 2 Trunk Failover to support active/standby configurations of network adapter teaming on compute nodes
  - Hot Links provides basic link redundancy with fast recovery for network topologies that require Spanning Tree to be turned off
- VLAN support
  - Up to 1024 VLANs supported per switch, with VLAN numbers ranging from 1 to 4095 (4095 is used for management module's connection only)
  - 802.1Q VLAN tagging support on all ports
  - Private VLANs
- Security
  - VLAN-based, MAC-based, and IP-based ACLs
  - 802.1x port-based authentication
  - Multiple user IDs and passwords
  - User access control
  - Radius, TACACS+ and LDAP authentication and authorization
- Quality of Service (QoS)
  - Support for IEEE 802.1p, IP ToS/DSCP, and ACL-based (MAC/IP source and destination addresses, VLANs) traffic classification and processing
  - Traffic shaping and re-marking based on defined policies
  - Eight Weighted Round Robin (WRR) priority queues per port for processing qualified traffic
- IP v4 Layer 3 functions
  - Host management
  - IP forwarding
  - IP filtering with ACLs, up to 896 ACLs supported
  - VRRP for router redundancy
  - Support for up to 128 static routes
  - Routing protocol support (RIP v1, RIP v2, OSPF v2, BGP-4), up to 2048 entries in a routing table
  - Support for DHCP Relay
  - Support for IGMP snooping and IGMP relay
  - Support for Protocol Independent Multicast (PIM) in Sparse Mode (PIM-SM) and Dense Mode (PIM-DM).
- IP v6 Layer 3 functions
  - IPv6 host management (except default switch management IP address)
  - IPv6 forwarding
  - Up to 128 static routes
  - Support for OSPF v3 routing protocol
  - IPv6 filtering with ACLs
- Virtualization
  - VMready

- Manageability
  - Simple Network Management Protocol (SNMP V1, V2, and V3)
  - HTTP browser GUI
  - Telnet interface for CLI
  - SSH
  - Serial interface for CLI
  - Scriptable CLI
  - Firmware image update (TFTP and FTP)
  - Network Time Protocol (NTP) for switch clock synchronization
- Monitoring
  - Switch LEDs for external port status and switch module status indication
  - Remote Monitoring (RMON) agent to collect statistics and proactively monitor switch performance
  - Port mirroring for analyzing network traffic passing through the switch
  - Change tracking and remote logging with the syslog feature
  - Support for the sFLOW agent for monitoring traffic in data networks (separate sFLOW analyzer required elsewhere)
  - POST diagnostics

For more information, see the IBM Redbooks Product Guide for the IBM Flex System EN2092 1Gb Ethernet Scalable Switch, available from:

http://www.redbooks.ibm.com/abstracts/tips0861.html?Open

12.8.4 IBM Flex System FC5022 16Gb SAN Scalable Switch

The IBM Flex System FC5022 16Gb SAN Scalable Switch is a high-density, 48-port 16 Gbps Fibre Channel switch that is used in the Enterprise Chassis. The switch provides 28 internal ports to compute nodes by way of the midplane, and 20 external SFP+ ports. These SAN switch modules deliver an embedded option for IBM Flex System users deploying storage area networks in their enterprise. They offer end-to-end 16 Gb and 8 Gb connectivity.

The N_Port Virtualization mode streamlines the infrastructure by reducing the number of domains to manage while enabling the ability to add or move servers without impact to the SAN. Monitoring is simplified via an integrated management appliance, or clients using end-to-end Brocade SAN can leverage the Brocade management tools.

Figure 12-39 shows the IBM Flex System FC5022 16Gb SAN Scalable Switch.

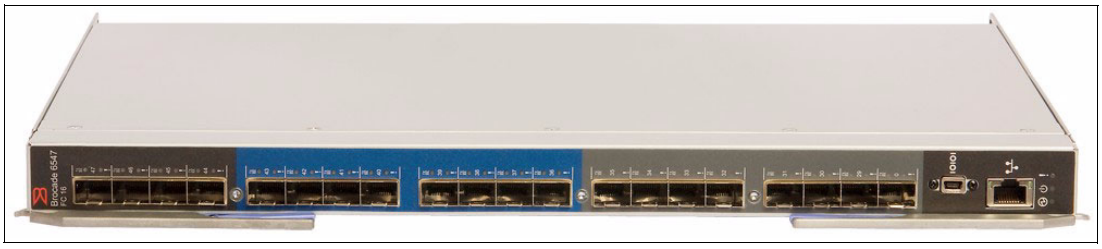

Figure 12-39 IBM Flex System FC5022 16Gb SAN Scalable Switch

Two versions are available as listed in Table 12-7: a 12-port switch module and a 24-port switch with the Enterprise Switch Bundle (ESB) software. The port count can be applied to internal or external ports using a a feature called Dynamic Ports on Demand (DPOD).

Table 12-7 IBM Flex System FC5022 16Gb SAN Scalable Switch part numbers

Part number	Feature codes[a]	Description	Ports enabled
88Y6374	A1EH / 3770	IBM Flex System FC5022 16Gb SAN Scalable Switch	12
90Y9356	A1EJ / 3771	IBM Flex System FC5022 24-port 16Gb ESB SAN Scalable Switch	24

a. The first feature code listed is for configurations ordered through System x sales channels. The second feature code is for configurations ordered through the IBM Power Systems channel.

Table 12-8 provides a feature comparison by model for FC5022 switches

Table 12-8 Feature comparison by model

Feature	FC5022 16Gb ESB Switch (90Y9356)	FC5022 16Gb SAN Scalable Switch (88Y6374)
Number of active ports	24	12
Full fabric	Included	Included
Access Gateway	Included	Included
Advanced zoning	Included	Included
Enhanced Group Management	Included	Included
ISL Trunking	Included	Not available
Adaptive Networking	Included	Not available
Advanced Performance Monitoring	Included	Not available
Fabric Watch	Included	Not available
Extended Fabrics	Included	Not available
Server Application Optimization	Included	Not available

With Dynamic Ports on Demand (DPOD), ports are licensed as they come online. With the FC5022 16Gb SAN Scalable Switch, the first 12 ports reporting (on a first-come, first-served basis) on boot-up are assigned licenses. These 12 ports may be any combination of external or internal Fibre Channel (FC) ports. After all licenses have been assigned, you can manually move those licenses from one port to another. Because this is dynamic, no defined ports are reserved except ports 0 and 29. The FC5022 16Gb ESB Switch has the same behavior, the only difference is the number of ports.

The part number for the switch includes the following items:

- One IBM Flex System FC5022 16Gb SAN Scalable Switch or IBM Flex System FC5022 24-port 16Gb ESB SAN Scalable Switch
- Important Notices Flyer
- Warranty Flyer
- Documentation CD-ROM

The switch does not include a serial management cable, however IBM Flex System Management Serial Access Cable, 90Y9338, is supported and contains two cables, a mini-USB-to-RJ45 serial cable and a mini-USB-to-DB9 serial cable, either of which can be used to connect to the switch locally for configuration tasks and firmware updates.

Transceivers

The switch comes without SFP+, they must be ordered separately to provide outside connectivity. Table 12-9 lists supported SFP+ options.

Table 12-9 Supported SFP+ transceivers

Part number	Feature code[a]	Description
88Y6416	5084 / 5370	Brocade 8Gb SFP+ SW Optical Transceiver
88Y6393	A22R / 5371	Brocade 16Gb SFP+ Optical Transceiver

a. The first feature code listed is for configurations ordered through System x sales channels. The second feature code is for configurations ordered through the IBM Power Systems channel.

Benefits

The switches offer the following key benefits:

- Exceptional price/performance for growing SAN workloads

  The FC5022 16Gb SAN Scalable Switch delivers exceptional price/performance for growing SAN workloads through a combination of market-leading 1,600 MB/sec throughput per port and an affordable high-density form factor. The 48 FC ports produce an aggregate 768 Gbps full-duplex throughput, plus any external eight ports can be trunked for 128 Gbps Inter-Switch Links (ISLs). As 16 Gbps port technology dramatically reduces the number of ports and associated optics/cabling required through 8/4 Gbps consolidation, the cost savings as well as simplification benefits are substantial.

- Accelerating fabric deployment and serviceability with diagnostic ports

  Diagnostic Ports (D_Ports) are a new port type supported by the FC5022 16Gb SAN Scalable Switch that enables administrators to quickly identify and isolate 16 Gbps optics, port and cable problems, reducing fabric deployment and diagnostic times. If the optical media is found to be the source of the problem, it can be transparently replaced as 16 Gbps optics are hot-pluggable.

- A building block for virtualized, private cloud storage

  The FC5022 16Gb SAN Scalable Switch supports multi-tenancy in cloud environments through VM-aware end-to-end visibility and monitoring, Quality of Service (QoS), and fabric-based advanced zoning features. The FC5022 16Gb SAN Scalable Switch enables secure distance extension to virtual private or hybrid clouds with dark fibre support, as well as in-flight encryption and data compression. Internal fault-tolerant and enterprise-class RAS features help minimize downtime to support mission-critical cloud environments.

