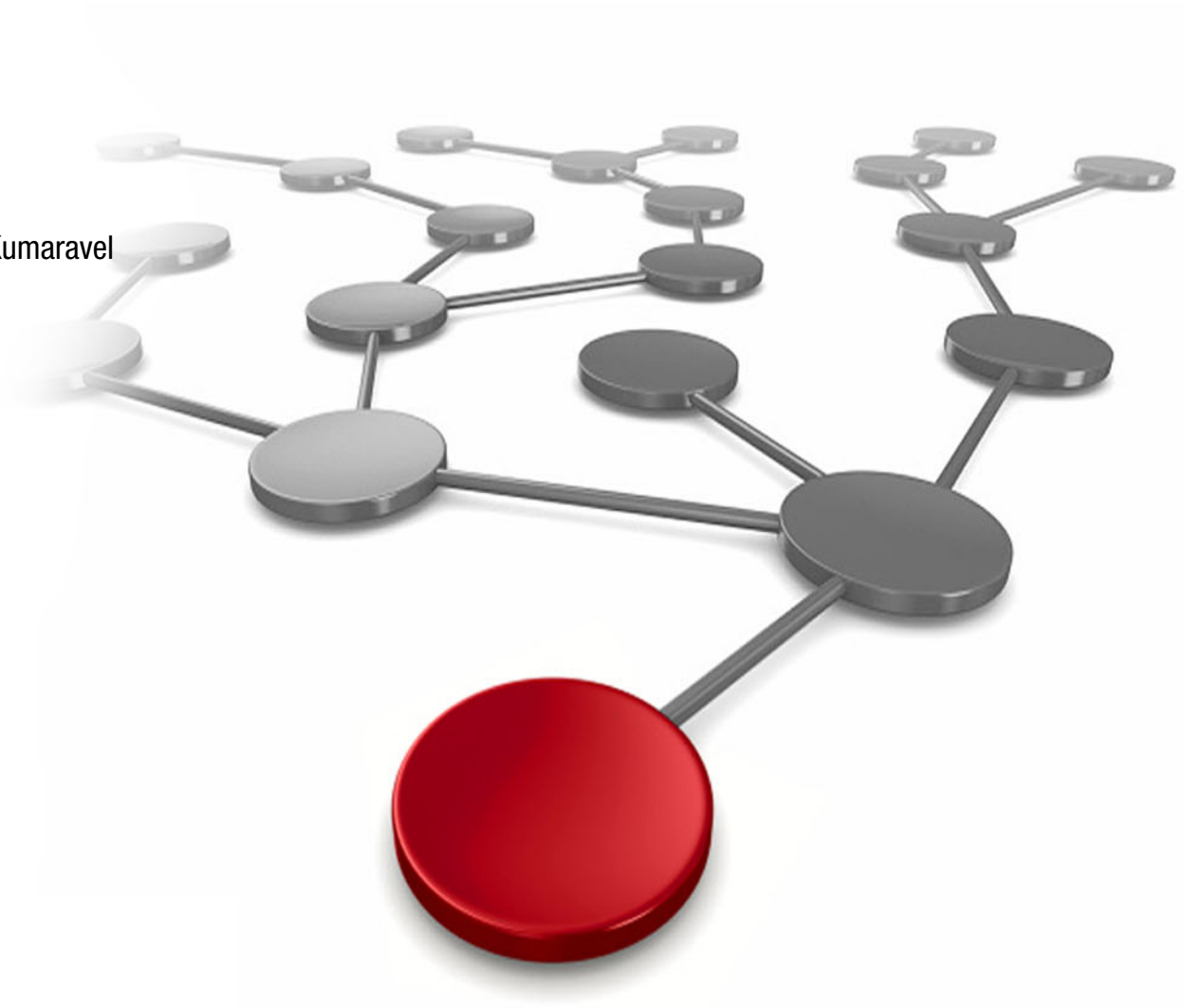
Redbooks
ibm.com/redbooks

Introduction to Storage Area Networks

Jon Tate
Pall Beck
Hector Hugo Ibarra
Shanmuganathan Kumaravel
Libor Miklas



IBM

Redbooks

International Technical Support Organization

Introduction to Storage Area Networks

December 2017

SG24-5470-08

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

Ninth Edition (December 2017)

This edition applies to the products in the IBM Storage Area Networks (SAN) portfolio.

Contents

Notices ix
Trademarks x

Preface xi
Authors xii
Now you can become a published author, too! xiv
Comments welcome xiv
Stay connected to IBM Redbooks xiv

Summary of changes xv
December 2017, Ninth Edition xv

Chapter 1. Introduction 1
1.1 Networks 2
1.1.1 The importance of communication 2
1.2 Interconnection models 2
1.2.1 The open systems interconnection model 2
1.2.2 Translating the OSI model to the physical world 4
1.3 Storage 5
1.3.1 Storing data 5
1.3.2 Redundant Array of Independent Disks 6
1.4 Storage area networks 11
1.5 Storage area network components 13
1.5.1 Storage area network connectivity 14
1.5.2 Storage area network storage 14
1.5.3 Storage area network servers 14
1.6 The importance of standards or models 14

Chapter 2. Storage area networks 17
2.1 Storage area networks 18
2.1.1 The problem 18
2.1.2 Requirements 19
2.2 Using a storage area network 20
2.2.1 Infrastructure simplification 20
2.2.2 Information lifecycle management 21
2.2.3 Business continuity 22
2.3 Using the storage area network components 22
2.3.1 Storage 22
2.3.2 Storage area network connectivity 23
2.3.3 Servers 28
2.3.4 Putting the components together 32

Chapter 3. Fibre Channel internals 33
3.1 Fibre Channel architecture 34
3.1.1 Small Computer Systems Interface 34
3.1.2 Limitations of the Small Computer System Interface 35
3.1.3 Fibre Channel advantages 38
3.2 Layers 40
3.3 Optical cables 43

3.3.1 Attenuation 43
3.3.2 Maximum power 43
3.3.3 Fiber in the storage area network 44
3.3.4 Dark fiber 49
3.4 Classes of service 49
3.4.1 Class 1 50
3.4.2 Class 2 50
3.4.3 Class 3 50
3.4.4 Class 4 50
3.4.5 Class 5 50
3.4.6 Class 6 51
3.4.7 Class F 51
3.5 Fibre Channel data movement 51
3.5.1 Byte-encoding schemes 52
3.6 Data transport 54
3.6.1 Ordered set 54
3.6.2 Frames 55
3.6.3 Sequences 57
3.6.4 Exchanges 57
3.6.5 In order and out of order 58
3.6.6 Latency 59
3.6.7 Open fiber control 59
3.7 Flow control 60
3.7.1 Buffer to buffer 60
3.7.2 End to end 60
3.7.3 Controlling the flow 60
3.7.4 Performance 60

Chapter 4. Ethernet and system networking concepts 63
4.1 Ethernet 64
4.1.1 Shared media 64
4.1.2 Ethernet frame 65
4.1.3 How Ethernet works 65
4.1.4 Speed and bandwidth 67
4.1.5 10 GbE 68
4.1.6 10 GbE copper versus fiber 68
4.1.7 Virtual local area network 71
4.1.8 Interface virtual local area network operation modes 73
4.1.9 Link aggregation 75
4.1.10 Spanning Tree Protocol 75
4.1.11 Link Layer Discovery Protocol 78
4.1.12 Link Layer Discovery Protocol Type Length Values (LLDP TLVs) 78
4.2 Storage area network IP networking 80
4.2.1 The multiprotocol environment 80
4.2.2 Fibre Channel switching 80
4.2.3 Fibre Channel routing 80
4.2.4 Tunneling 80
4.2.5 Routers and gateways 81
4.2.6 Internet Storage Name Service 81
4.3 Delving deeper into the protocols 81
4.3.1 Fibre Channel over Internet Protocol (FCIP) 81
4.3.2 Internet Fibre Channel Protocol 82
4.3.3 Internet Small Computer System Interface 83

4.3.4 Routing considerations 85
4.3.5 Packet size 85
4.3.6 TCP congestion control 85
4.3.7 Round-trip delay 86
4.4 Multiprotocol solution briefs 87
4.4.1 Dividing a fabric into subfabrics 87
4.4.2 Connecting a remote site over IP 88
4.4.3 Connecting hosts by using Internet Small Computer System Interface 88

Chapter 5. Topologies and other fabric services 89
5.1 Fibre Channel topologies 90
5.1.1 Point-to-point topology 90
5.1.2 Arbitrated loop topology 91
5.1.3 Switched fabric topology 92
5.1.4 Single switch topology 93
5.1.5 Cascaded and ring topology 94
5.1.6 Mesh topology 95
5.1.7 Core-edge topology 96
5.1.8 Edge-core-edge topology 96
5.2 Port types 97
5.2.1 Common port types 97
5.2.2 Expansion port types 98
5.2.3 Diagnostic port types 99
5.3 Addressing 101
5.3.1 Worldwide name 101
5.3.2 Tape device worldwide node name and worldwide port name 105
5.3.3 Port address 106
5.3.4 The 24-bit port address 106
5.3.5 Loop address 108
5.3.6 The b-type addressing modes 108
5.3.7 FICON address 110
5.4 Fibre Channel Arbitrated Loop protocols 114
5.4.1 Fairness algorithm 115
5.4.2 Loop addressing 115
5.5 Fibre Channel port initialization and fabric services 116
5.5.1 Fabric login (FLOGI) 116
5.5.2 Port login (PLOGI) 117
5.5.3 Process login (PRLI) 118
5.6 Fabric services 119
5.6.1 Management server 120
5.6.2 Time server 120
5.6.3 Simple name server 120
5.6.4 Fabric login server 120
5.6.5 Registered state change notification service 120
5.7 Routing mechanisms 121
5.7.1 Spanning tree 121
5.7.2 Fabric shortest path first 121
5.8 Zoning 122
5.8.1 Hardware zoning 123
5.8.2 Software zoning 126
5.8.3 Logical unit number masking 128

Chapter 6. Storage area network as a service for cloud computing 129

6.1 The cloud 130
6.1.1 Private and public cloud 131
6.1.2 Cloud computing components 131
6.1.3 Cloud computing models 132
6.2 Virtualization and the cloud 134
6.2.1 Cloud infrastructure virtualization 135
6.2.2 Cloud platforms 135
6.2.3 Storage virtualization 138
6.3 SAN virtualization 139
6.3.1 IBM b-type Virtual Fabrics 139
6.3.2 Cisco virtual storage area network 141
6.3.3 N_Port ID Virtualization 142
6.4 Building a smarter cloud 144
6.4.1 Automated tiering 144
6.4.2 Thin provisioning 145
6.4.3 Data deduplication 146
6.4.4 New generation management tools 149
6.4.5 Business continuity and disaster recovery 149
6.4.6 Storage on demand 149

Chapter 7. Fibre Channel products and technology 151
7.1 The environment 152
7.2 Storage area network devices 153
7.2.1 Fibre Channel bridges 154
7.2.2 Arbitrated loop hubs and switched hubs 154
7.2.3 Switches and directors 156
7.2.4 Multiprotocol routing 156
7.2.5 Service modules 157
7.2.6 Multiplexers 157
7.3 Components 157
7.3.1 Application-specific integrated circuit 157
7.3.2 Fibre Channel transmission rates 158
7.3.3 SerDes 158
7.3.4 Backplane and blades 158
7.4 Gigabit transport technology 159
7.4.1 Fibre Channel cabling 159
7.4.2 Transceivers 163
7.4.3 Host bus adapters 165
7.5 Inter-switch links 167
7.5.1 Cascading 168
7.5.2 Hops 168
7.5.3 Fabric shortest path first 168
7.5.4 Non-blocking architecture 170
7.5.5 Latency 171
7.5.6 Oversubscription 171
7.5.7 Congestion 171
7.5.8 Trunking or port-channeling 172

Chapter 8. Management 173
8.1 Management principles 174
8.1.1 Management types 174
8.1.2 Connecting to storage area network management tools 176
8.1.3 Storage area network fault isolation and troubleshooting 176

8.2 Management interfaces and protocols 177
8.2.1 Storage Networking Industry Association initiative 177
8.2.2 Simple Network Management Protocol 179
8.2.3 Service Location Protocol 180
8.2.4 Vendor-specific mechanisms 180
8.3 Management features 182
8.3.1 Operations 183
8.4 Vendor management applications 183
8.4.1 Storage Networking SAN b-type 184
8.4.2 Cisco 185
8.5 SAN multipathing software 187

Chapter 9. Security 193
9.1 Security in the storage area network 194
9.2 Security principles 195
9.2.1 Access control 195
9.2.2 Auditing and accounting 195
9.2.3 Data security 195
9.2.4 Securing a fabric 196
9.2.5 Zoning, masking, and binding 197
9.3 Data security 198
9.4 Storage area network encryption 198
9.4.1 Basic encryption definition 198
9.4.2 Data-in-flight 200
9.4.3 Data-at-rest 201
9.4.4 Digital certificates 201
9.4.5 Encryption algorithm 201
9.4.6 Key management considerations and security standards 202
9.4.7 b-type encryption methods 203
9.4.8 Cisco encryption methods 205
9.5 Encryption standards and algorithms 207
9.6 Security common practices 208

Chapter 10. Solutions 209
10.1 Introduction 210
10.2 Basic solution principles 210
10.2.1 Connectivity 210
10.2.2 Adding capacity 211
10.2.3 Data movement and copy 211
10.2.4 Upgrading to faster speeds 214
10.3 Infrastructure simplification 215
10.3.1 The origin of the complexity 216
10.3.2 Storage pooling 216
10.3.3 Consolidation 219
10.3.4 Migration to a converged network 221
10.4 Business continuity and disaster recovery 225
10.4.1 Clustering and high availability 225
10.4.2 LAN-free data movement 227
10.4.3 Disaster backup and recovery 228
10.5 Information lifecycle management 229
10.5.1 Information lifecycle management 230
10.5.2 Tiered storage management 230
10.5.3 Long-term data retention 232

10.5.4 Data lifecycle management . . . 232
10.5.5 Policy-based archive management . . . 233

Chapter 11. Storage area networks and green data centers . . . 235
11.1 Data center constraints . . . 236
11.1.1 Energy flow in the data center . . . 237
11.2 Data center optimization . . . 238
11.2.1 Strategic considerations . . . 239
11.3 Green storage . . . 239
11.3.1 Information lifecycle management . . . 240
11.3.2 Storage consolidation and virtualization . . . 241
11.3.3 On-demand storage provisioning . . . 243
11.3.4 Hierarchical storage and tiering . . . 243
11.3.5 Data compression and data deduplication . . . 245

Chapter 12. IBM Fibre Channel storage area network product portfolio . . . 247
12.1 Classification of IBM SAN products . . . 248
12.2 SAN Fibre Channel networking . . . 248
12.3 Entry SAN switches . . . 249
12.3.1 IBM Storage Networking SAN24B-6 . . . 249
12.3.2 IBM System Networking SAN24B-5 . . . 250
12.3.3 IBM System Storage SAN24B-4 Express . . . 251
12.4 Mid-range SAN switches . . . 252
12.4.1 Cisco MDS 9396S 16G Multilayer Fabric Switch . . . 252
12.4.2 IBM System Networking SAN96B-5 . . . 253
12.4.3 IBM Storage Networking SAN64B-6 . . . 255
12.4.4 IBM System Storage SAN48B-5 . . . 255
12.4.5 Cisco MDS 9148S 16G Multilayer Fabric Switch for IBM System Storage . . . 256
12.5 Enterprise SAN directors . . . 257
12.5.1 IBM Storage Networking SAN512B-6 and SAN256B-6 . . . 257
12.5.2 Cisco MDS 9718 Multilayer Director . . . 259
12.5.3 Cisco MDS 9710 Multilayer Director . . . 261
12.5.4 IBM System Storage SAN384B-2 and SAN768B-2 . . . 263
12.5.5 Cisco MDS 9706 Multilayer Director for IBM System Storage . . . 266
12.6 Extension switches . . . 268
12.6.1 IBM System Storage SAN42B-R . . . 268
12.6.2 Cisco MDS 9250i Multiservice Fabric Switch . . . 269
12.6.3 IBM System Storage SAN06B-R . . . 270

Chapter 13. Certification . . . 273
13.1 The importance of certification . . . 274
13.2 IBM Professional Certification Program . . . 274
13.2.1 About the program . . . 274
13.2.2 Certifications by product . . . 275
13.2.3 Mastery tests . . . 275
13.3 Storage Networking Industry Association certifications . . . 275
13.4 Brocade certification . . . 276
13.5 Cisco certification . . . 276
13.6 Open Group certification . . . 276

Related publications . . . 277
IBM Redbooks . . . 277
Online resources . . . 277
Help from IBM . . . 278

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

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 grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

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 jurisdictions 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 provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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 actual people or business enterprises 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. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

AFS™
AIX®
DB2®
Domino®
DS8000®
Easy Tier®
ECKD™
FICON®
GPFS™
HACMP™
HyperFactor®
IBM®
IBM z™
IBM z Systems®
Informix®
Insight™
Lotus®
OS/390®
Power Systems™
PowerHA®
ProtecTIER®
Redbooks®
Redbooks (logo)  ®
Storwize®
System Storage®
System z®
System z9®
Tivoli®
z Systems®
z/OS®
z9®

The following terms are trademarks of other companies:

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

Microsoft, 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.