- Simplified and optimized interconnect with Brocade Access Gateway

  The FC5022 16Gb SAN Scalable Switch can be deployed as a full-fabric switch or as a Brocade Access Gateway, which simplifies fabric topologies and heterogeneous fabric connectivity. Access Gateway mode utilizes N_Port ID Virtualization (NPIV) switch standards to present physical and virtual servers directly to the core of SAN fabrics. This makes it transparent to the SAN fabric, greatly reducing management of the network edge.

- Maximizing investments

  To help optimize technology investments, IBM offers a single point of serviceability backed by industry-renowned education, support and training. In addition, the IBM 16/8 Gbps SAN Scalable Switch is in the ServerProven® program, enabling compatibility among a

variety of IBM and partner products. IBM recognizes that customers deserve the most innovative, expert integrated systems solutions.

Features and specifications

The FC5022 16Gb SAN Scalable Switches have the following features and specifications:

- Internal ports
  - 28 internal full-duplex 16 Gb FC ports (up to 14 internal ports can be activated with Port-on-Demand feature, remaining ports are reserved for future use)
  - Internal ports operate as F_ports (fabric ports) in native mode or in access gateway mode
  - Two internal full-duplex 1 GbE ports connected to the chassis management module
- External ports
  - 20 external ports for 16 Gb SFP+ or 8 Gb SFP+ transceivers supporting 4 Gb, 8 Gb and 16 Gb port speeds (SFP+ modules are not included and must be purchased separately - see Table 2). Ports are activated with Port-on-Demand feature.
  - External ports can operate as F_ports (fabric ports), FL_ports (fabric loop ports), or E_ports (expansion ports) in native mode or as N_Ports (Node Ports) in access gateway mode
  - One external 1 GbE port (1000BASE-T) with RJ-45 connector for switch configuration and management
  - One RS-232 serial port (mini-USB connector) that provides an additional means to configure the switch module
- Access gateway mode (N_Port ID Virtualization - NPIV) support
- Power-on self-test diagnostics and status reporting
- Inter-Switch Link (ISL) Trunking (licensable), which allows up to eight ports (at 16, 8, or 4 Gbps speeds) to combine to form a single, logical ISL with a speed of up to 128 Gbps (256 Gbps full duplex) for optimal bandwidth utilization, automatic path failover, and load balancing.
- Brocade Fabric OS (FOS), which delivers distributed intelligence throughout the network and enables a wide range of value-added applications, such as Brocade Advanced Web Tools and Brocade Advanced Fabric Services (on certain models).
- Supports up to 768 Gbps I/O bandwidth
- 420 million frames switch per second, 0.7 microseconds latency
- 8,192 buffers for up to 3,750 Km extended distance at 4 Gbps FC (Extended Fabrics license required)
- In-flight 64 Gbps Fibre Channel compression and decompression support on up to two external ports (no license required)
- In-flight 32 Gbps encryption and decryption on up to two external ports (no license required)
- 48 Virtual Channels (VCs) per port
- Port mirroring to monitor ingress or egress traffic from any port within the switch
- Two I2C connections able to interface to redundant management modules
- Hot pluggable — up to 4 hot pluggable switches per chassis
- Single fuse circuit
- Four temperature sensors

- Managed with Brocade Web Tools
- Supports a minimum of 128 domains in Native mode and Interoperability mode
- Nondisruptive code load in Native mode and Access Gateway mode
- 255 N_port logins per physical port
- D_port support on external ports
- Class 2 and Class 3 frames
- SNMP v1 and v3 support
- SSH v2 support
- SSL support
- NTP client support (NTP V3)
- FTP support for firmware upgrades
- SNMP/MIB monitoring functionality contained within the Ethernet Control MIB-II (RFC1213-MIB)
- End-to-end optics and link validation
- Sends switch events and syslogs to the Chassis Management Module (CMM)
- Traps identify cold start, warm start, link up/link down and authentication failure events
- Support for IPv4 and IPv6 on the management ports

The FC5022 16Gb SAN Scalable Switches come standard with the following software features:

- Brocade Full Fabric mode: Enables high performance 16 Gb or 8 Gb fabric switching
- Brocade Access Gateway mode: leverages NPIV to connect to any fabric without adding switch domains to reduce management complexity
- Dynamic Path Selection: enables exchange-based load balancing across multiple Inter-Switch Links for superior performance
- Brocade Advanced Zoning: segments a SAN into virtual private SANs to increase security and availability
- Brocade Enhanced Group Management: enables centralized and simplified management of Brocade fabrics through IBM Network Advisor

Enterprise Switch Bundle (ESB) software licenses

The IBM Flex System FC5022 24-port 16Gb ESB SAN Scalable Switch includes a complete set of licensed features that maximizes performance, ensures availability, and simplifies management for the most demanding applications and expanding virtualization environments.

This switch comes with 24 port licenses that can be applied to either internal or external links on this switch.

This switch also includes the following Enterprise Switch Bundle (ESB) software licenses:

- Brocade Extended Fabrics

  Provides up to 1000km of switches fabric connectivity over long distances.

- Brocade ISL Trunking

  Provides the ability to aggregate multiple physical links into one logical link for enhanced network performance and fault tolerance.

- Brocade Advanced Performance Monitoring

  Enables performance monitoring of networked storage resources. This license includes the TopTalkers feature.

- Brocade Fabric Watch

  Monitors mission-critical switch operations. Fabric Watch now includes new Port Fencing capabilities.

- Adaptive Networking

  Adaptive Networking provides a rich set of capabilities to the data center or virtual server environments. It ensures high priority connections to obtain the bandwidth necessary for optimum performance, even in congested environments. It optimizes data traffic movement within the fabric via Ingress Rate Limiting, Quality of Service, and Traffic Isolation Zones

- Server Application Optimization (SAO)

  This license optimizes overall application performance for physical servers and virtual machines. SAO, when deployed with Brocade Fibre Channel HBAs, extends Brocade Virtual Channel (VC) technology from fabric to the server infrastructure. This delivers application-level, fine-grain Quality of Service (QoS) management to the HBAs and related server applications.

Supported Fibre Channel standards

The switches support the following Fibre Channel standards:

- FC-AL-2 INCITS 332: 1999
- FC-GS-5 ANSI INCITS 427 (includes the following):
- FC-GS-4 ANSI INCITS 387: 2004
- FC-IFR INCITS 1745-D, revision 1.03 (under development)
- FC-SW-4 INCITS 418:2006 (includes the following):
- FC-SW-3 INCITS 384: 2004
- FC-VI INCITS 357: 2002
- FC-TAPE INCITS TR-24: 1999
- FC-DA INCITS TR-36: 2004 (includes the following):
- FC-FLA INCITS TR-20: 1998
- FC-PLDA INCIT S TR-19: 1998
- FC-MI-2 ANSI/INCITS TR-39-2005
- FC-PI INCITS 352: 2002
- FC-PI-2 INCITS 404: 2005
- FC-PI-4 INCITS 1647-D, revision 7.1 (under development)
- FC-PI-5 INCITS 479: 2011
- FC-FS-2 ANSI/INCITS 424:2006 (includes the following):
- FC-FS INCITS 373: 2003
- FC-LS INCITS 433: 2007
- FC-BB-3 INCITS 414: 2006 (includes the following):
- FC-BB-2 INCITS 372: 2003

- FC-SB-3 INCITS 374: 2003 (replaces FC-SB ANSI X3.271: 1996; FC-SB-2 INCITS 374: 2001)
- RFC 2625 IP and ARP Over FC
- RFC 2837 Fabric Element MIB
- MIB-FA INCITS TR-32: 2003
- FCP-2 INCITS 350: 2003 (replaces FCP ANSI X3.269: 1996)
- SNIA Storage Management Initiative Specification (SMI-S) Version 1.2 (includes the following):
- SNIA Storage Management Initiative Specification (SMI-S) Version 1.03 ISO standard IS24775-2006. Replaces (ANSI INCITS 388: 2004)
- SNIA Storage Management Initiative Specification (SMI-S) Version 1.1.0
- SNIA Storage Management Initiative Specification (SMI-S) Version 1.2.0

For more information, see the IBM Redbooks Product Guide for the IBM Flex System FC5022 16Gb SAN Scalable Switch, available from:

http://www.redbooks.ibm.com/abstracts/tips0870.html?Open

12.8.5 IBM Flex System FC3171 8Gb SAN Switch

This 8Gb SAN switch from QLogic is a full-fabric Fibre Channel switch module that can be converted to a pass-thru module when configured in transparent mode.

Figure 12-40 IBM Flex System FC3171 8Gb SAN Switch

The I/O module has 14 internal ports and 6 external ports. All ports are licensed on the switch as there are no port licensing requirements. Ordering information is listed in Table 12-10.

Table 12-10 FC3171 8Gb SAN Switch

Part number	Feature code[a]	Product Name
69Y1930	A0TD / 3595	IBM Flex System FC3171 8Gb SAN Switch

a. The first feature code listed is for configurations ordered through System x sales channels. The second feature code is for configurations ordered through the IBM Power Systems channel.

There are no SFP supplied as standard, the SFP modules and cables listed in Table 12-11 are supported.

Table 12-11 FC3171 8Gb SAN Switch supported SFP modules and cables

Part number	Feature codes[a]	Description
44X1964	5075 / 3286	IBM 8Gb SFP+ SW Optical Transceiver
39R6475	4804 / 3238	4Gb SFP Transceiver Option

a. The first feature code listed is for configurations ordered through System x sales channels. The second feature code is for configurations ordered through the IBM Power Systems channel.

It is possible to re-configure the FC3171 8Gb SAN Switch to become a Pass-Thru module using the switch GUI or CLI and then the Module can be converted back to a full function SAN switch at some future date. The switch requires a reset when turning transparent mode on or off.

The switch can be configured via either command line, or via QuickTools.

- Command Line: Access the switch via the console port through the Chassis Management Module or through the Ethernet Port. This method requires a basic understanding of the CLI commands.
- QuickTools: Requires a current version of the JRE on your workstation before pointing a web browser to the switch's IP address. The IP Address of the switch must be configured. QuickTools does not require a license and code is included.

On this switch when in Full Fabric mode, access to all of the Fibre Channel Security features is provided. Security includes additional services available, being Secure Socket Layer (SSL) and Secure Shell (SSH). In addition, RADIUS servers may be used for device and user authentication. Once SSL/SSH is enabled, then the Security features are available to be configured. This allows the SAN administrator to configure which devices are allowed to login to the Full Fabric Switch module, by creating security sets with security groups. These are configured on a per switch basis. The security features are not available when in pass-thru mode.

FC3171 8Gb SAN Switch specifications and standards:

- Fibre Channel standards:
  - C-PH version 4.3
  - FC-PH-2
  - FC-PH-3
  - FC-AL version 4.5
  - FC-AL-2 Rev 7.0
  - FC-FLA
  - FC-GS-3
  - FC-FG
  - FC-PLDA
  - FC-Tape
  - FC-VI
  - FC-SW-2
  - Fibre Channel Element MIB RFC 2837
  - Fibre Alliance MIB version 4.0
- Fibre Channel protocols:
  - Fibre Channel service classes: Class 2 and class 3
  - Operation modes: Fibre Channel class 2 and class 3, connectionless
- External port type:
  - Full fabric mode: Generic loop port (GL_port)
  - Transparent mode: Transparent fabric port (TF_port)
- Internal port type:
  - Full fabric mode: Fabric port (F_port)
  - Transparent mode: Transparent host port/NPIV mode (TH_port)
  - Support for up to 44 host NPIV logins

- Port characteristics:
  - External ports are automatically detected and self- configuring
  - Port LEDs illuminate at startup
  - Number of Fibre Channel ports: 6 external ports and 14 internal ports
  - Scalability: Up to 239 switches maximum depending on your configuration
  - Buffer credits: 16 buffer credits per port
  - Maximum frame size: 2148 bytes (2112 byte payload)
  - Standards-based FC FC-SW2 Interoperability
  - Support for up to a 255 to 1 port-mapping ratio
  - Media type: Small form-factor pluggable plus (SFP+) module
- 2Gb specifications
  - 2 Gb fabric port speed: 1.0625 or 2.125 Gbps (gigabits per second)
  - 2 Gb fabric latency: Less than 0.4 msec
  - 2 Gb fabric aggregate bandwidth: 80 Gbps at full duplex
- 4Gb specifications
  - 4 Gb switch speed: 4.250 Gbps
  - 4 Gb switch fabric point-to-point: 4 Gbps at full duplex
  - 4 Gb switch fabric aggregate bandwidth: 160 Gbps at full duplex
- 8Gb specifications
  - 8 Gb switch speed: 8.5 Gbps
  - 8 Gb switch fabric point-to-point: 8 Gbps at full duplex
  - 8 Gb switch fabric aggregate bandwidth: 320 Gbps at full duplex
- Nonblocking architecture to prevent latency
- System processor: PowerPC®

For more information, see the IBM Redbooks Product Guide for the IBM Flex System FC3171 8Gb SAN Switch, available from:

http://www.redbooks.ibm.com/abstracts/tips0866.html?Open

12.8.6 IBM Flex System FC3171 8Gb SAN Pass-thru

The IBM Flex System FC3171 8Gb SAN Pass-thru I/O module is an 8 Gbps Fibre Channel pass-thru SAN module that has 14 internal ports and six external ports. It is shipped with all ports enabled.

Figure 12-41 on page 304 shows the switch.

Figure 12-41 IBM Flex System FC3171 8Gb SAN Pass-thru

Ordering information is listed in Table 12-12 on page 305.

Table 12-12 FC3171 8Gb SAN Pass-thru part number

Part number	Feature code[a]	Description
69Y1934	A0TJ / 3591	IBM Flex System FC3171 8Gb SAN Pass-thru

a. The first feature code listed is for configurations ordered through System x sales channels. The second feature code is for configurations ordered through the IBM Power Systems channel.

Note: If there is a potential future requirement to enable full fabric capability, then this switch should not be purchased and instead the FC3171 8Gb SAN Switch should be considered.

There are no SFPs supplied with the switch and must be ordered separately. Supported transceivers and fiber optic cables are listed in Table 12-13.

Table 12-13 FC3171 8Gb SAN Pass-thru supported modules and cables

Part Number	Feature Code	Description
44X1964	5075 / 3286	IBM 8Gb SFP+ SW Optical Transceiver
39R6475	4804 / 3238	4Gb SFP Transceiver Option

The FC3171 8Gb SAN Pass-thru can be configured using either command line or QuickTools.

- Command Line: Access the module via the console port through the Chassis Management Module or through the Ethernet Port. This method requires a basic understanding of the CLI commands.
- QuickTools: Requires a current version of the JRE on your workstation before pointing a web browser to the modules IP address. The IP Address of the module must be configured. QuickTools does not require a license and code is included.

The pass-thru module supports the following standards:

- Fibre Channel standards:
  - C-PH version 4.3
  - FC-PH-2
  - FC-PH-3
  - FC-AL version 4.5
  - FC-AL-2 Rev 7.0
  - FC-FLA
  - FC-GS-3
  - FC-FG
  - FC-PLDA
  - FC-Tape
  - FC-VI
  - FC-SW-2
  - Fibre Channel Element MIB RFC 2837
  - Fibre Alliance MIB version 4.0
- Fibre Channel protocols:
  - Fibre Channel service classes: Class 2 and class 3
  - Operation modes: Fibre Channel class 2 and class 3, connectionless
- External port type: Transparent fabric port (TF_port)
- Internal port type: Transparent host port/NPIV mode (TH_port)
  - Support for up to 44 host NPIV logins

- Port characteristics:
  - External ports are automatically detected and self- configuring
  - Port LEDs illuminate at startup
  - Number of Fibre Channel ports: 6 external ports and 14 internal ports
  - Scalability: Up to 239 switches maximum depending on your configuration
  - Buffer credits: 16 buffer credits per port
  - Maximum frame size: 2148 bytes (2112 byte payload)
  - Standards-based FC FC-SW2 Interoperability
  - Support for up to a 255 to 1 port-mapping ratio
  - Media type: Small form-factor pluggable plus (SFP+) module
- Fabric point-to-point bandwidth: 2 Gbps or 8 Gbps at full duplex
- 2Gb Specifications
  - 2 Gb fabric port speed: 1.0625 or 2.125 Gbps (gigabits per second)
  - 2 Gb fabric latency: Less than 0.4 msec
  - 2 Gb fabric aggregate bandwidth: 80 Gbps at full duplex
- 4Gb Specifications
  - 4 Gb switch speed: 4.250 Gbps
  - 4 Gb switch fabric point-to-point: 4 Gbps at full duplex
  - 4 Gb switch fabric aggregate bandwidth: 160 Gbps at full duplex
- 8Gb Specifications
  - 8 Gb switch speed: 8.5 Gbps
  - 8 Gb switch fabric point-to-point: 8 Gbps at full duplex
  - 8 Gb switch fabric aggregate bandwidth: 320 Gbps at full duplex
- System processor: PowerPC
- Maximum frame size: 2148 bytes (2112 byte payload)
- Nonblocking architecture to prevent latency

For more information, see the IBM Redbooks Product Guide for the IBM Flex System FC3171 8Gb SAN Pass-thru, available from:

http://www.redbooks.ibm.com/abstracts/tips0866.html?Open

13

Certification

In this chapter we will provide an insight into some of the various professional certifications that exist and that are appropriate to the topics in this book.