Preface

The superabundance of data that is created by today's businesses is making storage a strategic investment priority for companies of all sizes. As storage takes precedence, the following major initiatives emerge:

- Flatten and converge your network: IBM® takes an open, standards-based approach to implement the latest advances in the flat, converged data center network designs of today. IBM Storage 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 demand 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% each year, this outlook is where a storage area network (SAN) enters the arena. SANs are the leading storage infrastructure for the global economy of today. SANs offer simplified storage management, scalability, flexibility, and availability; and improved data access, movement, and backup.

Welcome to the cognitive era.

The smarter data center with the improved economics of IT can be achieved by connecting servers and storage with a high-speed and intelligent network fabric. A smarter data center that hosts IBM Storage solutions can provide an environment that is smarter, faster, greener, open, and easy to manage.

This IBM Redbooks® publication provides an introduction to SAN and Ethernet networking, and how these networks help to achieve a smarter data center. This book is intended for people who are not very familiar with IT, or who are just starting out in the IT world.

Also, be sure to see the IBM Storage Redbooks portal for the latest material from the International Technical Support Organization (ITSO):

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

Authors

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 (ITSO), San Jose Center. Before Jon joined the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2 support for IBM storage products. Jon has 32 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, at the time of writing, was a SAN Technical Lead in IBM Nordic. He has 18 years of experience working with storage, both for dedicated clients and for large shared environments. Those environments include clients from the medical and financial sector, which includes several of the largest shared SAN environments in Europe. He is a member of the SAN and SAN Volume Controller best practices community in the Nordics and in EMEA. In his current job role, he is a member of the Solutions Consultant Express+ (SCE+) Storage Deployment Specialists, responsible for SCE+ storage deployments around the globe. Pall is also a member of a team that helps in critical situations and performs root cause analyzes. He is coauthor of the *Implementing SVC 5.1* and *SVC Advanced Copy Services 4.2* Redbooks publications. Pall has a diploma as an Electronic Technician from Odense Tekniske Skole in Denmark and an IR in Reykjavik, Iceland. He is also an IBM Certified IT Specialist.

Since writing this book, Pall has left IBM.

Hector Hugo Ibarra at the time of writing, was an Infrastructure IT Architect who specialized in cloud computing and storage solutions for IBM Argentina. Hector was the ITA Leader for The VMware Center of Competence and he specialized in virtualization technologies and assisted global IBM clients in deploying virtualized infrastructures across the world. Since 2009, he worked as the Leader for the Argentina Delivery Center Strategy and Architecture Services department, where major projects are driven.

Since writing this book, Hector has left IBM.

Shanmuganathan Kumaravel is an IBM Technical Services Specialist for the ITD-SSO MR Storage team of IBM India. He has supported SAN and disk products from both IBM and Hewlett Packard since August 2008. Previously, Shan worked for HP product support, which provides remote support for 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 Systems Engineer (CSE).

Libor Miklas at the time of writing, was a Team Leader and an experienced IT Specialist, who worked at the IBM Delivery Center Central Europe in the Czech Republic. He has over 17 years of practical experience within the IT industry. For the last ten years, he has focused on backup and recovery and storage management. Libor and his team supported midrange and enterprise storage environments for various global and local clients, worldwide. He is an IBM Certified Deployment Professional of the IBM Tivoli® Storage Manager family of products and holds a Masters Degree in Electrical Engineering and Telecommunications.

Since writing this book, Libor has left IBM.

Thanks to the International Technical Support Organization, San Jose Center, for their contributions to this project.

Thanks to the authors of the previous 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
IBM

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

Jim Baldyga
Silviano Gaona
Brian Steffler
Marcus Thordal
Steven Tong
Brocade Communications Systems (a Broadcom Limited Company)

Special thanks to 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:

  https://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

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 SG24-5470-08
for Introduction to Storage Area Networks
as created or updated on October 9, 2018.

December 2017, Ninth Edition

This revision includes the following new and changed information.

New information

Chapter 12 contains the current IBM SAN portfolio as of the time of the update of this book (December 2017). Editorial changes and corrections are included.

Changed information

Products that were removed from the IBM portfolio are deleted from the book.

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 that are running 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 (SAN), which is regarded as the ultimate response to all of these needs.

1.1 Networks

A computer network, often simply called a *network*, is a collection of computers and devices that are interconnected by communication channels. These channels allow for the efficient sharing of resources, services, and information among the network.

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

1.1.1 The importance of communication

It is impossible to imagine the human world as stand-alone human beings, with nobody that talks or does anything for each other. Much more importantly, it is hard trying to imagine how a human being can work without using their senses. In our human world, we are sure you will agree with us that communication between individuals makes a significant difference in all aspects of life.

First of all, communication in any form is not easy, and we need many components. Factors consist of a common language, something to be communicated, a medium where the communication flows, and finally we need to be sure that whatever was communicated was received and understood. To do that in the human world, we use language as a communication protocol, and sounds and writing are the communication media.

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

A wealth of information is devoted to networking history and its evolution, and we do not intend to give a history lesson in this book. This publication focuses on the prevalent interconnection models, storage, and networking concepts.

1.2 Interconnection models

An interconnection model is a standard that is used to connect sources and targets in a network. Several well-known models in the IT industry are the open systems interconnection model (OSI), Department of Defense (DoD), TCP/IP protocol suite, and Fibre Channel (FC). Each model has its advantages and disadvantages. Its model is applied where it has the maximum benefit in terms of performance, reliability, availability, cost benefits, and so on.

1.2.1 The open systems interconnection model

The open systems interconnection model (OSI model) was a product of the open systems interconnection effort at the International Organization for Standardization (ISO). It is a way of subdividing a communications system into smaller parts that are called *layers*. Similar communication functions are grouped into logical layers. A layer provides services to its upper layer while it receives 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 this book, we focus on the Physical, DataLink, 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 relationship includes the layout of pins, voltages, cable specifications, and more.

Layer 2: DataLink Layer

The *DataLink Layer* provides the functional and procedural means to transfer data between network entities. This layer also detects and possibly corrects errors that might 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 another network. The Network Layer provides this functionality while it maintains the quality of service that is requested by the Transport Layer (in contrast to the DataLink Layer, which connects hosts within the same network). The Network Layer performs network routing functions. This layer might also perform fragmentation and reassembly, and report delivery errors. Routers operate at this layer by sending data throughout the extended network and making the Internet access possible.

Layer 4: Transport Layer

The *Transport Layer* provides transparent transfer of data between users, providing reliable data transfer services to the upper layers. The Transport Layer controls the reliability of a specific link through flow control, segmentation and desegmentation, and error control. Certain protocols are state-oriented and connection-oriented, which means that the Transport Layer can track the segments and retransmit the segments that fail. The Transport Layer also provides acknowledgment of successful data transfer.

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. Figure 1-1 shows a comparison table of various models.

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 Comparison table of OSI, TCP/IP, and FC models

The Fibre Channel model is covered later in this book.

1.2.2 Translating the OSI model to the physical world

To translate from theoretical models to reality, we introduce physical devices that perform certain tasks for each layer on each model.

Local area networks (*LANs*) are a good place to start. We define LANs as a small or large network that is limited within the same physical site. This site might be a traditional office or a corporate building.

In Figure 1-2, you see a basic network where computers and a printer are interconnected by using physical cables and interconnection devices.

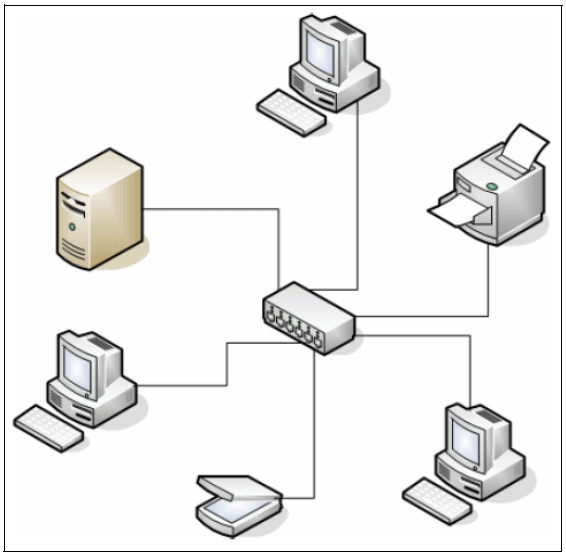


Figure 1-2 Basic network topology

We must keep in mind that any model we choose defines the devices, cables, connectors, and interface characteristics that we must implement to make it work. We must also support the protocols for each model layer.

All of the network components are categorized into five groups:

- End devices: An end device is a computer system that has a final purpose, such as desktop computers, printers, storage, or servers.
- Network interface: A network interface is an interface between the media and end devices that can interact with other network interfaces and understands an interconnection model.
- Connector: The connector is the physical element at the end of the media that allows a connection to the network interface.
- Media: Media is the physical path that is used to transmit an electrical or optical signal. It might be wired or wireless, copper, or a fiber optic cable.
- Network devices: These devices are used to interconnect multiple end devices as a single point of interconnection, route communication through separate networks, or provide network security. Examples of network devices are switches, routers, firewalls, and directors.

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 Storage

To understand what storage is, and because understanding it is a key point for this book, we start from a basic *hard disk drive* (*HDD*). We then progress through to storage systems that are high performance, fault tolerant, and highly available. During this explanation, instructional examples are used that might sometimes not reflect reality. However, the basic examples make it easier to understand for individuals who are just entering the world of storage systems.

Note: We are aware that solid-state drives (SSD) and Flash arrays form an important part of any data center today, but for this basic example, we use the HDD as our building block.

1.3.1 Storing data

Data is stored on HDDs on which data can be read and written. Depending on the methods that are used to run those tasks, and the HDD technology on which the HDDs were built, the read and write function can be faster or slower. The evolution of HDDs is incredible. We can store hundreds of gigabytes on a single HDD, which allows us to keep all of the data we can ever imagine. Even though this approach seems to bring us only advantages so far, one question might be what happens if for any reason we are unable to access the HDD?

The first solution might be to have a secondary HDD where we can manually copy our primary HDD to our secondary HDD. Immediately, we can see that our data is safe. But, how often must 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, must we copy the entire amount of data from one HDD to the other, or must we copy only what changes?

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

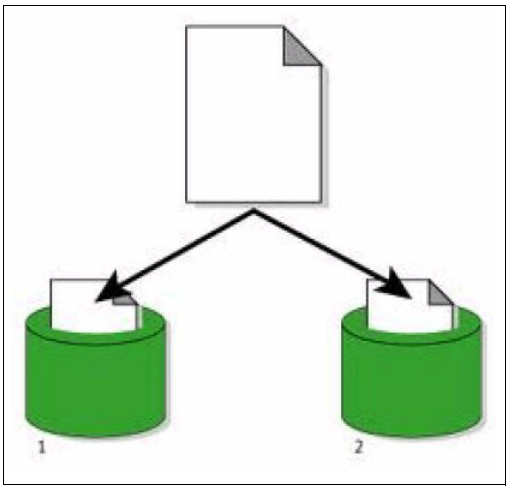


Figure 1-3 Manual copy of data

1.3.2 Redundant Array of Independent Disks

Fortunately, technology exists that can help us. That technology 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 that it changes to provide us with a reliable fault tolerant system. It is also clear that it cannot be done in a manual way. A *RAID controller* can maintain disks in synchronization and can also manage all of the writes and reads (input/output (I/O)) to and from the disks.

Figure 1-4 shows a diagram of our RAID system. The A, B, and C values in the diagram represent user data, such as documents or pictures.

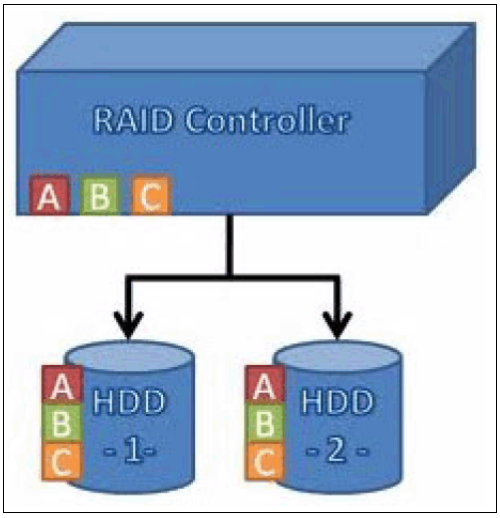


Figure 1-4 Typical RAID scenario

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

Stand-alone disks provide the following advantages:

- Redundancy to disk failure
- Faster reading of data because the data can be read from either disk

Stand-alone disks provide the following disadvantages:

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

Is any other RAID type available that can improve things further, while it conserves the advantages and removes the disadvantages of RAID 1? Yes, and this type of RAID is known as *RAID 5*. This scenario consists of dividing the user data into *N*-1 parts (where *N* is the number of disks that are used to build the RAID) and then calculating a parity part. This parity part permits RAID to rebuild the user data if a disk failure occurs.

RAID 5 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 of the disks. If a disk fails, the RAID requires a rebuild and the parity information is used to re-create the lost data. Figure 1-5 shows this example.

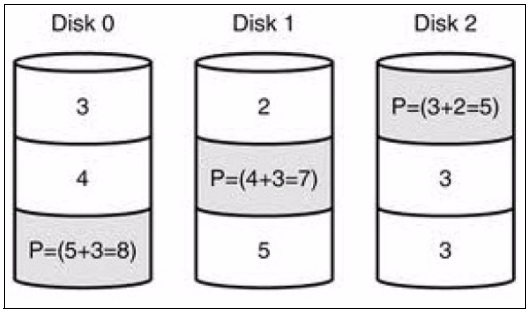


Figure 1-5 Example of RAID 5 with parity

RAID 5 requires a minimum of three disks. In theory, no limitations exist to add disks. This RAID type combines data safety with the 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 by using the *hot* spare. However, the failed disk must be replaced as soon as possible. A *spare disk* is an empty disk that is used by the RAID controller only when a disk fails. Figure 1-6 shows this example.

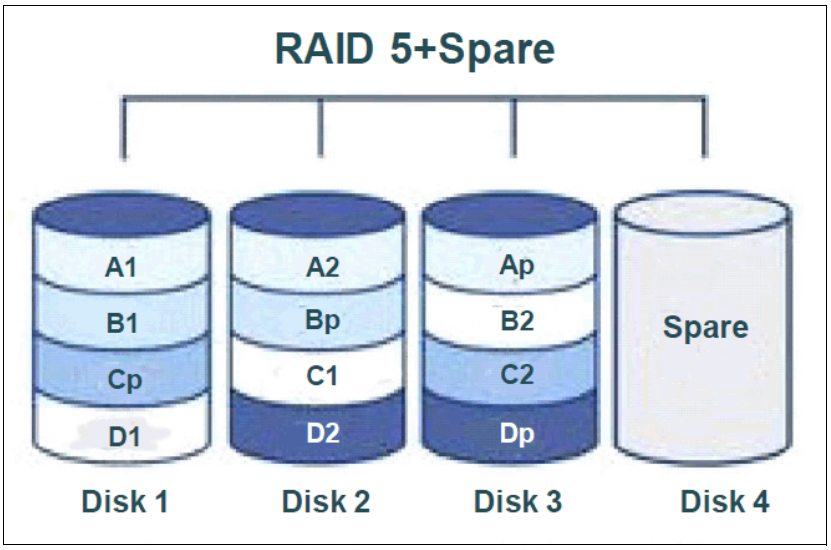


Figure 1-6 RAID 5 with a hot spare

RAID 5 has better performance for I/O than RAID 1. Depending on the number of disks that are used to build the RAID, the array disk space utilization is more than two-thirds. RAID 5 is also managed by a RAID controller that performs the same role as in RAID 1.

Table 1-1 shows a brief comparison among the most common RAID levels.

RAID types: RAID 1 and RAID 5 are the most common RAID levels. However, many other levels are not covered in this book. Levels that are not described include RAID 0, 3, 4, and 6; or nested (hybrid) types, such as RAID 0+1 or RAID 5+1. These hybrid types are used in environments where reliability and performance are key points to be covered from the storage perspective.

Table 1-1 RAID level comparison

Level	Description	Minimum number of drives	Fault tolerance
RAID 0	Block-level striping without parity or mirroring	2	None
RAID 1	Mirroring without parity or striping	2	n - 1 drive failures
RAID 2	Bit-level striping with Hamming code for error correction	3	One drive failure
RAID 3	Byte-level striping with dedicated parity	3	One drive failure
RAID 4	Block-level striping with dedicated parity	3	One drive failure
RAID 5	Block-level striping with distributed parity	3	One drive failure
RAID 6	Block-level striping with double distributed parity	4	Two drive failures

Our disk systems seem to be ready to support failures, and they are also high performance. But what if our RAID controller fails? We might not lose data, but the data is not accessible. Is a solution available to access this data?

It is almost the same scenario that we initially faced with only one disk as a storage system. This type of scenario is known as a *single point of failure* (*SPOF*). We must add redundancy by introducing a secondary RAID controller to our storage system.

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

RAID controller role: The RAID controller role in certain cases is performed by the software. This solution is less expensive than a hardware solution because it does not require controller hardware; however, it is a slower solution.

We now have several physical HDDs that are managed by two controllers.

Disk pools

When a logical storage volume needs to be provisioned to servers, first the storage RAID needs to be created. To create the RAID, select the available HDDs and group them together for a single purpose. The number of grouped HDDs depends on the RAID type that we choose and the space that is required for provisioning.