13.1 Why certification?

A very good question is why should individuals bother to certify when there is more than enough work to do anyway? These are some of the benefits to an individual:

- Validates your skills and your knowledge
- Gains peer recognition
- Potentially becoming more valuable to your company and in the marketplace
- Ratifying your skills as an industry professional

To an employer, these are some of the benefits:

- Great way of benchmarking employees' skill levels
- Gives confidence about the employees' ability to support storage networks.
- Demonstrates a standards-based, non-proprietary, vendor neutral, storage concepts

13.2 IBM Professional Certification Program

Today's marketplace is both crowded and complex. Individuals and businesses that don't stay ahead of the curve risk being left behind. To develop a solid, competitive advantage—and to remain ahead of that curve—technology specialists are turning to Professional Certification from IBM. Our extensive portfolio of integrated certifications includes servers, software, application and solution skills. The certification process is designed to prepare you and your company to meet e-business initiatives with real solutions.

The IBM Certification Program will assist in laying the groundwork for your personal journey to become a world-class resource to your customers, colleagues, and company, by providing you with the appropriate skills and accreditation needed to succeed.

13.2.1 About the program

IBM Professional Certification is both a journey and a destination. It's a business solution. A way for skilled IT professionals to demonstrate their expertise to the world. It validates your skills and demonstrates your proficiency in the latest IBM technology and solutions.

The certification requirements are tough, but it's not rocket science, either. It's a rigorous process that differentiates you from everyone else.

The mission of IBM Professional Certification is to:

- Provide a reliable, valid and fair method of assessing skills and knowledge.
- Provide IBM a method of building and validating the skills of individuals and organizations.
- Develop a loyal community of highly skilled certified professionals who recommend, sell, service, support and/or use IBM products and solutions.

13.2.2 Certifications by product

IBM has a myriad of certification courses that covers software, hardware, and products.

For a complete list of the courses available, refer to:

http://www-03.ibm.com/certify/certs/index.shtml

13.2.3 Mastery tests

Mastery tests are used to verify the mastery of knowledge covered in a course or a defined set of learning materials. They are not certification tests, which are designed to validate skills needed in a specific job-role. Rather, mastery tests help to assure an individual has achieved a foundation of knowledge and understanding of a subject matter.

Mastery tests supplement certifications as a method used by IBM to evaluate knowledge of IBM Sales and Technical Professionals. As with certifications, the successful completion of a mastery test may be required for participation in some IBM Business Partner activities.

For a complete list of the courses available, refer to:

http://www-03.ibm.com/certify/mastery_tests/index_bd.shtml

13.3 Storage Networking Industry Association certifications

The Storage Networking Industry Association (SNIA) provides vendor neutral certifications. There are different certification options within SNIA, and the certification program itself is called the Storage Networking Certification Program (SNCP).

SNCP provides a strong foundation of vendor-neutral, systems-level credentials that integrate with and complement individual vendor certifications.

The structure of the SNCP has been enhanced to reflect the advancement and growth of storage networking technologies over the past few years, and to provide for expanded offerings in the future. Through evolving and enhancing the SNCP, the SNIA is establishing a uniform standard by which individual knowledge and skill sets can be judged.

Before the establishment of the SNIA SNCP, there was no single standard by which to measure a professional's knowledge of storage networking technologies. Through its certification program, the SNIA is working to establish open standards for storage networking certification that IT organizations can trust.

Note: The SNCP Foundations exam (S10–101) has been withdrawn.

13.3.1 SNIA Certified Storage Professional (SCSP)

Attaining this vendor-neutral credential demonstrates that a storage networking professional has attained a foundation of knowledge and expertise in fundamental storage networking technologies and concepts, independent of any product-specific certifications.

13.3.2 SNIA Certified Storage Engineer (SCSE)

This vendor-neutral credential addresses storage managers and administrators responsible for effective management and administration, troubleshooting and diagnosing a SAN.

- This certification includes:
- Performance management
- Implementation of upgrades
- Installation of new SANs
- Backup and recovery
- Business continuance

13.3.3 SNIA Certified Storage Architect (SCSA)

This vendor-neutral credential has been designed for storage architects and storage networking professionals who:

- Assess
- Plan
- Design complex storage networking solutions

This certification validates the individual's abilities and allows them to leverage industry standards in their programs.

13.3.4 SNIA Certified Storage Networking Expert (SCSN-E)

This credential is a culmination of the above vendor-netural technical expertise. Combined with one or more vendor certifications, this credential positions the individual to work in a multi-vendor landscape, e.g storage managed services, partner or reseller environment.

13.3.5 SNIA Qualified Data Protection Associate

This vendor-netural credential is designed for individuals seeking to gain expertise and validate their knowledge in the storage data protection area.

This credential signifies that the candidate can:

- Recognize and describe concepts of data protection, restoration, and recovery methods
- Assess data protection planning and strategies
- Use management tools and practices
- Evaluate data protection methods and practices
- Assess security/confidentiality
- Troubleshoot potential pain points of data protection and recovery

13.3.6 SNIA Qualified Storage Virtualization Associate

This vendor-netural credential is designed for individuals seeking to gain expertise and validate their knowledge in storage virtualization.

This credential signifies that the candidate can:

- Define virtualization concepts
- Describe the benefits of virtualization
- Identify the potential pain points of virtualization
- Describe virtualization implementation strategies
- Explain administrative and management tasks required for virtualization

13.3.7 SNIA Qualified Storage Sales Professional

SNIA's Qualified Storage Sales Professional (SQSSP) credential is vendor neutral and is designed to validate storage concepts, terminology and basic customer need assessments.

In direct response to firms requesting specialized training for their sales and marketing professionals, this vendor–neutral training and credential provides individuals with the technologies in today’s complex data center.

It arms sales professionals by arming them with a broad set of information about a variety of different solutions in the storage ecosystem.

13.3.8 CompTIA Storage+ Powered by SNIA

CompTIA Storage+ Powered by SNIA is a vendor-neutral certification that validates the knowledge and skills required of IT storage professionals.

The CompTIA Storage+ Powered by SNIA certification exam covers the knowledge and skills required to configure basic networks to include archive, backup, and restoration technologies. Additionally, the successful candidate will be able to understand the fundamentals of business

continuity, application workload, system integration, and storage/system administration, while performing basic troubleshooting on connectivity issues and referencing documentation.

The exam is targeted toward IT storage professionals with at least twelve months of experience. Though it is not required, CompTIA A+, CompTIA Network+ or CompTIA Server+ certification is recommended. For more details refer to:

http://certification.comptia.org/getCertified/certifications/storage.aspx

13.4 Brocade certifications

Brocade has a large track of certification exams. Once completing three of the tracks outlined in the following, it is possible to acchieve the top credential of Brocade Distinguished Architect.

Each individual track has various exams in it as follows.

Brocade Certified Professional FICON Track

- Brocade Accredited Data Center Specialist Exam 160-130
- Brocade Accredited FICON Specialist Exam 160-140
- Brocade Certified Fabric Administrator (BCFA) Exam 143-410
- Brocade Certified Fabric Professional (BCFP) Exam 143-070
- Brocade Certified Architect For FICON (BCAF) Exam 143-120

Brocade Certified Professional Data Center Track

- Brocade Accredited Server Connectivity Specialist Exam 160-020
- Brocade Accredited Data Center Specialist Exam 160-130
- Brocade Certified Fabric Administrator (BCFA) Exam 143-410
- Brocade Certified Fabric Professional (BCFP) Exam 143-070
- Brocade Certified San Manager (BCSM) Exam 143-360
- Brocade Certified Fabric Designer (BCFD) Exam 143-260

Brocade Certified Professional Internetworking Track

- Brocade Accredited Internetworking Specialist Exam 160-120
- Brocade Certified Network Engineer Exam 150-120
- Brocade Certified Layer 4-7 Engineer Exam 150-320
- Brocade Certified Network Professional Exam 150-220
- Brocade Certified Layer 4-7 Professional Exam 150-420
- Brocade Accredited WLAN Specialist Exam 160-170
- Brocade Certified Network Designer Exam 150-510

Brocade Certified Professional Converged networking Track

- Brocade Accredited FCoE Specialist Exam 160-160
- Brocade Certified Ethernet Fabric Engineer Exam 150-610
- Brocade Certified FCoE Professional (BCFCoEP) Exam 143-510

13.4.1 Brocade Accredited Server Connectivity Specialist

This certification is for professionals with understanding of basic Brocade server adapter concepts, and can demonstrate knowledge of installation, maintenance and troubleshooting. This certification requires successful completion of:

- Brocade Accredited Server Connectivity Specialist Exam 160-020

13.4.2 Brocade Accredited Data Center Specialist

This certification is for professionals with understanding of basic Fibre Channel theory, terminology, hardware, and the various reasons and benefits from implementing a Storage Area Network (SAN). This certification requires successful completion of:

- Brocade Accredited Data Center Specialist Exam 160-130

13.4.3 Brocade Accredited FICON Specialist

This certification is for professionals who understand basic mainframe terminology and the relationship between FICON and Open Systems, can recognize the functions of FMS and RMF™ and can identify Brocade hardware and software products that support FICON.This certification requires successful completion of:

- Brocade Accredited FICON Specialist Exam 160-140

13.4.4 Brocade Accredited FCoE Specialist

This certification is for professionals who have an understanding of FCoE, CEE, the FCoE Initialization Protocol and the associated Brocade hardware.This certification requires successful completion of:

- Brocade Accredited FCoE Specialist Exam 160-160

13.4.5 Brocade Accredited Internetworking Specialist

This certification is for professionals who have an understanding of the basic internetworking terminology, hardware, routing concepts, the OSI 7 Layer Model and the TCP/IP protocol suite including IP addressing.This certification requires successful completion of:

- Brocade Accredited Internetworking Specialist Exam 160-120

13.4.6 Brocade Accredited WLAN Specialist

This certification is for professionals with ability to demonstrate knowledge of wireless concepts, and install, maintain and troubleshoot a Brocade Mobility solution.This certification requires successful completion of:

- Brocade Accredited WLAN Specialist Exam 160-170

13.4.7 Brocade Certified Fabric Administrator (BCFA)

This certification is for beginners who have basic FC protocol understanding and perform basic switch configurations and trobleshooting. The exam which is required to be completed for this certification is:

- The Brocade Certified Fabric Administrator (BCFA) Exam 143-410

13.4.8 Brocade Certified Fabric Professional (BCFP)

This certification is for professionals with understanding of advanced FC technologies like FC switching, routing in extended environment, and will be able to configure, administer and troubleshoot the FC router, extended fabrics and also implement, adaptive networking in SAN. The exam which is required to be completed for this certification is:

- The Brocade Certified Fabric Professional (BCFP) Exam 143-070

Note: Brocade has announced new BCFA and BCFP beta exam on 16 Gbps, the information related to the exams can be found at:

http://community.brocade.com/community/forums/education/certification

13.4.9 The Brocade Certified SAN Manager (BCSM)

This certification is for experts who are skilled with administering, configuring and troubleshooting the Brocade products with help of management tools. Also this focus on implementation of security enhancements, Monitoring and alerting of brocade SAN. Below is the exam which is required to be completed for this certification.

- The Brocade Certified San Manager (BCSM) Exam 143-360

13.4.10 Brocade Certified Fabric Designer (BCFD)

This certification is for skilled persons with ability to design Data Center Fabric, a to provide the implementation plans. This requires design skills with various criteria like reliability, availability and scalability also plan for integration of new devices into current infrastructure. The exam which is required to be completed for this certification is:

- The Brocade Certified Fabric Designer (BCFD) Exam 143-260

13.4.11 Brocade Certified Architect For FICON (BCAF)

This certification is for experts who have a good understanding in IBM System z I/O and are able to identify, design, implement and support Brocade products for mainframe FICON requirements. The exam which is required to be completed for this certification is:

- The Brocade Certified Architect For FICON (BCAF) Exam 143-120.

13.4.12 Brocade Certified FCoE Professional (BCFCoEP)

This certification is for professionals who have the ability to show skills on FCoE, and CEE concepts. They will also will be able to design, implement, support and troubleshoot Brocade's FCoE products. This is the exam which is required to be completed for this certification:

- The Brocade Certified FCoE Professional (BCFCoEP) Exam 143-510.

13.4.13 Brocade Certified Ethernet Fabric Engineer

This certification is for professional who have the ability to demonstrate knowledge of Ethernet fabric concepts, Brocade Ethernet fabric products and be able to install, configure, manage and troubleshoot Brocade Ethernet fabrics.This certification requires successful completion of below exam,

- Brocade Certified Ethernet Fabric Engineer Exam 150-610

13.4.14 Brocade Certified Network Engineer

This certification is for professional who have the ability to install and maintain IP switching and routing (Layer 2/3) networks based on Brocade products. This certification requires successful completion of:

- Brocade Certified Network Engineer Exam 150-120

13.4.15 Brocade Certified Layer 4-7 Engineer

This certification is for professionals with the ability to install, configure, maintain, and perform basic troubleshooting of Brocade Layer 4-7 application delivery products. This certification requires successful completion of:

- Brocade Certified Layer 4-7 Engineer Exam 150-320

13.4.16 Brocade Certified Network Professional

This certification is for professionals with the ability to install, configure, maintain, and troubleshoot Brocade Ethernet switches and routers in complex environments. This certification requires successful completion of:

- Brocade Certified Network Professional Exam 150-220

13.4.17 Brocade Certified Layer 4-7 Professional

This certification is for professionals with the ability to design, configure, administer and troubleshoot complex implementations of Brocade Layer 4-7 application delivery solutions. This certification requires successful completion of:

- Brocade Certified Layer 4-7 Professional Exam 150-420

13.4.18 Brocade Certified Network Designer

This certification is for professionals with the ability to design a campus or enterprise network using Brocade solutions. This certification requires successful completion of:

- Brocade Certified Network Designer 150-510

For further information on certification on Brocade, please refer to below link,

http://www.brocade.com/education/certification-accreditation/index.page

13.5 Cisco certification

Cisco has various certifications for different product categories. We concentrate on SAN and system networking here. Cisco has five levels of general IT certification: Entry, Associate, Professional, Expert and Architect.

Table 13-1 on page 316 list various Cisco certifications paths and their respective exams for different levels from Entry to Expert.

Table 13-1 Cisco Certification levels and paths

Certification Paths	Entry Level	Associate	Professional	Expert	Architect
Routing and Switching	CCENT	CCNA	CCNP	CCIE Routing and Switching	Cisco Certified Architect (CCAr)
Design	CCENT	CCNA and CCDA	CCDP	CCDE	
Network Security	CCENT	CCNA Security	CCSP and CCNP Security	CCIE Security	
Wireless	CCENT	CCNA Wireless	CCNP Wireless	CCIE Wireless	
Storage Networking	CCENT	CCNA	CCNP	CCIE Storage Networking	

13.5.1 Cisco Certified Entry Networking Technician (CCENT)

This is the entry level certification for professionals with the ability to install, operate and troubleshoot a small enterprise branch network, including basic network security. This certification requires successful completion of:

- Interconnecting Cisco Networking Devices Part 1 640-822 ICND1

13.5.2 Cisco Certified Network Associate (CCNA)

This certification is for professionals with the ability to install, configure, operate, and troubleshoot medium-size routed and switched networks, including implementation and verification of connections to remote sites in a WAN. This certification requires successful completion of:

- Cisco Certified Network Associate 640-802 CCNA

13.5.3 Cisco Certified Network Associate Security (CCNA Security)

This certification is for professionals with the ability to secure Cisco networks, to develop a security infrastructure, recognize threats and vulnerabilities to networks, and mitigate security threats. This certification requires successful completion of:

- Implementing Cisco IOS Network Security 640-553 IINS

13.5.4 Cisco Certified Network Associate Wireless (CCNA Wireless)

This certification is for professionals with associate-level knowledge and skills to configure, implement and support wireless LANs using Cisco equipment. This certification requires successful completion of:

- Implementing Cisco Unified Wireless Networking Essentials 640-721 IUWNE

13.5.5 Cisco Certified Design Associate (CCDA)

This certification is for professionals with the ability to design a Cisco converged network, to design routed and switched network infrastructures and services involving LAN, WAN, and broadband access. This certification requires successful completion of:

- Designing for Cisco Internetwork Solutions Exam - 640-864

13.5.6 Cisco Certified Network Professional (CCNP)

This certification is for professionals with the ability to plan, implement, verify and troubleshoot local and wide-area enterprise networks and work collaboratively with specialists on advanced security, voice, wireless and video solutions. This certification requires successful completion of:

- The Implementing Cisco IP Routing (ROUTE 642-902)
- Implementing Cisco IP Switched Networks (SWITCH 642-813)
- Troubleshooting and Maintaining Cisco IP Networks (TSHOOT 642-832)

13.5.7 CCNP Security certification

This certification is for professionals responsible for Security in Routers, Switches, Networking devices and appliances, as well as choosing, deploying, supporting and troubleshooting firewalls, VPNS, and IDS/IPS solutions for their networking environments. This certification requires successful completion of:

- Securing Networks with Cisco Routers and Switches (SECURE) v1.0 642-637
- Deploying Cisco ASA Firewall Solutions (FIREWALL v1.0) 642-617
- Deploying Cisco ASA VPN Solutions (VPN v1.0) 642-647
- Implementing Cisco Intrusion Prevention System v7.0 642-627

13.5.8 CCNP Wireless certification

This certification is for professionals with good expertise in designing, implementing, and operating Cisco Wireless networks and mobility infrastructures. The professionals also need to have the ability to assess and translate network business requirements into technical specifications which can be incorporated into successful installations. This certification requires successful completion of:

- Conducting Cisco Unified Wireless Site Survey CUWSS (642-731)
- Implementing Cisco Unified Wireless Voice Networks IUWVN (642-741)
- Implementing Cisco Unified Wireless Mobility Services IUWMS (642-746)
- Implementing Advanced Cisco Unified Wireless Security IAUWS (642-736)