To understand what is meant, a basic example is shown that uses the following assumptions:

- Ten HDDs, which are named A, B, C, D, E, F, G, H, I, and J, are included.
- Two RAID controllers, which are named RC1 and RC2, support any RAID level.
- Each RAID controller can manage any HDD.
- Each RAID controller can act as a backup of the other controller, at any time.

The following tasks can be performed:

- Select HDDs A, B, and F, and create a RAID 5 array that is managed by RC1. We call it G1.
- Select HDDs E, I, and J, and create a RAID 5 array that is managed by RC2. We call it G2.

Figure 1-7 shows these steps.

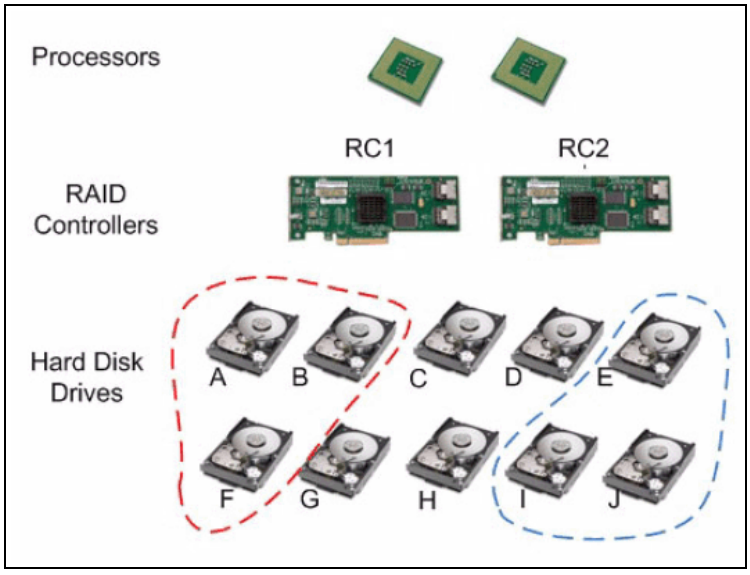


Figure 1-7 Disk pool creation

By issuing these simple steps, we create disk pools. These pools consist of grouping disks together for a single purpose, such as creating a RAID level, in our case, RAID 5.

In 1.3.2, “Redundant Array of Independent Disks” on page 6, we mentioned nested (hybrid) RAIDs, such as 5+0. These solutions are used when the amount of storage data is significant and important for business continuity. RAID 50 consists 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 subarray can fail without data loss.

Also, rebuild times are substantially shorter than the rebuild times of a single large RAID 5 array. See Figure 1-8.

Nested (hybrid) RAIDs: Nested or hybrid RAIDs are a combination of existing RAID levels that create a RAID to reap the benefits of two separate RAID levels.

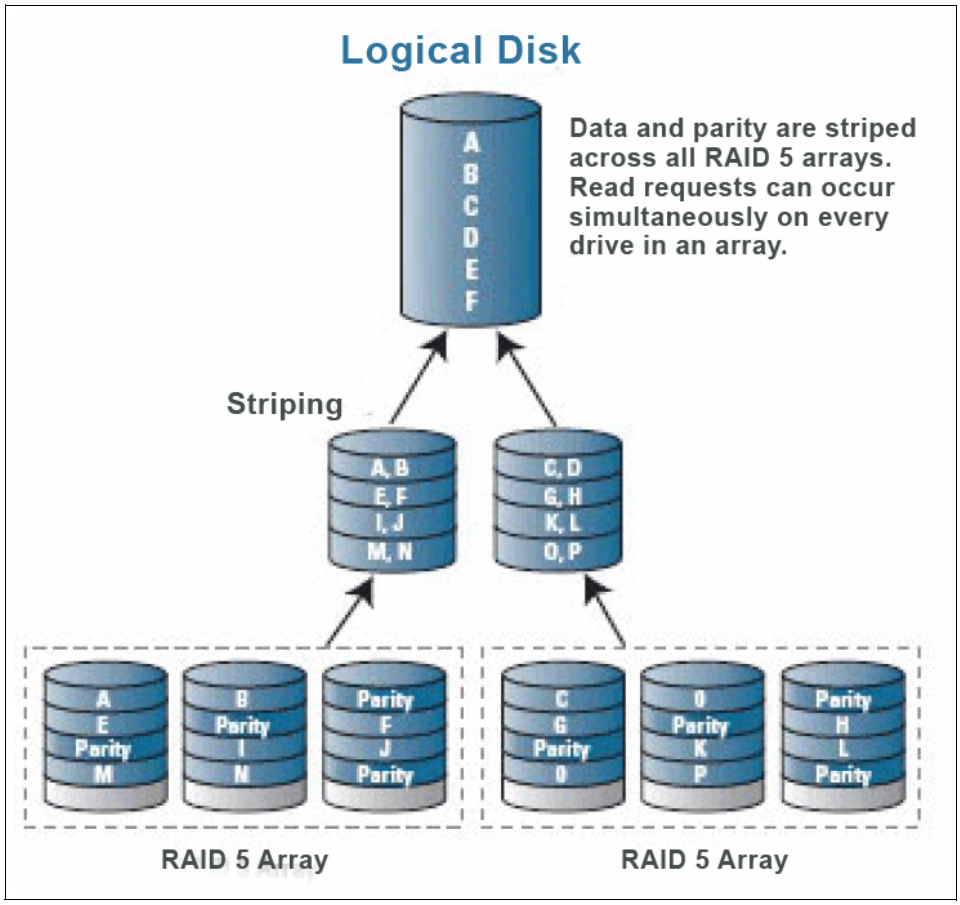


Figure 1-8 Nested (hybrid) RAID 5+0 or RAID 50

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

Storage systems

We are not 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 is a CPU that processes all of the required instructions to allow data to flow. Adding only one CPU creates a single point of failure (SPOF), so we add two CPUs.

We almost have an independent system. This system must be able to communicate with other systems in a network. Therefore, a minimum of two network interfaces are required to be able to avoid a SPOF.

Only one step is left. The last step is to put all of these hardware components into an enclosure. Figure 1-9 shows our storage system.

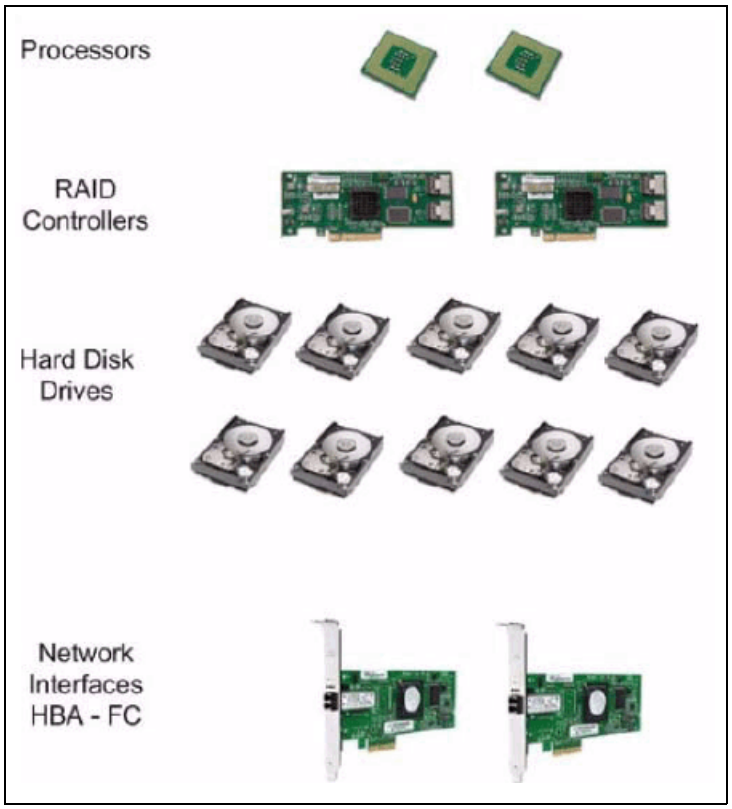


Figure 1-9 Basic storage system

This scenario is only an example: This basic storage configuration is presented as an example. However, a configuration can include as many CPUs, RAID controllers, network interfaces, and HDDs as needed.

1.4 Storage area networks

The Storage Networking Industry Association (SNIA) defines the *storage area network* (*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. It also includes a management layer, which organizes the connections, storage elements, and computer systems so that data transfer is secure and robust. The term *SAN* is typically (but not necessarily) identified with block I/O services rather than file access services.

In simple terms, a SAN is a specialized, high-speed network that attaches servers and storage devices. The SAN is sometimes referred to as the *network behind the servers*. A SAN allows an *any-to-any* connection across the network, by using interconnect elements, such as switches and directors. The SAN eliminates the traditional dedicated connection between a server and storage, and the concept that the server effectively *owns and manages* the storage devices.

The SAN also eliminates any restriction to the amount of data that a server can access. Traditionally, a server is limited by the number of storage devices that attach 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. A network might include many storage devices, including disk, tape, and optical storage. Additionally, the storage utility might be located far from the servers that it uses.

The SAN can be viewed as an extension to the storage *bus* concept. This concept enables storage devices and servers to interconnect by using similar elements, such as LANs and wide area networks (WANs).

The diagram in Figure 1-10 shows a tiered overview of a SAN that connects multiple servers to multiple storage systems.

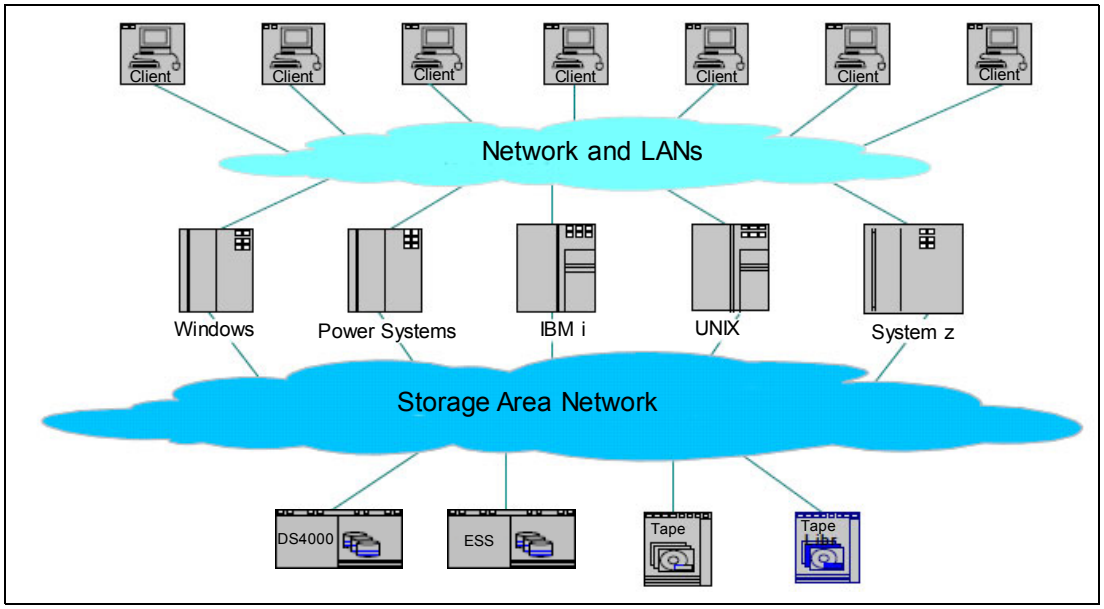


Figure 1-10 A SAN

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

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

- Server to storage: This method is the traditional model of interaction with storage devices. The advantage is that the same storage device might be accessed serially or concurrently by multiple servers.
- Server to server: A SAN might 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, therefore freeing up server processor cycles for other activities, such as application processing. Examples include a disk device that backs up its data to a tape device without server intervention, or a remote device mirroring across the SAN.

SANs allow applications that move data to perform better, for example, by sending data directly from the source device to the target device with minimal server intervention. SANs also enable new network architectures where multiple hosts access multiple storage devices that connect to the same network.

The use of 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 offloaded from servers and moved onto a separate network.
- Centralized and consolidated storage: Simpler management, scalability, flexibility, and availability are possible.
- Data transfer and vaulting to remote sites: A remote copy of data is enabled for disaster protection and against malicious attacks.
- Simplified centralized management: A single image of storage media simplifies management.

1.5 Storage area network components

Fibre Channel (FC) is the predominant architecture on which most SAN implementations are built. IBM Fibre Channel connection (FICON®) is the standard protocol for IBM z/OS® systems and Fibre Channel Protocol (FCP) is the standard protocol for open systems. The SAN components that are described in the following sections are FC-based (Figure 1-11).

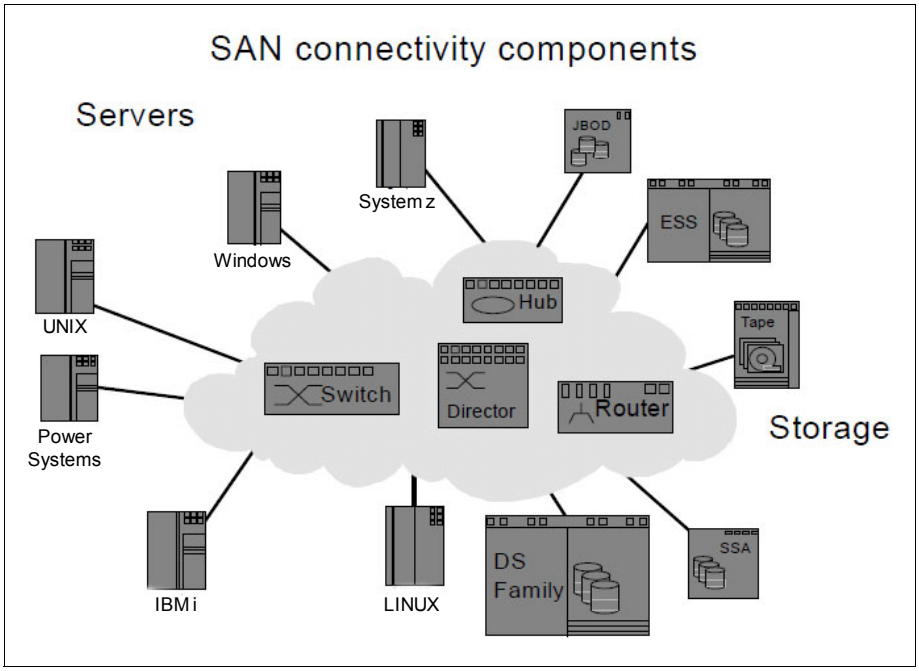


Figure 1-11 SAN components

1.5.1 Storage area network connectivity

The first element to consider in any SAN implementation is the connectivity of the storage and server components, which typically use FC. The components that are listed in Figure 1-11 on page 13 are typically used for LAN and WAN implementations. SANs, such as LANs, interconnect the storage interfaces together into many network configurations and across longer distances.

Much of the terminology that is used for SAN has its origins in Internet Protocol (IP) network terminology. In certain cases, the industry and IBM use different terms for the same thing and, in other cases, for different things.

1.5.2 Storage area network storage

The SAN liberates the storage device so that the storage device is not on a particular server bus and attaches it directly to the network. Storage is *externalized*. It can be functionally distributed across the organization. The SAN also enables the centralization of storage devices and the clustering of servers, which potentially can help you achieve easier and less expensive centralized administration that lowers the total cost of ownership (TCO).

The storage infrastructure is the foundation on which information relies. Therefore, the storage infrastructure must support the business objectives and business model of a company. 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.

Important: Remember that a good SAN begins with a good design. This statement is not only a maxim, but it must be your philosophy when you design or implement a SAN.

1.5.3 Storage area network servers

The server infrastructure is the underlying reason for all SAN solutions. This infrastructure includes a mix of server platforms, such as Microsoft Windows, UNIX (and its various versions), and z/OS. With initiatives, such as server consolidation and Internet commerce, the need for SANs increases, 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 because these standards are the basis for the wide acceptance of SANs. Widely accepted standards potentially allow for the heterogeneous, cross-platform, and multivendor deployment of SAN solutions.

Because all vendors accepted these SAN standards, ideally no problems will occur when you connect products from different vendors in the same SAN. However, nearly every vendor has an interoperability lab where it tests all kinds of combinations between their products and the products of other vendors. Several of the most important aspects of these tests are reliability, error recovery, and performance. If a combination passes 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 thinks that industry standards must be in place and, if necessary, redefined for SANs to be a major part of the IT business mainstream.

Probably the most important industry standards organization for SANs is the SNIA. IBM is a founding member and a board officer in SNIA. The SNIA, other standards organizations, and IBM are active participants in many of these organizations.

2

Storage area networks

In Chapter 1, “Introduction” on page 1, we introduced the basics by presenting a network and storage system. We also defined a standard storage area network (SAN) and briefly described the underlying technologies and concepts of a SAN implementation.

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

And although SAN technology is different, many of the concepts can also be applied in the Ethernet networking environment, which is covered in more depth later in this book.

2.1 Storage area networks

This section describes the major motivators that drive SAN implementations and presents several key benefits that this technology can bring to data-dependent businesses.

2.1.1 The problem

Distributed clients and servers are frequently chosen to meet specific application needs. They might, therefore, run different operating systems, such as Windows Server, various UNIX offerings, IBM VMware vSphere, or VMS. They might also run different database software, for example, IBM DB2®, Oracle, IBM Informix®, or SQL Server. Therefore, they have different file systems and different data formats.

The management of this multi-platform, multivendor, networked environment is increasingly complex and costly. Software tools for multiple vendors and appropriately skilled human resources must be maintained to handle data and storage resource management on the many different systems in the enterprise. Surveys that are published by industry analysts consistently show that management costs that are associated with distributed storage are much greater. The costs are shown to be much greater than the cost of managing consolidated or centralized storage. This comparison includes the 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. It is difficult to establish whether the price you pay per gigabyte (GB) is competitive, compared to the market price of disk storage. Disks and tape drives, which are directly attached to one client or server, cannot be used by other systems, leading to inefficient use of hardware resources. Organizations often discover that they need 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. You also need to manage information over its entire lifecycle, from conception to intentional destruction.

Information that is stored on one system cannot readily be made available to other users. One exception is to create duplicate copies and move the copy to the storage that is attached to another server. Movement of large files of data might result in significant degradation of performance of the LAN and wide area network (WAN), causing conflicts with mission-critical applications. Multiple copies of the same data might lead to inconsistencies between one copy and another copy.

Data that is spread on multiple small systems is difficult to coordinate and share for enterprise-wide applications. Examples of enterprise-wide applications include Internet commerce, enterprise resource planning (ERP), data warehouse, and business intelligence (BI).

Backup and recovery operations across a LAN might also cause serious disruption to normal application traffic. Even when you use fast Gigabit Ethernet transport, the sustained throughput from a single server to tape is about 25 GB per hour. It takes approximately 12 hours to fully back up a relatively moderate departmental database of 300 GBs. This time frame might exceed the available window of time in which the backup must be completed. And, it might 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 that result from the distributed model of computing do not match the needs of the enterprise.

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

2.1.2 Requirements

With this scenario in mind, you might consider several requirements for the storage infrastructures of today:

- Unlimited and just-in-time scalability: Businesses require the capability to flexibly adapt to the rapidly changing demands for storage resources without performance degradation.
- System simplification: Businesses require an easy-to-implement infrastructure with a minimum amount of management and maintenance. The more complex the enterprise environment, the more costs that 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 resource is essentially an investment protection requirement that allows for the configuration of a storage resource for one set of systems. It later configures 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 the data of one system from other systems.
- Encryption: When sensitive data is stored, it must be read or written only from certain authorized systems. If for any reason the storage system is stolen, data must never be available to be read from the system.
- Hypervisors: This requirement is for the support of the server, application, and desktop virtualization hypervisor features for cloud computing.
- Speed: Storage networks and devices must be able to manage the high number of gigabytes and intensive I/O that are required by each business industry.
- Availability: This requirement implies both the protection against media failure and the ease of data migration between devices, without interrupting application processing. This requirement 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 the following enhanced backup and recovery capabilities:
  - Serverless backup. This capability is the ability to back up your data without using the computing processor of your servers.
  - Synchronous copy. This capability ensures that your data is at two or more places before your application goes to the next step.
  - Asynchronous copy. This capability ensures that your data is at two or more places within a short time. The disk subsystem controls the data flow.

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

2.2 Using a storage area network

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

2.2.1 Infrastructure simplification

Four major methods exist by which infrastructure simplification can be achieved. An overview is provided for each of the major methods of infrastructure simplification:

- *Consolidation*

  Concentrating the 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 to make complexity nearly transparent. At the same time, storage virtualization can offer a composite view of storage assets. This feature might help reduce capital and administrative costs, and it provides users with 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 can help protect data as storage needs grow. As soon as day-to-day tasks are automated, storage administrators might be able to spend more time on critical, higher-level tasks that are unique to the 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 might be better able to respond to the information needs of your users.

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

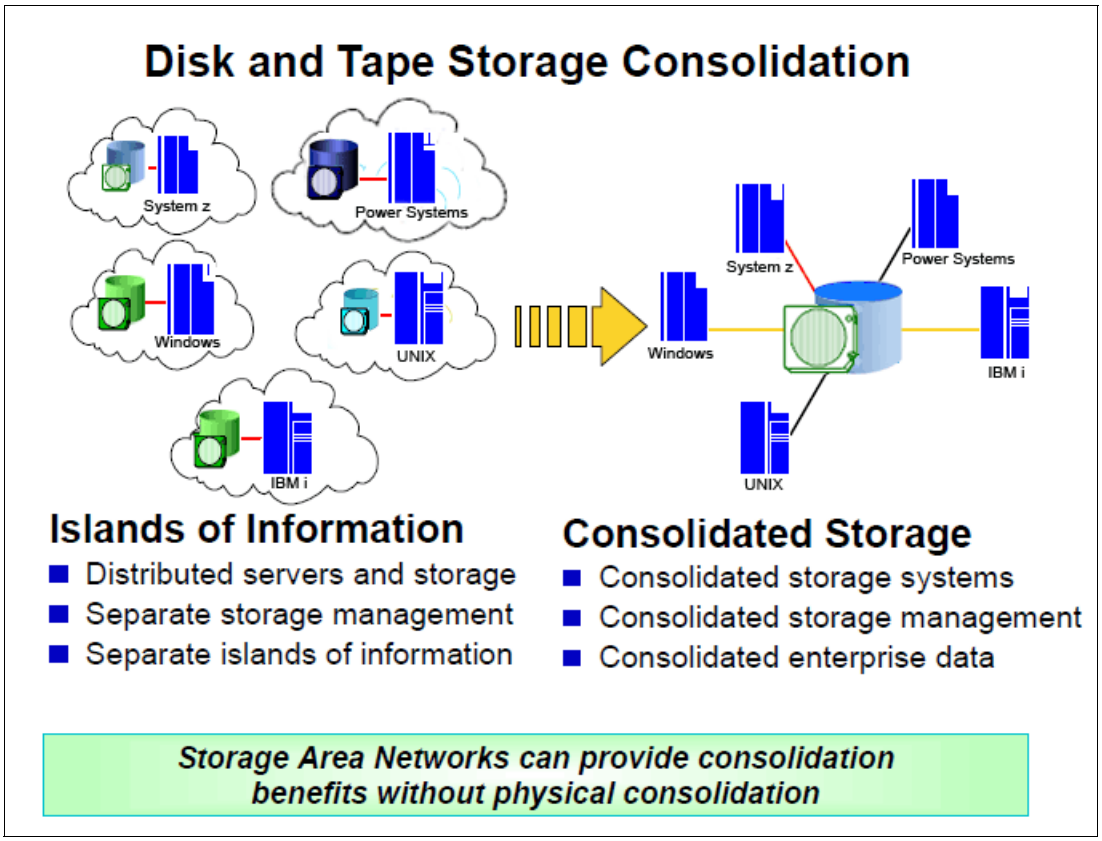


Figure 2-1 Disk and tape storage consolidation

Simplified storage environments have fewer elements to manage. This type of environment leads to increased resource utilization and simplified storage management. This environment 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 is 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 lifecycle, from conception until intentional disposal. The ILM process manages this information 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 because it integrates applications and data into a single-view system, in which the information resides. This single-view location can be managed more efficiently.

IBM Tivoli Productivity Center For Data was designed to support ILM.

2.2.3 Business continuity

The business climate in today's on-demand era is highly competitive. Clients, employees, suppliers, and IBM 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 client privacy and data security so that vital information assets are not compromised. Also, factor in the legal and regulatory requirements, the inherent demands of participating in the global economy, and accountability. All of a sudden, the IT manager can be overwhelmed.

Currently, with natural disasters seemingly occurring with more frequency, a disaster recovery (DR) plan is essential. 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.

Many specific vendor solutions require a SAN that runs in the background, such as IBM VMware Site Recovery Manager (SRM), for business continuity.

A sound and comprehensive business continuity strategy is now considered a business imperative, and SANs play a key role in this continuity. By deploying a consistent and safe infrastructure, SANs make it possible to meet any availability requirements.

2.3 Using the storage area network components

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

2.3.1 Storage

This section briefly describes the major types of storage devices that are available in the market.

Storage systems

By being contained in a single *box*, a storage system (hard disk drive (HDD), solid-state drive (SSD), or Flash) typically has a central control unit that manages all of the I/O. This configuration simplifies the integration of the system with other devices, such as other disk systems or servers.

We introduced you to the components of a storage system in Chapter 1, "Introduction" on page 1. Depending on the specific functionality that is offered by a particular storage system, you can make a storage system behave as a small, midsize, or enterprise solution. The decision about the type of storage system that is more suitable for a SAN implementation depends on the performance capacity and availability requirements for the particular SAN. We describe the product portfolio in Chapter 12, "IBM Fibre Channel storage area network product portfolio" on page 247.

Tape systems

Tape systems, similar to disk systems, are devices that consist of all of the necessary apparatus to manage the use of tapes for storage. In this case, however, the serial nature of a tape makes it impossible for them to be treated in parallel. This treatment is because Redundant Array of Independent Disks (RAID) devices are leading to a simpler architecture to manage and use.

Three types of tape systems exist: drives, autoloaders, and libraries. An overview of each type of system is provided.

Tape drives

As with disk drives, tape drives are the means by which tapes can connect 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 that can manage tapes and perform automatic backup operations. They are typically connected to high-throughput devices that require constant data backup.

Tape libraries

Tape libraries are devices that can manage multiple tapes simultaneously. They can be viewed as a set of independent tape drives or autoloaders. They are typically deployed in systems that require massive storage capacity, or that need a type of data separation that results in multiple single-tape systems. Because a tape is not a random-access media, tape libraries cannot provide parallel access to multiple tapes as a way to improve performance. However, they can provide redundancy as a way to improve data availability and fault-tolerance.

The circumstances under which each of these systems, or even a disk system, might be used strongly depends on the specific requirements of a particular SAN implementation. However, disk systems are used for online storage because of their superior performance. Tape systems are ideal for offline, high-throughput storage, because of the lower cost of storage per byte.

The next section describes the prevalent connectivity interfaces, protocols, and services for building a SAN.

2.3.2 Storage area network connectivity

SAN connectivity consists of hardware and software components that interconnect storage devices and servers. The Fibre Channel model for SANs is introduced.

Standards and models for storage connectivity

Networking is governed by adherence to standards and models. Data transfer is also governed by standards. By far the most common standard is Small Computer System Interface (SCSI).

SCSI is an American National Standards Institute (ANSI) standard that is one of the leading I/O buses in the computer industry.

An industry effort was started to create a stricter standard to allow devices from separate vendors to work together. This effort is recognized in the ANSI SCSI-1 standard. The SCSI-1 standard (circa 1985) is rapidly becoming obsolete. The current standard is SCSI-2. The SCSI-3 standard is in the production stage.

The SCSI bus is a parallel bus, which comes in several variants (Figure 2-2).

Fibre Channel: For more information about parallel and serial data transfer, see Chapter 3, “Fibre Channel internals” on page 33.

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). It was developed more or less in parallel with ANSI SCSI-1.

The SCSI bus not has data lines and also several control signals. An elaborate protocol is part of the standard to allow multiple devices to share the bus efficiently.

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

As always, the demands and needs of the market push for new technologies. In particular, a push exists for faster communications without limitations on distance or the number of connected devices.

Fibre Channel is a serial interface (primarily implemented with fiber-optic cable). Fibre Channel is the primary architecture for most SANs. To support this interface, many vendors in the marketplace produce Fibre Channel adapters and other Fibre Channel 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 shows the evolution of Fibre Channel speeds. Fibre Channel is described in greater depth throughout this publication.

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 2-3 Fibre Channel evolution

Figure 2-4 shows an overview of the Fibre Channel model. The diagram shows the Fibre Channel, which is divided into four lower layers (FC-0, FC-1, FC-2, and FC-3) and one upper layer (FC-4). FC-4 is where the upper-level protocols are used, such as SCSI-3, Internet Protocol (IP), and Fibre Channel connection (FICON).

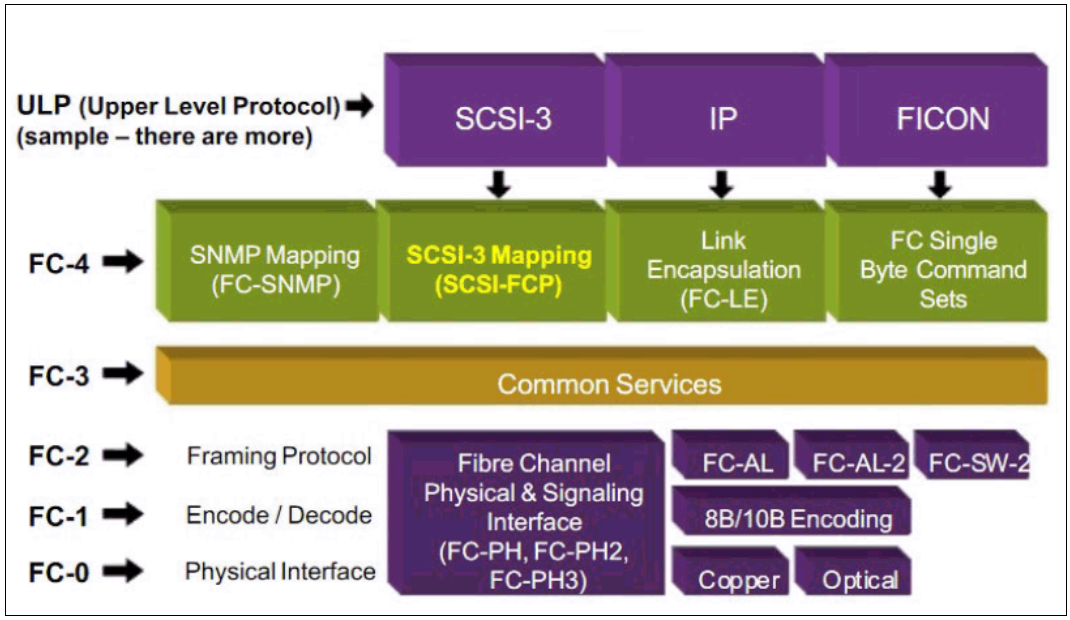


Figure 2-4 Fibre Channel model overview

Options for storage connectivity

In this section, we 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 shows an example of each networking stack.

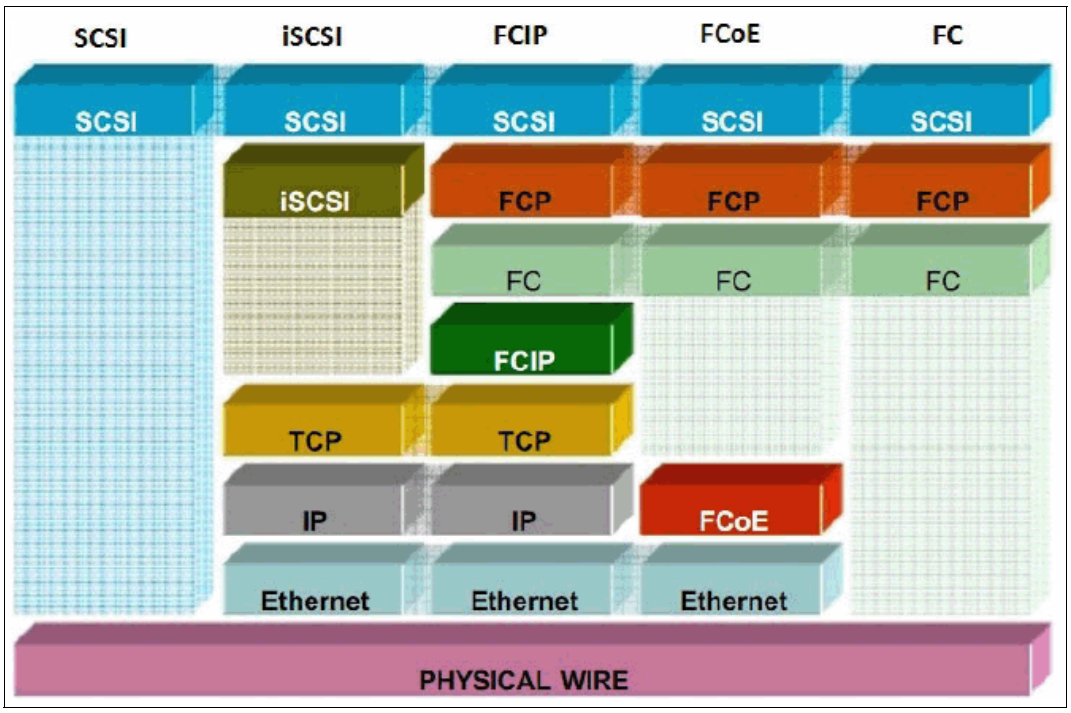


Figure 2-5 Networking stack comparison

Lower-level layers

As Figure 2-5 shows, only three stacks can directly interact with the physical wire: Ethernet, SCSI, and Fibre Channel. Because of this configuration, these models are considered the *lower-level layers*. All of the other stacks are combinations of the layers, such as Internet SCSI (iSCSI), Fibre Channel over IP (FCIP), and Fibre Channel over Ethernet (FCoE), which are also called the *middle-level layers*.

We assume that you have a basic knowledge of Ethernet, which is typically used on conventional server-to-server or workstation-to-server network connections. The connections build up a common-bus topology by which every attached device can communicate with every other attached device by using this common bus. Ethernet speed is increasing as it becomes more pervasive in the data center. Key concepts of Ethernet are describer later in this book.

Middle-level layers

This section consists of the transport protocol and session layers.

Fibre Channel over Ethernet (FCoE): FCoE is described later in this book. It is a vital model for the Converged Network Adapter (CNA).