13.5.9 Cisco Certified Design Professional (CCDP)

This certification is for professionals with advanced knowledge of network design concepts and principles with the ability to discuss, design, and create advanced addressing and routing, security, network management, data center of multi-layered enterprise architectures that include virtual private networking and wireless domains. This certification requires successful completion of:

- Implementing Cisco IP Routing 642-902
- Implementing Cisco IP Switched Networks 642-813
- Designing Cisco Network Service Architectures 642-874

13.5.10 Cisco CCIE Routing and Switching

This is for professionals who are expert-level network engineers and plan, operate and troubleshoot complex, converged network infrastructures. This certification requires successful completion of:

- CCIE Routing and Switching Written Exam #350-001, v4.0
- CCIE Routing & Switching v4.0 Lab Exam

13.5.11 Cisco CCIE Security

This certification is for professionals who have the knowledge and skills to implement, maintain and support extensive Cisco Network Security Solutions using the latest industry best practices and technologies. This certification requires successful completion of:

- CCIE Security Written Exam 350-018
- CCIE Security Lab Exam v3.0

13.5.12 Cisco CCIE Wireless certification

This certification is for professionals with a broad theoretical knowledge of wireless networking and a solid understanding of wireless local area networking (WLAN) technologies from Cisco. This certification requires successful completion of:

- 350-050 CCIE Wireless Exam
- CCIE Wireless Lab Exam

13.5.13 Cisco Certified Design Expert (CCDE)

This certification is for professionals with the ability to design large enterprise networks and develop solutions which address planning, design, integration, optimization, operations, security and support the infrastructure. This certification requires successful completion of:

- CCDE Written Exam 352-001
- CCDE Practical Exam

13.5.14 Cisco CCIE Storage Networking

This certification is for professionals with a good understanding on FC protocols, Cisco products, management and troubleshooting tools from entry level to high end products and applications. This certification requires successful completion of:

- CCIE Storage Networking Written Exam 350-04
- CCIE Storage Networking Lab Exam

13.5.15 Cisco Certified Architect

CCDE certification is a pre-requisite for this certification, and professionals who apply for this certification need to appear for an interview with a Cisco board of members for the validation of the architect role, and during which a skills assessment will be done.

13.5.16 Cisco specialization tracks

Apart from the general IT certification tracks Cisco also have certifications for various specializations. For example, for data center professionals there are certifications with various specializations such as sales, design and support. These are shown below.

Data Center Networking Infrastructure

- Cisco Data Center Networking Infrastructure Design Specialist
- Cisco Data Center Networking Infrastructure Sales Specialist
- Cisco Data Center Networking Infrastructure Support Specialist

Data Center Storage Networking

- Cisco Data Center Storage Networking Design Specialist
- Cisco Data Center Storage Networking Sales Specialist
- Cisco Data Center Storage Networking Support Specialist

For further details on Cisco certifications refer to:

http://www.cisco.com/web/learning/le3/learning_career_certifications_and_learning_paths_home.html

13.6 The Open Group certifications

The Open Group provides Certification Programs for people, products and services that meet their standards. For enterprise architects and IT specialists, the certification programs provide a worldwide professional credential for knowledge, skills and experience. For IT products, Open Group Certification Programs offer a worldwide guarantee of conformance.

13.6.1 The Open Group Certified IT Specialists (Open CITS)

This is a vendor neutral, Open group, global certification program for IT specialists. This certification applies to various domains of IT industry and not specific only to Storage domain. Depending upon the professionals skillset and experience there are three levels of certifications for IT specialists as detailed below:

- Level 1: Certified IT Specialist — able to perform as a contributing Specialist with assistance/supervision, with a wide range of appropriate skills
- Level 2: Master Certified IT Specialist — able to perform independently as lead Specialist, and take responsibility for delivery of solutions
- Level 3: Distinguished IT Specialist — delivering leadership, scope, depth and breadth of impact.

13.6.2 The Open Group Certified Architect (Open CA)

This is for architect professionals in IT, business and enterprise architecture. This also has three levels of certifications as detailed below:

- Level 1 Certified — Professionals with the ability to perform with assistance/supervision and who have a wide range of appropriate skills, as a contributing architect.
- Level 2: Master — Professionals with the ability to perform independently and take responsibility for delivery of systems and solutions as lead architect.
- Level 3: Distinguished — Professionals who have significant breadth and depth of impact on the business through the application of IT architecture.

13.6.3 Open Group Certification

For more information refer to:

http://www3.opengroup.org/certifications

13.7 Juniper Networks Certification Program

Juniper networks has Junos based and non-Junos based platform specific certifications.

13.7.1 JNCP Junos based certification tracks

JNCP has three tracks of professional certifications based on the Junos platform.

- Service Provider Routing and Switching
- Enterprise Routing and Switching
- Junos Security

Table 13-2 Junos tracks

Certification Track	Associate JNCIA	Specialist JNCIS	Professional JNCIP	Expert JNCIE
Service Provider and Switching	JNCIA- Junos	JNCIS-SP	JNCIP-SP	JNCIE-SP
Enterprise Routing & Switching	JNCIA- Junos	JNCIS-ENT	JNCIP-ENT	JNCIE-ENT
Junos Security	JNCIA-Junos	JNCIS-SEC	JNCIP-SEC	JNCIE-SEC

13.7.2 Service Provider Routing and Switching track

These are the certifications in the Service Provider Routing and Switching track.

Juniper Networks Certified Internet Associate (JNCIA–Junos)

Designed for experienced networking professionals with beginner-intermediate knowledge of networking, this written exam verifies the candidate's understanding of the Juniper Networks Junos OS, networking fundamentals and basic routing and switching.

This certification requires successful completion of:

- Juniper Networks Junos Associate (JNCIA-Junos) Exam code: JN0-101

Juniper Networks Certified Internet Specialist (JNCIS-SP)

Designed for experienced networking professionals with beginner to intermediate knowledge of routing and switching implementations in Junos, this written exam verifies the candidate's basic understanding of routing and switching technologies and related platform configuration and troubleshooting skills.

This certification requires successful completion of:

- JNCIA-Junos
  - (Acceptable substitutions: JNCIA-ER, JNCIS-ER, JNCIA-EX, JNCIA-M, JNCIS-M) and:
- Juniper Networks Certified Internet Specialist (JNCIS-SP) Exam code: JN0-360

Juniper Networks Certified Internet Professional (JNCIP–SP)

Designed for experienced networking professionals with advanced knowledge of the Juniper Networks Junos OS, this written exam verifies the candidate's understanding of advanced routing technologies and related platform configuration and troubleshooting skills.

This certification requires successful completion of:

- JNCIS-SP
  - (Acceptable substitution: JNCIS-M) and:
- Juniper Networks Certified Internet Professional (JNCIP-SP) Exam code: JN0-660

Juniper Networks Certified Internet Expert (JNCIE–SP)

At the pinnacle of the Service Provider Routing and Switching certification track is the one-day JNCIE-SP Lab Exam. This exam is designed to validate the networking professionals' ability to implement, troubleshoot and maintain Juniper Networks service provider networks. The 8-hour format of this exam requires that candidates build a service provider network consisting of multiple MX series routers. Successful candidates will perform system configuration on all devices, implement various protocols, policies and VPNs, HA capabilities, and Class of Services on 8 MX Series Ethernet Services Routers.

This certification requires successful completion of:

- JNCIP-SP
  - (Acceptable substitution: JNCIP-M) and:
- Juniper Networks Certified Internet Expert (JNCIE–SP) Lab Exam code: JPR-960

13.7.3 Enterprise Routing and Switching track

These are the certifications in the Enterprise Routing and Switching track.

Juniper Networks Certified Internet Specialist (JNCIS-ENT)

Designed for experienced networking professionals with beginner to intermediate knowledge of routing and switching implementations in Junos, this written exam verifies the candidate's basic understanding of routing and switching technologies and related platform configuration and troubleshooting skills.

This certification requires successful completion of:

- JNCIA-Junos
  - (Acceptable substitutions: JNCIA-ER, JNCIS-ER, JNCIA-EX, JNCIA-M, JNCIS-M) and:
- Juniper Networks Certified Internet Specialist (JNCIS-ENT) - Exam code: JN0-343

Juniper Networks Certified Internet Professional (JNCIP-ENT)

Designed for experienced networking professionals with advanced knowledge of the Juniper Networks Junos OS, this written exam verifies the candidate's understanding of advanced enterprise routing and switching technologies and related platform configuration and troubleshooting skills.

This certification requires successful completion of:

- JNCIS-ENT or JNCIS-ER and:
- Juniper Networks Certified Internet Professional (JNCIP-ENT) Exam code: JN0-643

Juniper Networks Certified Internet Expert (JNCIE-ENT)

At the pinnacle of the Enterprise Routing and Switching certification track is the one-day JNCIE-ENT practical exam. This exam is designed to validate the networking professionals' ability to deploy, configure, manage and troubleshoot Junos-based enterprise routing and switching platforms. Throughout this 8-hour practical exam, candidates will build an enterprise network infrastructure consisting of multiple routers and switching devices. Successful candidates will perform system configuration on all devices, configure protocols and features like IPV6 ,OSPF V2 , OSPF V3, BGP, MSDP, PIM, SSM, RSTP,LLDP, 802.1x , CoS, routing policies.