Internet Small Computer System Interface

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

iSCSI enables the implementation of IP-based SANs, enabling clients to use the same networking technologies, for both storage and data networks. Because iSCSI uses TCP/IP, iSCSI is also 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.

Fibre Channel Protocol

The *Fibre Channel Protocol* (*FCP*) is the interface protocol of SCSI on Fibre Channel (FC). It is a gigabit speed network technology that is primarily used for storage networking. Fibre Channel is standardized in the T11 Technical Committee of the International Committee of Information Technology Standards (INCITS), an ANSI-accredited standards committee. FCP started for use primarily in the supercomputer field, but FCP is now the standard connection type for SANs in enterprise storage. Despite its name, Fibre Channel signaling can run on both twisted-pair copper wire and fiber optic cables.

Fibre Channel over IP

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 tunneled 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 then transports it over a TCP socket. TCP/IP services are used to establish connectivity between remote SANs. Congestion control and management and also data error and data loss recovery are handled by TCP/IP services and do not affect Fibre Channel fabric services.

The major consideration with FCIP is that it does not replace Fibre Channel with IP; it allows deployments of Fibre Channel fabrics by using IP tunneling. You might assume that the industry decided that Fibre Channel-based SANs are appropriate. Another possible assumption is that the IP connection is only needed to facilitate any distance requirement that is beyond the current scope of an FCP SAN.

Fibre Channel connection

Fibre Channel connection (*FICON*) architecture is an enhancement of, rather than a replacement for, the traditional IBM Enterprise Systems Connection (ESCON) architecture. A SAN is Fibre Channel-based (FC-based). Therefore, FICON is a prerequisite for IBM 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 can achieve data rates up to 200 MBps full duplex and extend the channel distance (up to 100 km (62 miles)). FICON can also increase the number of control unit images for each link and the number of device addresses for each control unit link. The protocol can also retain the topology and switch management characteristics of ESCON.

Higher-level layers

This section consists 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 by using an adapter, 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 onboard intelligence. They merely ran the I/O requests of the server. The subsequent evolution led to the introduction of control units (CUs). These units are storage offload servers that contain a limited level of intelligence. The CUs can perform functions, such as I/O request caching for performance improvements or dual copying data (RAID 1) for availability. Many advanced storage functions are developed and implemented inside the CU.

Network-attached storage

Network-attached storage (*NAS*) is basically a LAN-attached file server that serves files by using a network protocol, such as *Network File System* (*NFS*). NAS refers to storage elements that connect to a network and provide file access services to computer systems. An NAS storage element consists of an engine that implements the file services (by using access protocols, such as NFS or Common Internet File System (CIFS)) and one or more devices, on which data is stored. NAS elements might be attached to any type of network.

From a SAN perspective, a SAN-attached NAS engine is treated just like any other server. However, NAS does not provide any of the activities that a server in a server-centric system typically provides, such as email, authentication, or file management.

NAS allows more hard disk storage space to be added to a network that already uses servers, without shutting them down for maintenance and upgrades. With an NAS device, storage is not a part of the server. Instead, in this storage-centric design, the server still handles all of the processing of the data, but an NAS device delivers the data to the user.

An NAS device does not need to be located within the server, but an NAS device can exist anywhere in the LAN. An NAS device can consist of multiple networked NAS devices. These units communicate to a host by using Ethernet and file-based protocols. This method is in contrast to the disk units that are already described, which use Fibre Channel Protocol (FCP) 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).

To bridge the two worlds and open up new configuration options for clients, certain vendors, including IBM, sell NAS units that act as a gateway between IP-based users and SAN-attached storage. This configuration allows the connection of the storage device and shares the storage device between your high-performance database servers (attached directly through FC) and your users (attached through IP). These users do not have strict performance requirements.

NAS is an ideal solution for serving files that are stored on the SAN to users in cases where it is impractical and expensive to equip 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 server platform (IBM z™ Systems, UNIX, IBM AIX®, HP-UX, Sun Solaris, Linux, IBM i, and Microsoft Windows Server) implements SAN solutions by using various interconnections and storage technologies. The following sections review these solutions and the implementation on each platform.

Mainframe servers

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

The IBM System z® is a processor and operating system mainframe set. Historically, the IBM z Systems® servers supported many operating systems, such as z/OS, IBM OS/390®, VM, VSE, and Transaction Processing Facility (TPF), which were enhanced over the years. The processor to storage device interconnection also evolved from a bus and tag interface to ESCON channels, and now to FICON channels. Figure 2-6 shows the various processor-to-storage interfaces.

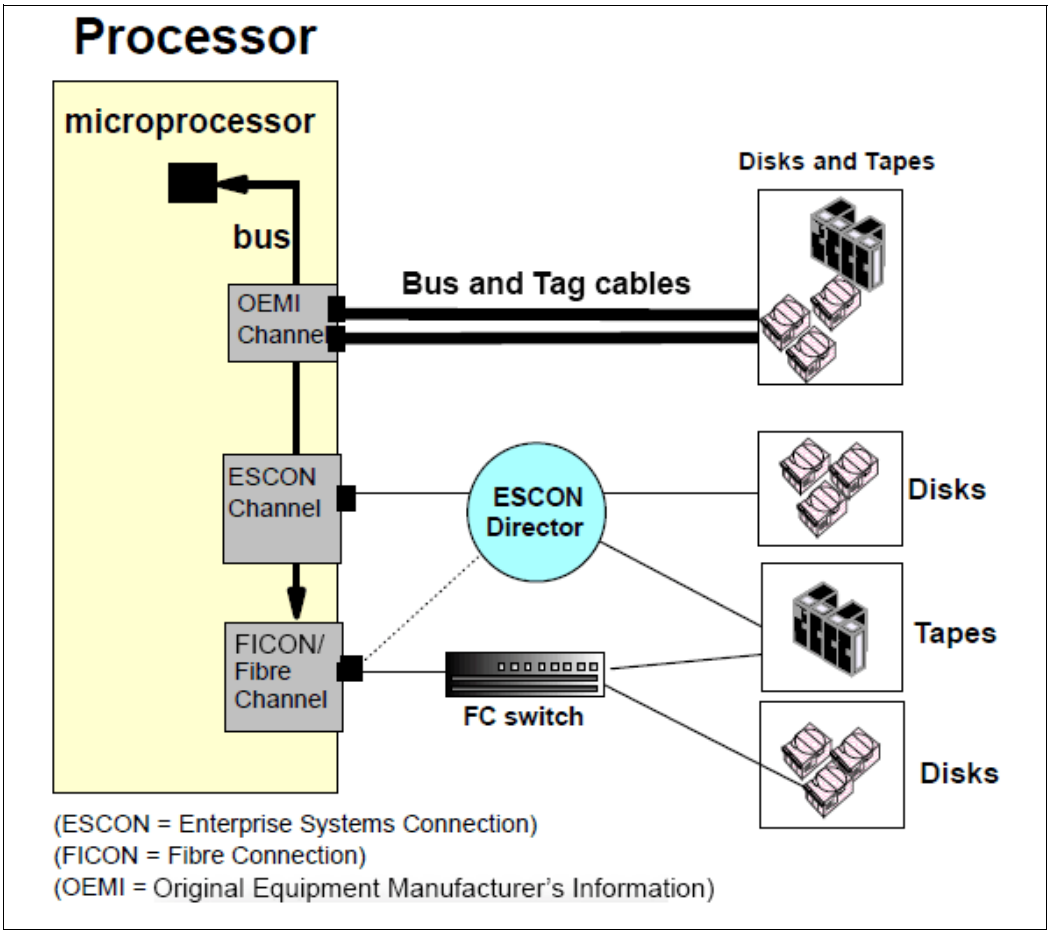


Figure 2-6 Processor-to-storage interface connections

Because of architectural differences, and strict data integrity and management requirements, the implementation of FICON is somewhat behind that of FCP on open systems. However, at the time of writing this book, FICON caught up with FCP SANs, and they coexist amicably.

For the latest news about IBM z Systems® FICON connectivity, see this website:

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

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

UNIX servers

Originally designed for high-performance computer systems, such as mainframes, today's UNIX operating systems appear on many hardware platforms, ranging from Linux personal computers to dedicated large-scale workstations. Because of the popularity and maturity of UNIX, UNIX also plays an important role on both existing and earlier IT infrastructures.

The IBM Power Systems™ servers run a UNIX operating system that is called *AIX*. The Power Systems servers offer various processor-to-storage interfaces, including SCSI, serial-attached SCSI (SAS), and Fibre Channel. The Serial Storage Architecture (SSA) interconnection is primarily used for disk storage. Fibre Channel adapters can connect to tape and disk. Figure 2-7 shows the various processor-to-storage interconnection options for the Power Systems family.

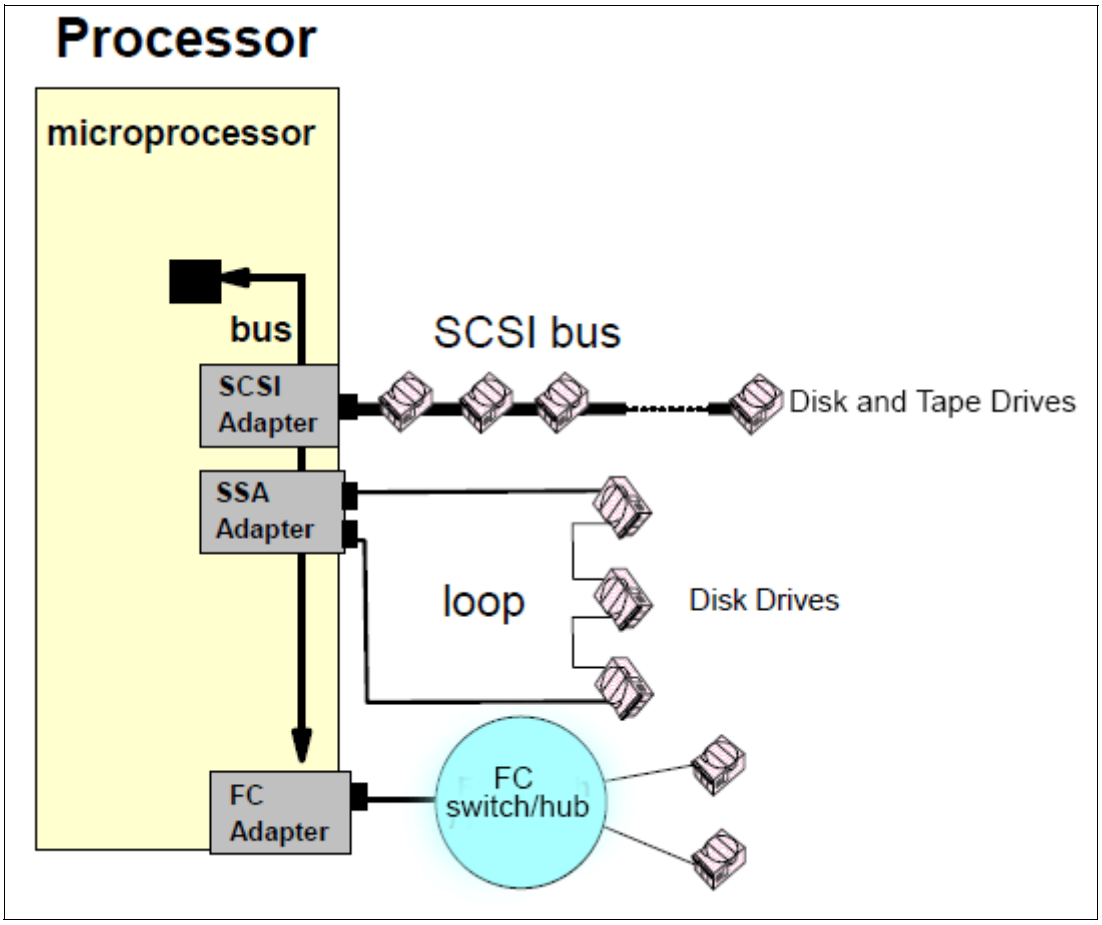


Figure 2-7 Power Systems processor-to-storage interconnections

The various UNIX system vendors in the market deploy variations of the UNIX operating system. Each product offers unique enhancements. The various vendors' UNIX operation system implementations often support separate file systems, such as the journaled file system (JFS), enhanced journaled file system (JFS2), and the IBM Andrew File System (AFS™). The server-to-storage interconnection is similar to Power Systems, as shown in Figure 2-7.

For the latest IBM Power Systems storage products, see this website:

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

Microsoft Windows servers

Based on the reports of various analysts about 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 Windows servers will host mission-critical applications that benefit from SAN solutions, such as disk and tape pooling, tape sharing, multipathing, and remote copy.

The processor-to-storage interfaces on Intel servers that support the Microsoft Windows Server operating system are similar to the interfaces that are supported on UNIX servers, including SCSI and Fibre Channel.

Single-level storage

Single-level storage (*SLS*) is probably the most significant differentiator in a SAN solution implementation on an IBM i server. This IBM i differentiator is a factor when compared to other systems such as z/OS, UNIX, and Windows. In IBM i, both the major storage (memory) and the auxiliary storage (disks) are treated as a large virtual address space that is known as SLS.

Figure 2-8 compares the IBM i SLS addressing with the way that Windows or UNIX systems work, by 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 are possible. Because a single page table maps all virtual addresses to physical addresses, task switching is efficient. SLS further eliminates the need for address translation, therefore speeding up data access.

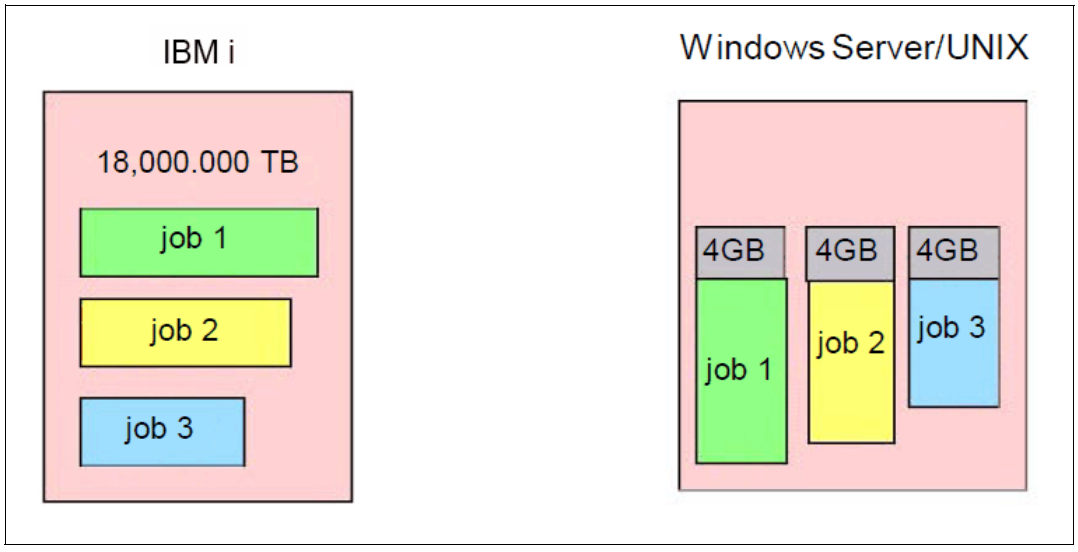


Figure 2-8 IBM i versus 32-bit Windows Server or UNIX storage addressing

The IBM i SAN support was rapidly expanded. IBM i servers now support attachment to switched fabrics and to most of the IBM SAN-attached storage products.

For more information, see this IBM i SAN website:

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

2.3.4 Putting the components together

After looking at all of these technologies and platforms, we can understand why it is a challenge to implement true heterogeneous storage and data environments across different hardware and operating system platforms. Examples of these environments include 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, with its operating system, treats data differently at various levels in the system architecture, therefore creating several of these challenges:

- Different attachment interfaces and protocols, such as SCSI, ESCON, and FICON
- Different data formats, such as extended count key data (IBM ECKD™), blocks, clusters, and sectors
- Different file systems, such as Virtual Storage Access Method (VSAM), JFS, JFS2, AFS, and Windows Server New Technology File System (NTFS)
- IBM i and 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

Figure 2-9 shows a brief summary of these differences for several systems.

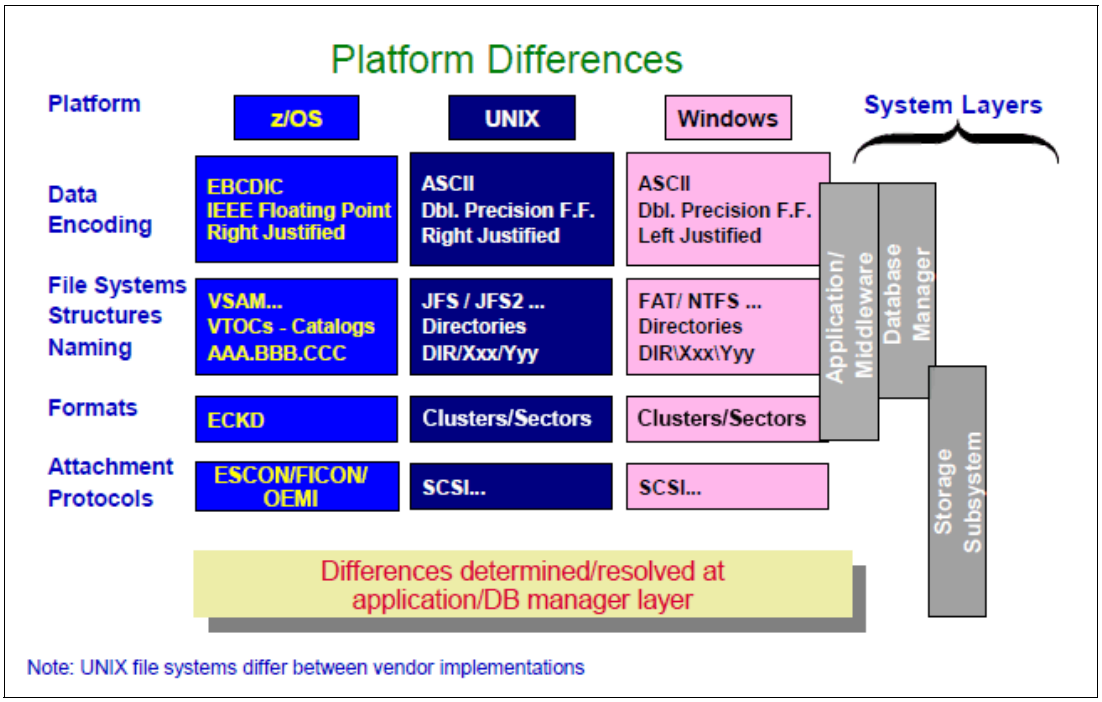


Figure 2-9 Hardware and operating system differences

3

Fibre Channel internals

Fibre Channel (FC) is the predominant architecture on 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. Many products on the market take advantage of the high-speed and high-availability characteristics of the Fibre Channel architecture.

Fibre Channel was developed through industry cooperation, unlike Small Computer System Interface (SCSI), which was developed by a vendor and submitted for standardization afterward.

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) to refer to the standard. The US English spelling (fiber) is retained to refer generically to fiber optics and cabling.

Certain people refer to Fibre Channel architecture as the fibre version of SCSI. Fibre Channel is an architecture that is used to carry intelligent peripheral interface (IPI) traffic, Internet Protocol (IP) traffic, Fibre Channel connection (FICON) traffic, and Fibre Channel Protocol (FCP) SCSI traffic. Fibre Channel architecture might also carry traffic that uses other protocols, all on the standard Fibre Channel transport.

An analogy might be Ethernet, where IP, Network Basic Input/Output System (NetBIOS), and Systems Network Architecture (SNA) are all used simultaneously over a single Ethernet adapter. This configuration is possible because these protocols all have mappings to Ethernet. Similarly, many protocols are mapped onto Fibre Channel.

FICON is the standard protocol for z/OS, and FICON will replace all Enterprise Systems Connection (ESCON) environments over time. FCP is the standard protocol for open systems. Both FICON and FCP use the Fibre Channel architecture to carry the traffic.

3.1 Fibre Channel architecture

Before we delve into the internals of Fibre Channel, we describe why Fibre Channel became the predominant storage area network (SAN) architecture.

3.1.1 Small Computer Systems Interface

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

In addition to being a physical transport, SCSI is also a protocol. This protocol specifies commands and controls for sending blocks of data between the host and the attached devices. The SCSI commands are issued by the host operating system in response to user requests for data. Certain operating systems, for example, Microsoft Windows, treat all attached peripheral devices as SCSI devices, and they issue SCSI commands to handle all read and write operations. SCSI was used in direct-attached storage (DAS) with internal and external devices that connected through the SCSI channel in daisy chain fashion.

Figure 3-1 shows typical SCSI device connectivity.

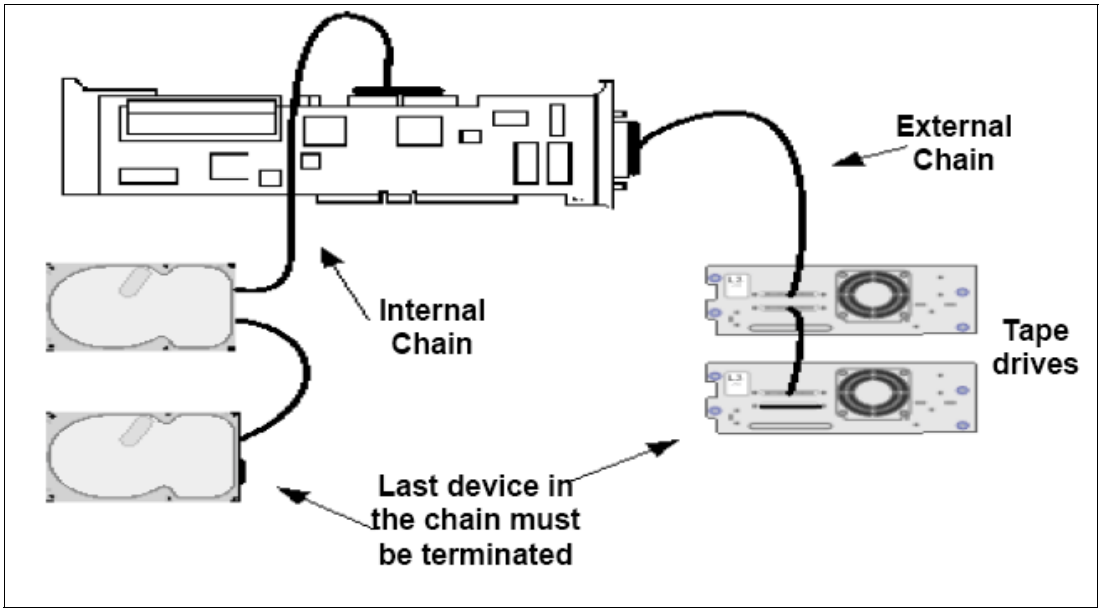


Figure 3-1 SCSI device connectivity

3.1.2 Limitations of the Small Computer System Interface

Several limitations of SCSI are described in the following topics.

Scalability limitations

The amount of data that is available to the server is determined by the number of devices that can attach to the bus. The amount is also determined by the number of buses that are 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, commonly no more than four or five devices are attached in this way. This factor limits the scalability in terms of the number of devices that can connect to the server.

Reliability and availability limitations

SCSI shares aspects with bus and tag; for example, the cables and connectors are bulky, relatively expensive, and prone to failure. Access to data is lost in a failure of any of the SCSI connections to the disks. Data is also lost in the reconfiguration or servicing of a disk device that is attached to the SCSI bus because all of 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 transferred bits, and the bus cycle time (measured in megahertz (MHz)). Decreasing the cycle time increases the transfer rate. However, because of limitations that are inherent in the bus architecture, decreasing the cycle time might also reduce the distance over which the data can be successfully transferred. The physical transport was originally a parallel cable that consisted of eight data lines to transmit 8 bits in parallel, plus control lines. Later implementations widened the parallel data transfers to 16-bit paths (wide SCSI) to achieve higher bandwidths.

A SCSI propagation delay in sending data in parallel along multiple lines leads to a phenomenon that is known as *skew*. Skew means that all bits might not arrive at the target device at the same time. Figure 3-2 shows this result.

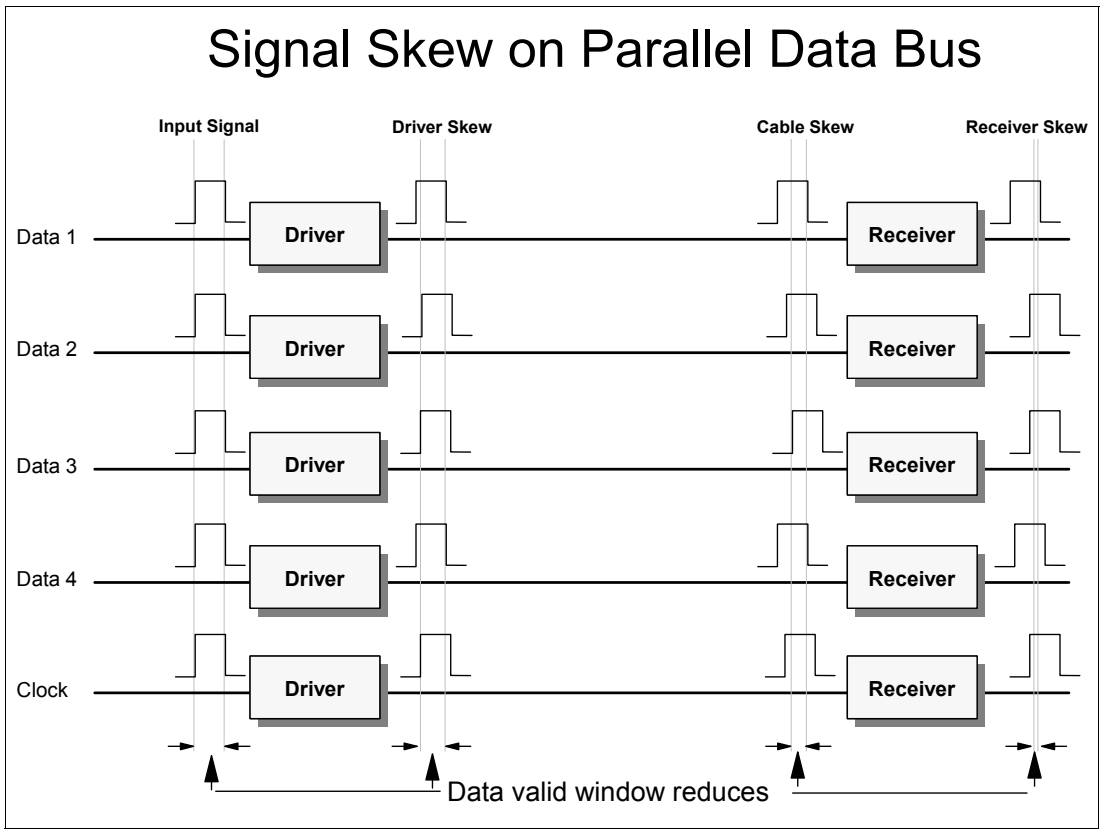


Figure 3-2 A SCSI propagation delay results in skew

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 2 meters (6.5 ft) - 25 meters (82 ft). The distance depends 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. Figure 3-3 shows the SCSI distance limitations. These limitations might severely restrict the total GB capacity of the disk storage that can be attached to an individual server.

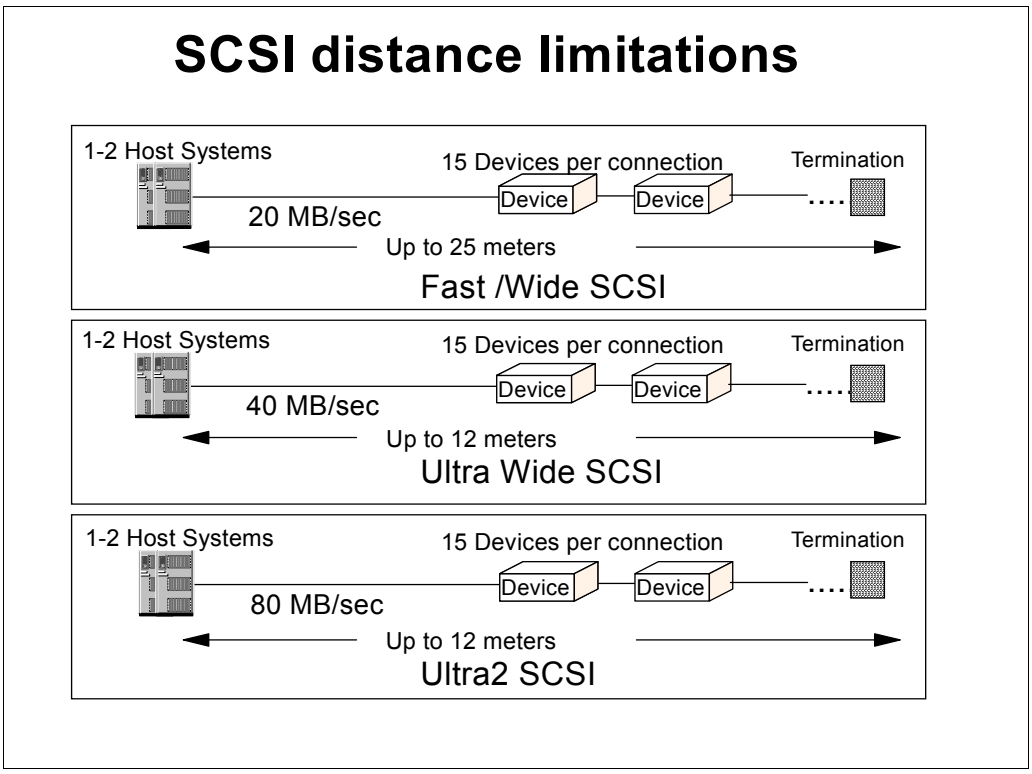


Figure 3-3 SCSI bus distance limitations

Device sharing

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

Figure 3-4 shows a multi-drop bus structure configuration.

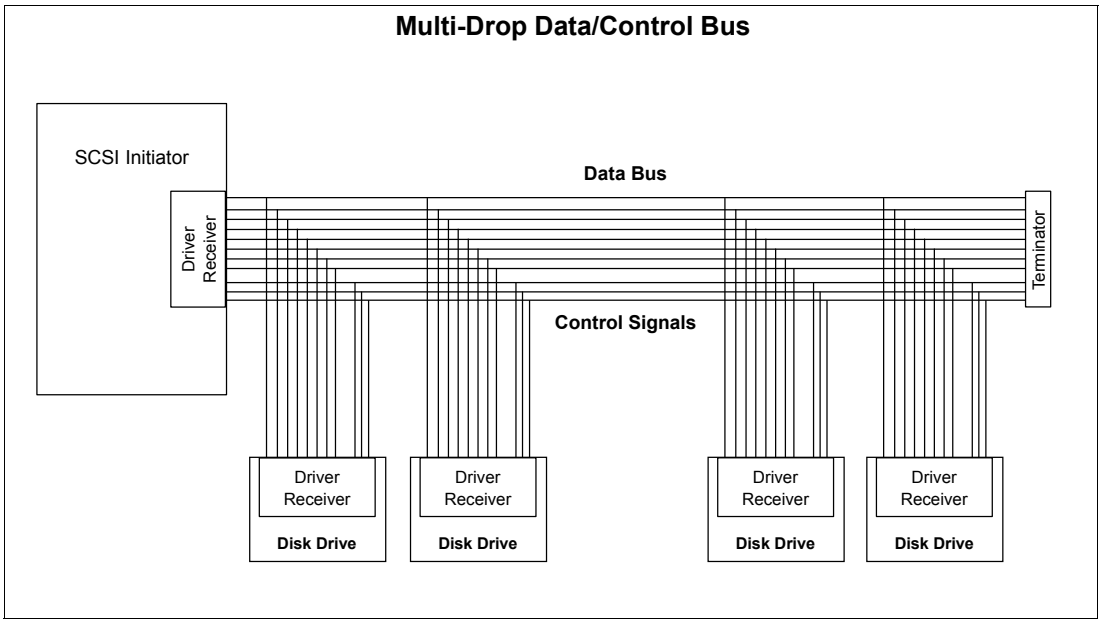


Figure 3-4 Multi-drop bus structure

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

Normally, only a single server can access data on a specific disk with 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 the device that 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 loss of movement represents an additional performance overhead that reduces bandwidth utilization, substantially reducing the effective data rate that is 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 several significant speed, distance, and availability limitations. These limits make it increasingly less suitable for many applications in today's networked IT infrastructure. However, the SCSI protocol is deeply embedded in the way that commonly encountered operating systems handle user requests for data. Therefore, requiring a move to new protocols is a major inhibitor to progress.

3.1.3 Fibre Channel advantages

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

Because of Fibre Channel's channel-like qualities, hosts and applications see storage devices that are attached to the SAN as though they are locally attached storage. Because of Fibre Channel's network characteristics, Fibre Channel can support multiple protocols and a broad range of devices. And, Fibre Channel 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 multiple layer network that is based on a series of American National Standards Institute (ANSI) standards that define characteristics and functions for moving data across the network. These standards include the definitions of physical interfaces, for example:

- Cabling, distances, and signaling
- Data encoding and link controls
- Data delivery in terms of frames
- Flow control and classes of service
- Common services
- 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 significant hardware processing to deliver high performance.

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

Serial transfer does not suffer from the problem of skew, so speed and distance are not restricted in the same way that parallel data transfers are restricted. Figure 3-5 shows the process of parallel data transfers versus serial data transfers.