This certification requires successful completion of:

- Juniper Networks Certified Internet Expert (JNCIE-ENT) Exam code: JPR-943

13.7.4 Junos Security track

These are the certifications in the Security track.

Juniper Networks Certified Internet Specialist (JNCIS-SEC)

Designed for experienced networking professionals with intermediate knowledge of the Juniper Networks Junos software for SRX Series devices, this written exam verifies the candidate's understanding of security technologies and related platform configuration and troubleshooting skills.

This certification requires successful completion of:

- JNCIA-Junos

– (Acceptable substitutions: JNCIA-ER, JNCIS-ER, JNCIA-M, JNCIS-M or JNCIA-EX) and:

- Juniper Networks Certified Internet Specialist (JNCIS-SEC) Exam code: JN0-332

Juniper Networks Certified Internet Professional (JNCIP-SEC)

Designed for experienced networking professionals with advanced knowledge of the Juniper Networks Junos software for SRX Series devices, this written exam verifies the candidate's understanding of advanced security technologies and related platform configuration and troubleshooting skills.

This certification requires successful completion of:

- JNCIS-SEC and:
- Juniper Networks Certified Internet Professional (JNCIP-SEC) - Exam code: JN0-632

Juniper Networks Certified Internet Expert (JNCIE-SEC)

At the pinnacle of the Junos Security certification track is the one-day JNCIE-SEC practical exam. This exam is designed to validate the networking professionals' ability to deploy, configure, manage and troubleshoot JUNOS-based security platforms. Throughout this 8-hour practical exam, candidates will build a secure enterprise network consisting of multiple firewall devices interconnected via IPSec VPNs. Successful candidates will perform system configuration on all devices, configure secure management capabilities, install complex policies and attack prevention features, HA capabilities, IPS features.

This certification requires successful completion of:

- JNCIP-SEC and:
- Juniper Networks Certified Internet Expert (JNCIE-SEC) Exam code: JPR-932

13.8 Non Junos Certification Tracks

Non Junos platforms have certification tracks as shown in Table 13-3.

Table 13-3 Non Junos track

Certification Track	Associate JNCIA	Specialist JNCIS	Professional JNCIP
E-Series	JNCIA-E	JNCIS-E	JNCIP-E
Firewall/ VPN	JNCIA-FWV	JNCIS-FWV	
SSL	JNCIA-SSL	JNCIS-SSL	
IDP	JNCIA-IDP		
Unified Access Control	JNCIA-AC		
WX Series	JNCIA-WX		

13.8.1 E-Series certification Track

The Juniper Networks Certification Program (JNCP) E-series certification track is a multi-tiered program that allows participants to demonstrate, through a combination of written

proficiency exams and hands-on configuration and troubleshooting exams, competence with specific Juniper Networks technology. Successful candidates demonstrate thorough understanding of networking technology in general and Juniper Networks E-series platforms and operating system in particular.

Juniper Networks Certified Internet Associate (JNCIA- E)

Designed for experienced networking professionals with beginner to intermediate knowledge of the Juniper Networks E-series platforms, this written exam verifies the candidate's basic understanding of Internet technology and related platform configuration and troubleshooting skills. JNCIA-E exam topics are based on the content of the Introduction to Juniper Networks Routers–E-series and E-series B-RAS Configuration Basics instructor-led training courses. This exam is not a prerequisite for the JNCIS-E exam.

This certification requires successful completion of:

- Juniper Networks Certified Internet Associate (JNCIA- E) Exam code: JN0-120

Juniper Networks Certified Internet Specialist (JNCIS-E)

The JNCIS-E exam is designed for networking professionals with advanced knowledge of, and experience with, the Juniper Networks E-series platforms. The JNCIS-E exam tests for a wider and deeper level of knowledge than does the JNCIA-E exam. Sources of question content include the E-series platforms documentation set, on the job product experience, as well as the understanding of Internet technologies and design principles considered to be common knowledge at the Specialist level. Passing the JNCIS-E exam is a prerequisite for attempting the JNCIP-E practical exam.

This certification requires successful completion of:

- Juniper Networks Certified Internet Specialist (JNCIS-E) Exam code: JN0-130

Juniper Networks Certified Internet Professional (JNCIP-E)

The JNCIP-E exam is a one-day practical exam designed to validate the candidate's ability to successfully build an ISP consisting of multiple E-series virtual routers. This certification establishes the candidate's practical and theoretical knowledge of basic and advanced Internet technologies as well as the candidate's ability to effectively apply that knowledge in a hands-on environment. Candidates configure and troubleshoot routing scenarios utilizing various protocols and technologies on E-series platforms.

This certification requires successful completion of:

- Juniper Networks Certified Internet Professional (JNCIP-E) Exam code: CERT-JNCIP-E

13.8.2 Firewall/VPN certification Track

The Juniper Networks Certification Program (JNCP) Firewall/VPN certification track is a two-tiered program that allows participants to demonstrate competence with Juniper Networks Firewall with VPN products and the ScreenOS software.

Juniper Networks Certified Internet Associate (JNCIA-FWV)

Designed for experienced networking professionals with beginner to intermediate knowledge of Juniper Firewall/VPN products and ScreenOS software, this written exam verifies the candidate's basic understanding of Internet and security technology and related device configuration. JNCIA-FWV exam topics are based on the content of the Configuring Juniper Networks Firewall/IPSec VPN Products instructor-led training course. This exam is not a prerequisite for the JNCIS-FWV certification.

This certification requires successful completion of:

- Juniper Networks Certified Internet Associate (JNCIA-FWV) Exam code: JN0-522

Juniper Networks Certified Internet Specialist (JNCIS-FWV)

The JNCIS-FWV is designed for networking professionals with advanced knowledge of, and experience with, Juniper Firewall/VPN products and ScreenOS software. The JNCIS-FWV exam tests for a wider and deeper level of knowledge than does the JNCIA-FWV exam. Sources of question content include all ScreenOS training courses, the Firewall/VPN and ScreenOS documentation set, on-the-job product experience, as well as Internet technologies and design principles considered to be common knowledge at the Specialist level.

This certification requires successful completion of:

- Juniper Networks Certified Internet Specialist (JNCIS-FWV) Exam code: JN0-532

13.8.3 SSL certification track

The Juniper Networks Certification Program (JNCP) SSL certification track allows participants to demonstrate competence with Juniper Networks Secure Access products and their deployment.

Juniper Networks Certified Internet Associate (JNCIA-SSL)

Designed for experienced networking professionals with beginner-intermediate knowledge of the Juniper Networks Secure Access products and their deployment. JNCIA-SSL exam topics are based on the content of the Configuring Juniper Networks Secure Access instructor led training course.

This certification requires successful completion of:

- Juniper Networks Certified Internet Associate (JNCIA-SSL) Exam code: JN0-562

Juniper Networks Certified Internet Specialist (JNCIS-SSL)

Designed for experienced networking professionals with intermediate knowledge of the Juniper Networks Secure Access products and their deployment. JNCIS-SSL exam topics are based on the content of the Advanced Juniper Networks Secure Access instructor led training course.

This certification requires successful completion of:

- Juniper Networks Certified Internet Specialist (JNCIS-SSL) Exam code: JN0-570

13.8.4 Intrusion Detection and Prevention (IDP) Track

The Juniper Networks Certification Program (JNCP) IDP certification track allows participants to demonstrate competence with Juniper Networks NetScreen IDP products and their deployment.

Juniper Networks Certified Internet Associate (JNCIA-IDP)

Designed for experienced networking professionals with beginner to intermediate knowledge of the Juniper Networks IDP products and their deployment. JNCIA-IDP exam topics are based on the content of the Implementing Intrusion Detection and Prevention (IIDP) instructor-led training course.

This certification requires successful completion of:

- Juniper Networks Certified Internet Associate (JNCIA-IDP) Exam code: JN0-541

13.8.5 Unified Access Control (UAC) Track

The Juniper Networks Certification Program (JNCP) Unified Access Control certification track allows participants to demonstrate competence with Juniper Networks Unified Access Control products and their deployment.

Note: Effective April 13, 2012, the JN0-141 AC, Associate (JNCIS-AC) exam will End of Life (EOL). At that time, the Juniper Networks Certified Internet Associate, AC (JNCIA-AC) certification and the Access Control Track will become inactive and unsupported. However, once earned, a JNCIA-AC credential is valid for a period of two years.

The replacement exam is the JN0-314 Junos Pulse Access Control, Specialist (JNCIS-AC), which earns the Juniper Networks Certified Specialist Junos Pulse Access Control (JNCIS-AC) certification, will be available March 16, 2012.