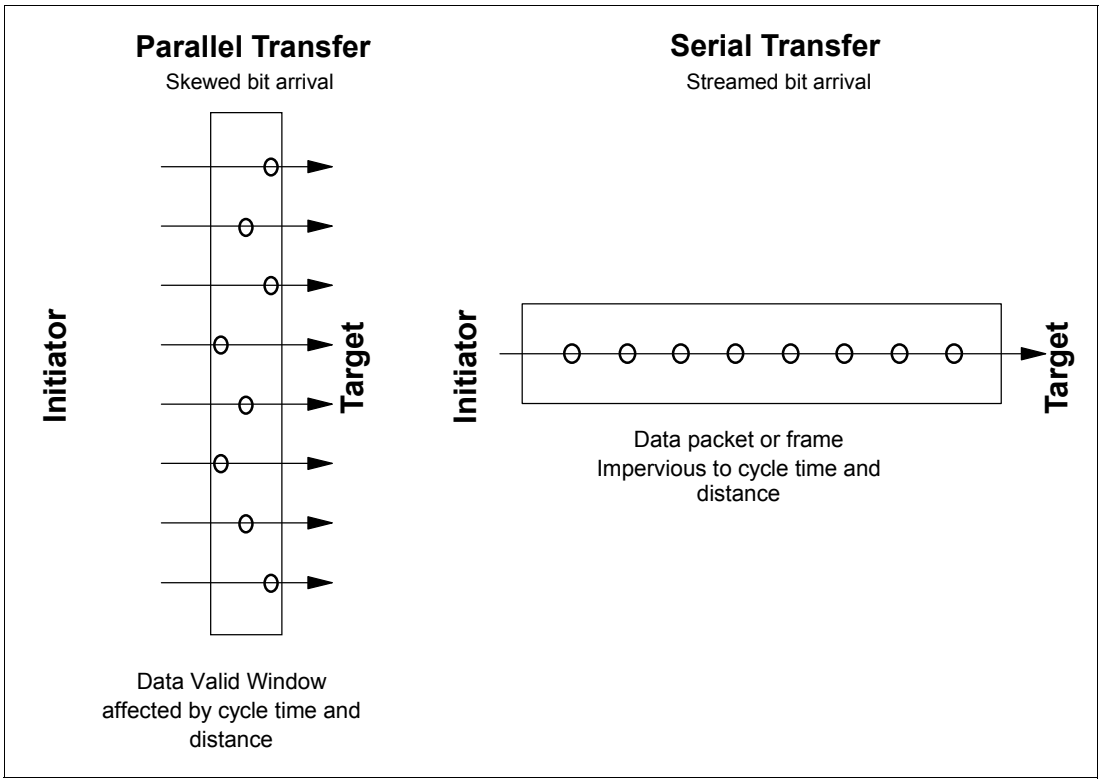


Figure 3-5 Parallel data transfers versus serial data transfers

Serial transfer enables simpler cabling and connectors, and also the 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 wide area networks (WANs) by using high-speed gateways, campus SANs can 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 a disk or tape drive, or a 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 use the same protocol, but each node might support several protocols or data types.

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

Using a credit-based flow control approach, Fibre Channel can deliver data as fast as the destination device buffer can receive it. And low transmission overhead enables high sustained utilization rates without the loss of data.

Therefore, Fibre Channel combines the best characteristics of traditional I/O channels with the characteristics 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
- The ability to package data in packets (frames, in Fibre Channel terminology)
- Flexibility in terms of the types of information that 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 that can be addressed when compared to I/O interfaces: a theoretical maximum of more than 15 million ports

Several factors make the Fibre Channel architecture ideal for the development of enterprise SANs. One example is the high degree of flexibility, availability, and scalability of the architecture. Other factors include 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.

The following topics describe several key concepts that are mentioned in the previous pages and that are behind Fibre Channel SAN implementations. We also introduce more Fibre Channel SAN terminology and jargon that you 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 International Organization for Standardization/open systems interconnection (ISO/OSI) seven-layer model, allows the development of one layer to remain independent of the adjacent layers. Although a Fibre Channel contains five layers, those layers follow the general principles that are stated in the ISO/OSI model.

The series of five layers that make up a Fibre Channel can be categorized into the following layers:

- Physical and signaling layer
- Upper layer

Fibre Channel is a layered protocol. Figure 3-6 shows the upper and physical layers.

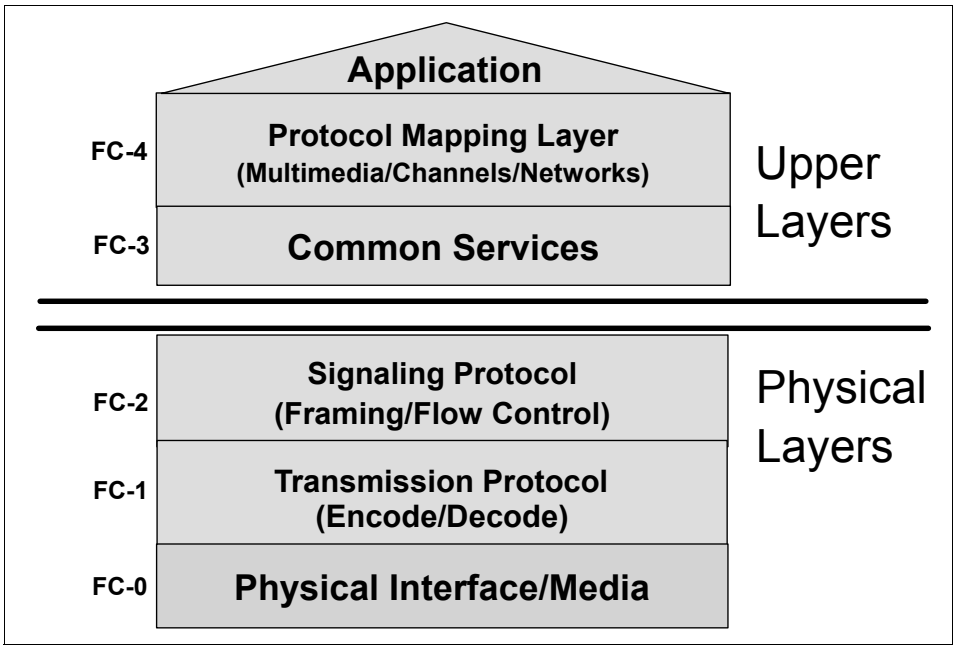


Figure 3-6 Fibre Channel upper and physical layers

The FC layers are briefly described next.

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 this level allows the use of many 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 be 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 it 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 that is required by serial transmission technologies.

Framing and signaling protocol: FC-2

Reliable communications result from the *FC-2* framing and signaling protocol of the FC. 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 *Fibre Channel Physical and Signaling interface* (*FC-PH*), provides the transport methods to determine the following factors:

- Topologies that are based on the presence or absence of a fabric
- Communication models
- Classes of service that are 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 that contain 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 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 the following features:

- Hunt groups

  A *hunt group* is a set of associated node ports (N_ports) that is attached to a single node. This set is assigned an alias identifier that allows any frames that contain the alias to be routed to any available N_port within the set. This process decreases the latency in waiting for an N_port to become available.

- Striping

  Striping is used to multiply bandwidth by 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 method includes the ability to broadcast to all nodes or a subset of nodes.

Upper-layer protocol mapping: FC-4

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

Through mapping rules, a specific FC-4 describes how upper-layer protocol (ULP) processes of the same FC-4 type interoperate.

A channel example is FCP. This protocol is used to transfer SCSI data over Fibre Channel. A networking example is sending IP packets between the 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 thin strand of silica glass and its geometry is quite like a human hair. In reality, it is a narrow, 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:

- Little light is lost in its journey along the fiber.
- Fiber can bend around corners and the light stays within it and is guided around the corners.

An optical fiber consists of two parts: the core and the cladding. See Figure 3-7 on page 45. The core is a narrow cylindrical strand of glass and the cladding is a tubular jacket that surrounds it. The core has a (slightly) higher refractive index than the cladding. Therefore, the boundary (interface) between the core and the cladding acts as a perfect mirror. Light traveling 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 *the kind of light.* The electromagnetic radiation that we call *light* exists at many wavelengths. These wavelengths go from invisible infrared through all of 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 because the wavelengths that are used are typically 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 is changed (degraded) by its passage. It emerges (depending on the distance) much weaker, lengthened in time (*smeared out*), and distorted in other ways. The reasons for this transformation are described in the following topics.

3.3.1 Attenuation

The pulse is 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 the scattering of the light. Both the rate of light absorption and the amount of scattering depend 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

A practical limit exists to the amount of power that can be sent on a fiber. This limit is about half a watt (in a standard single-mode fiber) because of several 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 it contains imperfections. Light traveling down such a fiber is changed in polarization. (In current optical systems, this change does not matter, but in future systems, it might 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 the 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 analog transmission systems where information is sent as a continuous series of changes in the signal, dispersion causes distortion.) Many kinds of dispersion exist, and each kind works differently. The three most important kinds of dispersion are described.

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; therefore, each wavelength travels at a different speed in the fiber. Therefore, certain wavelengths arrive before other wavelengths, and a signal pulse disperses (or smears out).

Modal dispersion

When you use a multimode fiber, the light can take many paths or *modes* as it travels within the fiber. The distance that is traveled by light in each mode differs from the distance that is traveled in other modes. When a pulse is sent, parts of that pulse (rays or quanta) take many different modes (usually all available modes). Therefore, certain components of the pulse arrive before other components of the pulse. The difference between the arrival time of light that takes the fastest mode, versus the arrival time of light that takes the slowest mode, obviously becomes greater as the distance becomes greater.

Waveguide dispersion

Waveguide dispersion is a complex effect, and it is caused by the shape and index profile of the fiber core. However, this effect 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 does not pick up noise from outside the system. However, various kinds of noise can come from components within the system. Mode partition noise can be a problem in single-mode fiber, and modal noise is a phenomenon in multimode fiber.

It is not our intention to delve any deeper into optical than the information that is already described.

3.3.3 Fiber in the storage area network

Fibre Channel can be run over optical or copper media, but fiber-optic cables offer a major advantage in noise immunity. For this reason, fiber-optic cabling is preferred. However, copper is also used. In the short term, a mixed environment likely needs to be tolerated and supported. Although, a mixed environment is less likely to be needed as SANs mature.

In addition to the noise immunity, fiber-optic cabling provides distinct advantages over copper transmission lines that make it an attractive medium for many applications. The following advantages are at the forefront:

- Greater distance capability than is generally possible with copper
- Insensitivity to induced electromagnetic interference (EMI)
- No emitted electromagnetic radiation, such as Radio Frequency Interference (RFI)
- No electrical connection between two ports

- Not susceptibility to crosstalk
- Compact and lightweight cables and connectors

However, fiber-optic and optical links have drawbacks. The drawbacks include the following considerations:

- 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 might be affected by dirt and other contamination.

Overall, optical fibers provide a high-performance transmission medium that was 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. Product flexibility needs to be considered when you plan a SAN. Copper cables tend to be used for short distances, up to 30 meters (98 feet), and they can be identified by their DB-9, 9-pin connector.

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

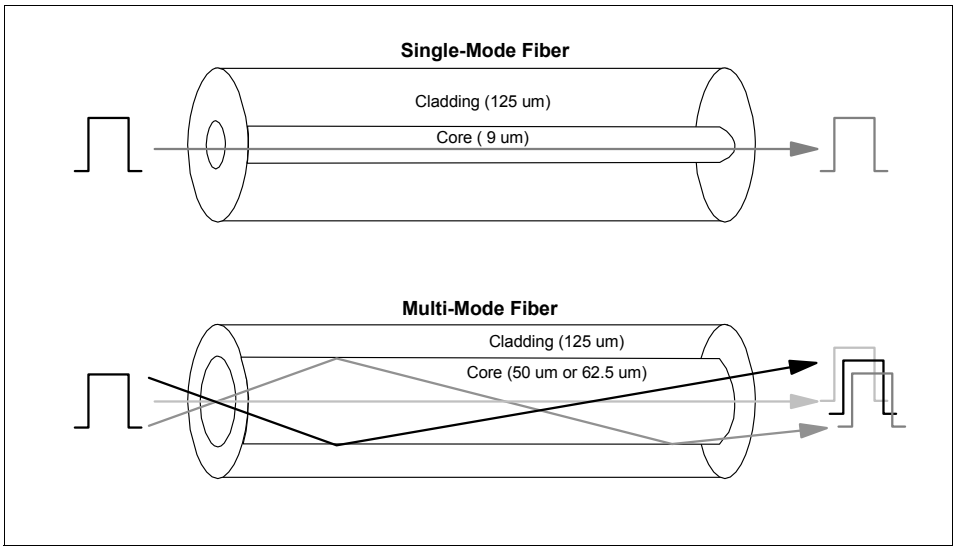


Figure 3-7 Cable types

The cable types are described:

- Multi-mode fiber 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 562.5-micron diameter is sufficiently large for injected light waves to be reflected off the core interior.

  Multi-mode fiber (MMF) allows more than one mode of light. Common multi-mode core sizes are 50 micron and 62.5 micron. MMF fiber is better suited for shorter-distance applications. Where costly electronics are heavily concentrated, the primary cost of the system is not the cable. In this case, MMF is more economical because it can be used with inexpensive connectors and laser devices, therefore reducing the total system cost.

- Single-mode fiber for longer distances

  Single-mode fiber (SMF) allows only one pathway, or mode, of light to travel within the fiber. The core size is typically 8.3 micron. SMFs are used in applications where low signal loss and high data rates are required. An example of this type of application is on long spans between two system devices or network devices, where repeater and amplifier spacing needs to be maximized.

Fibre Channel architecture supports both short wave and long wave optical transmitter technologies in the following ways:

- Short wave laser

  This technology uses a wavelength of 780 nanometers, and it is only compatible with MMF.

- Long wave laser

  This technology uses a wavelength of 1300 nanometers. It is compatible with both SMF and MMF.

Table 3-1 lists the cable types and their speed and distance.

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
Multi-mode 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 multi-mode fiber (MMF) details.

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

Table 3-2 Optical multimode designations

Multi-mode 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

To connect one optical device to another optical device, a form of fiber-optic link is required. If the distance is short, a standard fiber cable suffices. Over a slightly longer distance, for example from one building to the next building, you might need to lay a fiber link. You might need to lay this fiber underground or through a conduit. This process is not as simple as connecting two switches together in a single rack.

If the two units that need to be connected are in separate cities, the problem is much larger. Larger, in this case, is typically associated with more expensive. Because most businesses are not in the business of laying cable, they lease fiber-optic cables to meet their needs. When a company leases equipment, 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 used while the particular technology supports the need.

Certain forward-thinking services companies laid fiber-optic links beside their pipes and cables. For example, a water company might dig up a road to lay a main pipe. Other examples include an electric company that might take a power cable across a mountain range by using pylons. Or, a cable TV company might lay cable to all of the buildings in a city. While they perform the work to support their core business, they might also lay fiber-optic links.

But these cables are merely cables. They are not used in any way by the company who owns them. They remain dark until the user puts their own light down the fiber, therefore, the term dark fiber.

3.4 Classes of service

Applications might require different levels of service and guarantees for delivery, connectivity, and bandwidth. Certain applications need bandwidth that is dedicated to the application during the data exchange. An example of this type of application is a tape backup. Other applications might be *bursty* in nature and not require a dedicated connection, but they might insist that an acknowledgment is sent for each successful transfer. The Fibre Channel standards provide different classes of service to accommodate different application needs. Table 3-3 provides brief details of the separate classes of service.

Table 3-3 Fibre Channel classes of service

Class	Description	Requires an acknowledgment
1	Dedicated connection with full bandwidth	Yes
2	Connectionless switch-to-switch communication for frame transfer and delivery	Yes
3	Connectionless switch-to-switch communication for frame transfer and delivery	No
4	Dedicated connection with a fraction of bandwidth between ports by using virtual circuits	Yes
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 are established through the fabric during the transmission. Class 1 service 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. This class reserves full bandwidth for the connection between the two devices. It does not provide for a good utilization of the available bandwidth because it blocks another possible contender for the same device. Because of this blocking and the 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 because it allows the fabric to multiplex several messages on a frame-by-frame basis. While frames travel through the fabric, they can take separate routes, so class 2 service does not guarantee in-order delivery. Class 2 relies on upper-layer protocols to take care of the frame sequence. The use of acknowledgments reduces available bandwidth, which needs to be considered in large-scale busy networks.

3.4.3 Class 3

No dedicated connection is available 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 correct order. The upper-layer protocol also needs 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, which is similar to class 1. The major difference is that class 4 allocates only a fraction of the available bandwidth of the 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 the in-order delivery of frames and provides an acknowledgment of delivered frames. However, now the fabric is responsible for multiplexing frames of different VCs. Class 4 service is intended for multimedia applications, such as video, and for applications that allocate an established bandwidth by department within the enterprise. Class 4 is included in the FC-PH-2 standard.

3.4.5 Class 5

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

3.4.6 Class 6

Class 6 is a variant of class 1, and it is known as a *multicast class of service*. It provides dedicated connections for a reliable multicast. An N_port might request a class 6 connection for one or more destinations. A multicast server in the fabric establishes the connections, receives the acknowledgment from the destination ports, and sends the acknowledgment back to the originator. When a connection is established, the connection is retained and guaranteed by the fabric until the initiator ends the connection. Class 6 was designed for applications, such as audio and video, that require multicast functionality. Class 6 is included in the FC-PH-3 standard.

3.4.7 Class F

Class F service is defined in the Fibre Channel Switched Fabric (FC-SW) standard and the FC-SW-2 standard for use by switches that communicate through inter-switch links (ISLs). It is a connectionless service with notification of non-delivery between E_ports that are used for the control, coordination, and configuration of the fabric. Class F is similar to class 2. The major difference is that class 2 works with N_ports that send data frames. Class F is used by E_ports for the control and management of the fabric.

3.5 Fibre Channel data movement

To move data bits with integrity over a physical medium, a mechanism must exist to check that this movement happened and that integrity is not compromised. This review is provided by a reference clock, which ensures that each bit is received as it was transmitted. In parallel topologies, you can perform this review by using a separate clock or strobe line. While data bits are transmitted in parallel from the source, the strobe line alternates between high or low to signal to the receiving end that a full byte was sent. If 16-bit and 32-bit wide parallel cables are used, the strobe line indicates that multiple bytes were sent.