Juniper Networks Certified Internet Associate (JNCIA-AC)

Designed for experienced networking professionals with beginner to intermediate knowledge of the Juniper Networks Unified Access Control products and their deployment. JNCIA-AC exam topics are based on the content of the Configuring Unified Access Control instructor led training course.

This certification requires successful completion of:

- Juniper Networks Certified Internet Associate (JNCIA-AC) Exam code: JN0-141

13.8.6 WX certification track

The Juniper Networks Certification Program (JNCP) WX certification track allows participants to demonstrate competence with Juniper Networks WAN Acceleration platforms and their deployment.

Juniper Networks Certified Internet Associate (JNCIA-WX)

Designed for experienced networking professionals with beginner to intermediate knowledge of the Juniper Networks WAN Acceleration (WX) and WAN Acceleration Cache (WXC) platforms and their deployment. JNCIA-WX exam topics are based on the content of the Operating Juniper Networks WX Application Acceleration Platforms (OJWX) instructor led training course.

This certification requires successful completion of:

- Juniper Networks Certified Internet Associate (JNCIA-WX) Exam code: JN0-311

For further information on the Juniper certifications refer to:

http://www.juniper.net/us/en/training/certification/certification-tracks/

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

- *IBM TotalStorage: SAN Product, Design, and Optimization Guide*, SG24-6384
- *IBM TotalStorage SAN Volume Controller*, SG24-6423
- *Implementing an Open IBM SAN,* SG24-6116
- *Implementing the Cisco MDS 9000 in an Intermix FCP, FCIP, and FICON Environment*, SG24-6397
- *Introduction to SAN Distance Solutions*, SG24-6408
- *Introducing Hosts to the SAN Fabric*, SG24-6411
- *IP Storage Networking: IBM NAS and iSCSI Solutions*, SG24-6240
- *The IBM TotalStorage NAS Integration Guide*, SG24-6505
- *Implementing the IBM TotalStorage NAS 300G: High Speed Cross Platform Storage and Tivoli SANergy!*, SG24-6278
- *Using iSCSI Solutions' Planning and Implementation*, SG24-6291
- *IBM Storage Solutions for Server Consolidation*, SG24-5355
- *Implementing the Enterprise Storage Server in Your Environment*, SG24-5420
- *Implementing Linux with IBM Disk Storage*, SG24-6261
- *IBM Tape Solutions for Storage Area Networks and FICON*, SG24-5474
- *IBM Enterprise Storage Server*, SG24-5465
- *The IBM TotalStorage Solutions Handbook*, SG24-5250

The following publications from IBM Redbooks provide additional information about IBM Flex System. These are available from:

http://www.redbooks.ibm.com/portals/puresystems

- *IBM PureFlex System and IBM Flex System Products & Technology*, SG24-7984
- IBM *Flex System p260 and p460 Panning and Implementation Guide*, SG24-7989
- *IBM Flex System Networking in an Enterprise Data Center*, REDP4834

Chassis and Compute Nodes:

- *IBM Flex System Enterprise Chassis*, TIPS0863
- *IBM Flex System p260 and p460 Compute Node*, TIPS0880
- *IBM Flex System x240 Compute Node*, TIPS0860
- *IBM Flex System Manager*, TIPS0862

Switches:

- *IBM Flex System EN2092 1Gb Ethernet Scalable Switch*, TIPS0861

- *IBM Flex System Fabric EN4093 10Gb Scalable Switch*, TIPS0864
- *IBM Flex System EN4091 10Gb Ethernet Pass-thru Module*, TIPS0865
- *IBM Flex System FC5022 16Gb SAN Scalable Switch and FC5022 24-port 16Gb ESB SAN Scalable Switch*, TIPS0870
- *IBM Flex System IB6131 InfiniBand Switch*, TIPS0871
- *IBM Flex System FC3171 8Gb SAN Switch and Passthru*, TIPS0866

Adapters:

- *IBM Flex System EN2024 4-port 1Gb Ethernet Adapter*, TIPS0845
- *IBM Flex System FC5022 2-port 16Gb FC Adapter*, TIPS0891
- *IBM Flex System CN4054 10Gb Virtual Fabric Adapter and EN4054 4-port 10Gb Ethernet Adapter*, TIPS0868
- *IBM Flex System FC3052 2-port 8Gb FC Adapter*, TIPS0869
- *ServeRAID M5115 SAS/SATA Controller for IBM Flex System*, TIPS0884
- *IBM Flex System IB6132 2-port FDR InfiniBand Adapter*, TIPS0872
- *IBM Flex System EN4132 2-port 10Gb Ethernet Adapter*, TIPS0873
- *IBM Flex System IB6132 2-port QDR InfiniBand Adapter*, TIPS0890
- *IBM Flex System FC3172 2-port 8Gb FC Adapter*, TIPS0867

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

IBM Flex System education

The following are IBM educational offerings for IBM Flex System. Note that some course numbers and titles might have changed slightly after publication.

Note: IBM courses prefixed with NGTxx are traditional, face-to-face classroom offerings. Courses prefixed with NGVxx are Instructor Led Online (ILO) offerings. Courses prefixed with NGPxx are Self-paced Virtual Class (SPVC) offerings.

- *NGT10/NGV10/NGP10,* IBM Flex System - Introduction
- *NGT20/NGV20/NGP20,* IBM Flex System x240 Compute Node
- *NGT30/NGV30/NGP30,* IBM Flex System p260 and p460 Compute Nodes
- *NGT40/NGV40/NGP40,* IBM Flex System Manager Node
- *NGT50/NGV50/NGP50,* IBM Flex System Scalable Networking

For more information on these, and many other IBM System x educational offerings, visit the global IBM Training website located at:

http://www.ibm.com/training

Referenced Web sites

These Web sites are also relevant as further information sources:

- IBM TotalStorage hardware, software and solutions:
  http://www.storage.ibm.com
- IBM System Storage Storage Area Networks:
  http://www-03.ibm.com/servers/storage/san/
- Brocade:
  http://www.brocade.com
- Cisco:
  http://www.cisco.com
- QLogic:
  http://www.qlogic.com
- Emulex:
  http://www.emulex.com
- Finisar:
  http://www.finisar.co
- Tivoli:
  http://www.tivoli.com
- IEEE:
  http://www.ieee.org
- Storage Networking Industry Association:
  http://www.snia.org
- Fibre Channel Industry Association:
  http://www.fibrechannel.com
- SCSI Trade Association:
  http://www.scsita.org
- Internet Engineering Task Force:
  http://www.ietf.org
- American National Standards Institute:
  http://www.ansi.org
- Technical Committee T10:
  http://www.t10.org
- Technical Committee T11:
  http://www.t11.org

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

Numerics

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Z

(1.5" spine)
1.5"<-> 1.998"
789 <->1051 pages

(1.0" spine)
0.875"<->1.498"
460 <-> 788 pages

(0.5" spine)
0.475"<->0.873"
250 <-> 459 pages

(0.2"spine)
0.17"<->0.473"
90<->249 pages

(0.1"spine)
0.1"<->0.169"
53<->89 pages

To determine the spine width of a book, you divide the paper PPI into the number of pages in the book. An example is a 250 page book using Plainfield opaque 50# smooth which has a PPI of 526. Divided 250 by 526 which equals a spine width of .4752". In this case, you would use the .5” spine. Now select the Spine width for the book and hide the others: **Special>Conditional Text>Show/Hide>SpineSize(-->Hide:)>Set** . Move the changed Conditional text settings to all files in your book by opening the book file with the spine.fm still open and **File>Import>Formats** the Conditional Text Settings (ONLY!) to the book files.

(2.5” spine)
2.5”<->nnn.n”
1315<-> nnnn pages

(2.0” spine)
2.0” <-> 2.498”
1052 <-> 1314 pages

Introduction to Storage Area Networks and System Networking

Learn basic SAN and System Networking concepts

Introduce yourself to the business benefits

Discover the IBM System Networking portfolio

The plethora of data created by the businesses of today is making storage a strategic investment priority for companies of all sizes. As storage takes precedence, three major initiatives have emerged:

Flatten and converge your network
IBM takes an open, standards-based approach to implement the latest advances in today's flat, converged data center network designs. IBM System Networking solutions enable clients to deploy a high-speed, low-latency Unified Fabric Architecture.

Optimize and automate virtualization
Advanced virtualization awareness reduces the cost and complexity of deploying physical and virtual data center infrastructure.

Simplify management
IBM data center networks are easy to deploy, maintain, scale and virtualize, delivering the foundation of consolidated operations for dynamic infrastructure management.

Storage is no longer an afterthought. Too much is at stake. Companies are searching for more ways to efficiently manage expanding volumes of data, and to make that data accessible throughout the enterprise; this is propelling the move of storage into the network. Also, the increasing complexity of managing large numbers of storage devices and vast amounts of data is driving greater business value into software and services.

Welcome to the era of Smarter Networking for Smarter Data Centers!

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

SG24-5470-04 ISBN