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

This dispersion might result in frames that arrive at different times. This bit error rate (BER) is referred to as the *jitter budget*. No products are entirely jitter-free. This jitter budget is an important consideration when you select the components of a SAN.

Because serial data transports have only two leads, transmit and receive, clocking is not possible by 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 (covered in more detail in 3.5.1, “Byte-encoding schemes” on page 52) and clock and data recovery (CDR) logic to recover the clock. From this recovery, it determines the data bits that make up 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 rate equates to a maximum of one bit error every 16 minutes. However, actual occurrence is significantly less frequent.

3.5.1 Byte-encoding schemes

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

First developed by IBM, the 8b/10b encoding process converts 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 that are input to the encoder and the number of bits that are output from the encoder.

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

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

Figure 3-8 shows an encoding example.

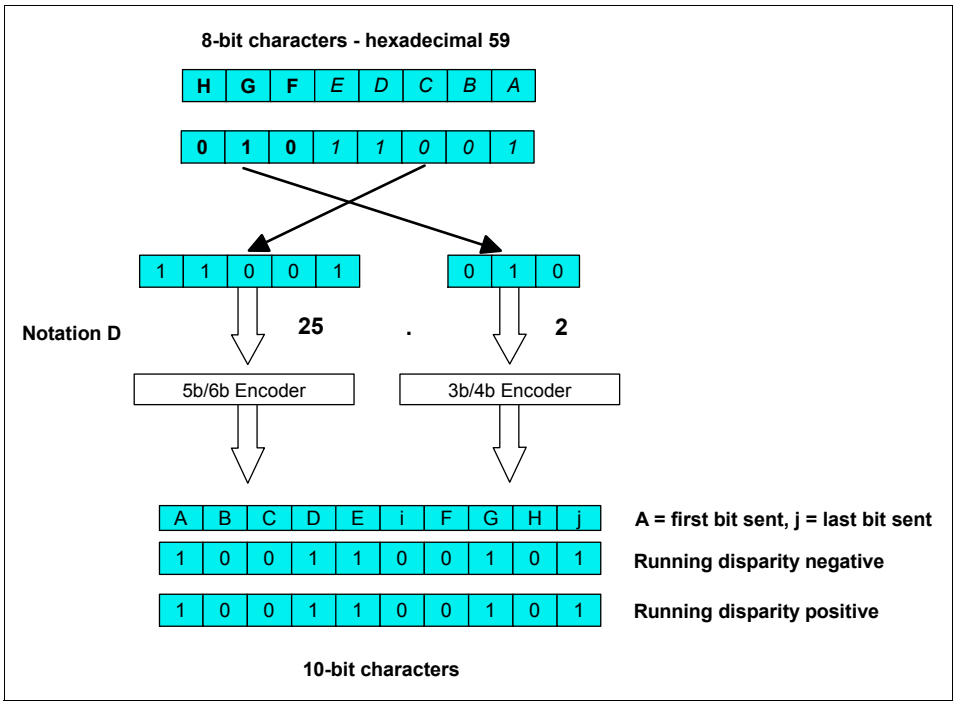


Figure 3-8 8b/10b encoding logic

The following steps occur in the encoding example that is shown in Figure 3-8:

1. Hexadecimal representation x'59' is converted to binary: 01011001.
2. The upper 3 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. The letter notation `D` (for data) is assigned and becomes D25.2.

Running disparity

The conversion of the 8-bit data bytes results in two 10-bit results. The encoder must choose one of these results to use. This decision is achieved by monitoring the running disparity of the previously processed character. For example, if the previous character showed a positive disparity, the next character that is issued might have an encoded value that represents negative disparity.

In the example that is used in Figure 3-8 on page 52, the encoded value, when the running disparity is either positive or negative, is the same. This outcome is legitimate. In certain cases, the encoded value differs, and in other cases, the encoded value is the same.

In Figure 3-8 on page 52, 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 shown:

- If 5 bits are set, the byte is said to have neutral disparity.
- If 4 bits are set and 6 bits are unset, the byte is said to have negative disparity.
- If 6 bits are set and 4 bits are unset, 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. Therefore, if the current running disparity is negative, the next byte that is sent must have one of these properties:

- Neutral disparity:
  - Keeping the current running disparity negative.
  - The subsequent byte needs to have either neutral or positive disparity.
- Positive disparity:
  - Making the new current running disparity neutral.
  - The subsequent byte has either positive, negative, or neutral disparity.

Number of bits: At any time or at the end of any byte, the number of set bits and unset bits that passes over a Fibre Channel link differ only by a maximum of two.

K28.5

In addition to the fact that many 8-bit numbers encode to *two* 10-bit numbers under the 8b/10b encoding scheme, other key features exist.

Certain 10-bit numbers cannot be generated from any 8-bit number. Therefore, it is not possible to see these particular 10-bit numbers as part of a flow of data. This outcome is useful because these particular 10-bit numbers can be used by the protocol for signaling or control.

These characters are referred to as *comma* characters. Instead of the prefix `D`, they use the prefix `K`.

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

Table 3-4 shows the two 10-bit encodings of K28.5.

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

Name of character	Encoding for current running disparity of either negative or positive	
	Negative	**Positive**
K28.5	001111 1010	110000 0101

All of the 10-bit bytes that are possible by using the 8b/10b encoding scheme have either 4, 5, or 6 bits that are set. The K28.5 character is special because it is the only character that is used in Fibre Channel that has 5 consecutive bits that are set or unset. All other characters have four or fewer consecutive bits of the same setting.

When you determine the significance of the bit settings, consider two things:

- The 1s and 0s actually represent light and dark on the fiber (assuming fiber optic medium). A 010 pattern effectively is a light pulse between two periods of darkness. A 0110 represents the same, except that the pulse of light lasts for twice the length of time.

  Because the two devices have their own clocking circuitry, the number of consecutive set bits, or consecutive unset bits, becomes important. For example, device 1 is sending to device 2, and the clock on device 2 is running 10% faster than the clock on device 1. If device 1 sent 20 clock cycles worth of set bits, device 2 counts 22 set bits. (This example is provided merely to illustrate the point.) The worst possible case that you can have in Fibre Channel is 5 consecutive bits of the same setting within 1 byte: the K28.5.

- Because K28.5 is the *only* character with five consecutive bits of the same setting, Fibre Channel hardware can look out for it specifically. Because K28.5 is used for control, this setting is useful. This setting allows the hardware to be designed for maximum efficiency.

64b/66b encoding

Communications of 10 Gbps and 16 Gbps use 64/66b encoding. Sixty-four bits of data are transmitted as a 66-bit entity. The 66-bit entity is made by prefixing one of two possible 2-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 8-bit type field follows, plus 56 bits of control information and data. The preambles *00* and *11* are not used, and they generate an error, if seen.

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

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

3.6 Data transport

For Fibre Channel devices to be able to communicate with each other, strict definitions must exist about the way that data is sent and received. Because of this need, certain data structures are defined. It is fundamental to understanding Fibre Channel that you have at least minimal knowledge of the way that data is moved around. You also need a basic understanding of the mechanisms that are used to accomplish this data movement.

3.6.1 Ordered set

Fibre Channel uses a command syntax, which is known as an *ordered set*, to move the data across the network. The ordered sets are 4-byte transmission words that contain data and special characters that 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. Eleven types of SOF and eight types of EOF delimiters are defined for the fabric and N_port sequence control.

The two primitive signals, idle and receiver ready (R_RDY), are ordered sets that are designated by the standard to have a special meaning. An *idle* is a primitive signal that is transmitted on the link to indicate that an operational port facility is 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, the set might indicate conditions that are 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 the consecutive detection of three instances of the same ordered set. The primitive sequences that are supported by the standard include the following settings:

- 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.
  - The port received and recognized the NOS protocol.
  - 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 detected a link failure. Or, the NOS might indicate an offline condition that is 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 recognizes a link reset sequence and performed the correct link reset.

Data transfer

To send data over Fibre Channel, though, we need more than merely 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 56. 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 that are administered by an even larger unit that is called an *exchange*.

Frame arrival: Certain classes of Fibre Channel communication guarantee that the frames arrive at the destination in the same order in which they were transmitted. Other classes do not. If the frames arrive in the same order in which they were sent, the delivery is considered an *in order* delivery of frames.

A frame consists of the following elements:

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

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

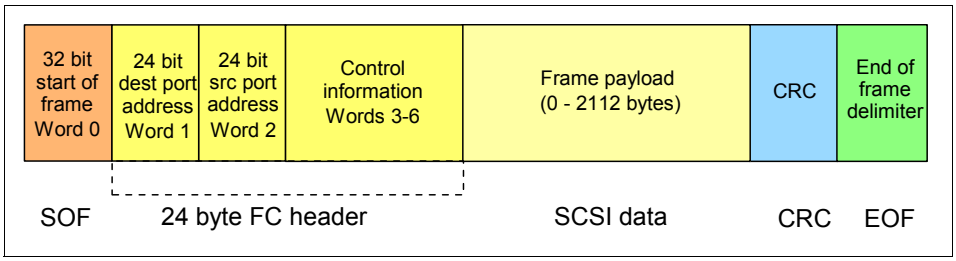


Figure 3-9 Fibre Channel frame structure

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 that is defined in Fibre Channel. This unit consists of four transmission characters, 4 x 10 or 40 bits. When information that is transferred is not an even multiple of 4 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 that are prefixed by a SOF delimiter and followed by an EOF delimiter. Table 3-5 shows the way that transmission words make up a frame.

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. The frame also includes control information that manages the frame, sequences, and exchanges that are associated with that frame.

Table 3-6 shows the structure of the frame header.

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	Sequence Count (SEQ_CNT)	
Word 4	Originator X_ID (OX_ID)		Responder X_ID (RX_ID)	
Word 5	Parameter			

The abbreviations in Table 3-6 are explained:

- Routing control (R_CTL): This field identifies the type of information that is contained in the payload and where in the destination node it might be routed.
- Destination ID: This field contains the address of the frame destination. This field is referred to as the D_ID.
- Source ID: This field contains the address where the frame comes from. This field is referred to as the S_ID.
- Type: The type field 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. The sequence ID is unique for a specific D_ID and S_ID pair while the sequence is open.
- DF_CTL: The data field control specifies whether optional headers are present at the beginning of the data field.
- SEQ_CNT: This count identifies the position of a frame within a sequence. This field is incremented by one for each subsequent frame that is transferred in the sequence.
- OX_ID: This field identifies the exchange ID that is assigned by the originator.
- RX_ID: This field identifies the exchange ID that is assigned to the responder.
- Parameter: The parameter field specifies the relative offset for data frames, or information that is 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. 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 identify the exchange ID. An exchange is responsible for managing a single operation that might 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 following information units (IUs) are relevant for this SCSI task:

- 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. Figure 3-10 shows the flow of the exchange, sequence, and frames.

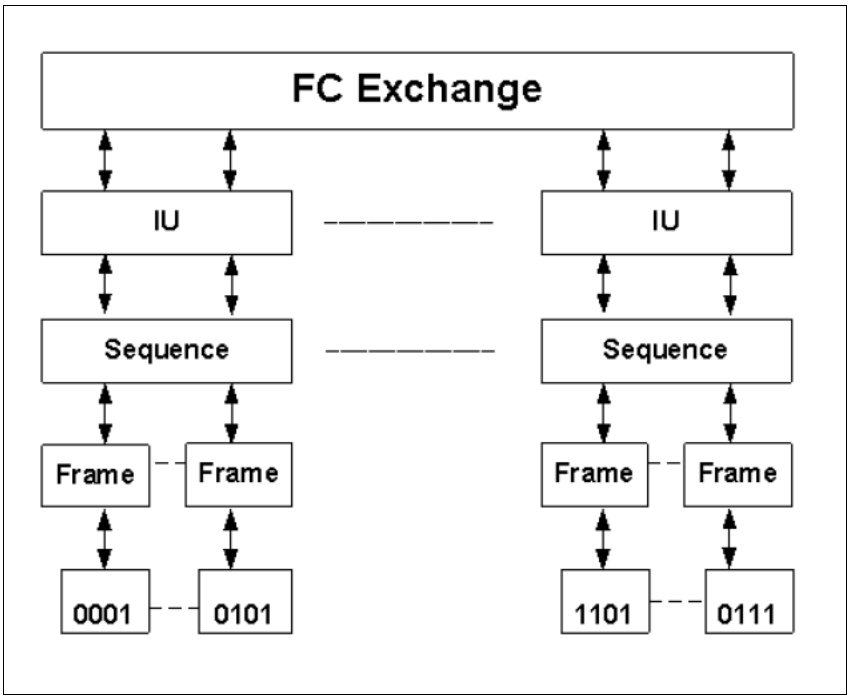


Figure 3-10 Fibre Channel (FC) exchange, sequence, and frame flow

3.6.5 In order and out of order

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

In certain cases, it is critical that the frames arrive in the correct order, and in other cases, it is not as important. In the latter case, which is considered *out of order*, the receiving port can reassemble the frames in the correct order before the port passes the data out to the application. It is, however, 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 that is requested and the action occurring.

Latency occurs almost everywhere because it takes time and energy to perform an action. The following areas highlight where you need to be aware of the latency in a SAN:

- Ports
- Switches and directors
- Inter-Chassis Links in a DCX director
- Long-distance links
- Inter-Switch Links
- Application-specific integrated circuits (ASICs)

3.6.7 Open fiber control

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

ATTENTION: Never look into a laser light source. And never look into the end of a fiber optic cable unless you know exactly where the other end is. You also need to know that no one can connect a light source to it.

To add a degree of safety, the concept of open fiber control (OFC) was developed. The following actions describe this concept:

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

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

OFC ensures that the laser does not emit light that will exceed the class 1 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 certain existing equipment might use OFC optics.

3.7 Flow control

Now that you know that data is sent in frames, you also must understand that devices need to temporarily store the frames as they arrive. The data frames must be stored until they are assembled in sequence and then delivered to the upper-layer protocol. Because of the potential high bandwidth of the Fibre Channel, it is possible to inundate and overwhelm a target device with frames. A mechanism must exist to prevent this situation. The ability of a device to accept a frame is called its *credit*. This credit is typically 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 are the maximum number of frame transfers that a port can support. During login, node ports (N_ports) and fabric ports (F_ports) at both ends of a link establish its buffer-to-buffer credit (BB_Credit). Each port states the maximum BB_Credit that it can offer and the lower of the two BB_Credits is used.

3.7.2 End to end

At login, all N_ports establish an end-to-end credit (EE_Credit) with each other. During data transmission, a port must not send more frames than the buffer of the receiving port can handle. The sending port must first get an indication from the receiving port that it processed a previously sent frame.

3.7.3 Controlling the flow

Two counters are used to accomplish successful flow control: The BB_Credit_CNT and EE_Credit_CNT, and both counters are initialized to 0 during login. Each time that a port sends a frame, the port increments the BB_Credit_CNT and EE_Credit_CNT by one. When the port receives a receiver ready (R_RDY) indication from the adjacent port, it decrements the BB_Credit_CNT by one, and when it receives an acknowledgment (ACK) from the destination port, it decrements the EE_Credit_CNT by one. At certain times, the BB_Credit_CNT might be equal to the BB_Credit, or the EE_Credit_CNT might become equal to the EE_Credit of the receiving port. If this situation happens, the transmitting port must stop sending frames until the related 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 the BB_Credit; only the EE_Credit is used (EE Flow Control). Class 3, however, is an unacknowledged service, so it uses only the BB_Credit (BB Flow Control), but the mechanism is the same in all cases.

3.7.4 Performance

You can see the importance of the number of buffers in overall performance. You need enough buffers to ensure that the transmitting port can continue sending frames without stopping to be able to use the full bandwidth. Using sufficient buffers is important with distance. At 1 Gbps, a frame occupies about 75 m (246 ft) - 4 km (2.48 miles) of fiber, which depends on the size of the data payload. In a 100 km (62 miles) link, many frames can be sent before the first frame reaches its destination. You need an ACK back to start replenishing the EE_Credit or an R_RDY indication to replenish the BB_Credit.

For a moment, consider frames with 2 KB of data. These frames occupy approximately 4 km (2.48 miles) of fiber. You are able to send about 25 frames before the first frame arrives at the far end of the 100 km (62 miles) link. You are able to send another 25 frames before the first R_RDY or ACK indication is received. Therefore, you need at least 50 buffers to allow for non-stop transmission at a 100 km (62 miles) distance with frames of this size. If the frame size is reduced, more buffers are required to allow non-stop transmission. In brief, the buffer credit management is critical in long-distance communication. Therefore, the correct buffer credit allocation is important to obtain optimal performance. Incorrect allocation of the buffer credit might result in a delay of transmission over the Fibre Channel link. As a preferred 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 introduce you to Ethernet and system networking concepts. We also describe the storage area network (SAN) Internet Protocol (IP) networking options and how we arrive at converged networks.

4.1 Ethernet

In Chapter 1, “Introduction” on page 1, we briefly introduced the network and the importance of the models. The Ethernet standard fits into layer 2 of the open systems interconnection (OSI) model. The standard refers to the media access layer that devices are connected to (the cable) and compete for access by using the Carrier Sense Multiple Access with Collision Detection (CSMA/CD) protocol.

Ethernet is a standard communications protocol that is 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 Institute of Electrical and Electronics Engineers (IEEE) proposed a working group in February 1980 to standardize network protocols. The third subcommittee worked on a version that is nearly identical to Ethernet, although insignificant variances exist. Therefore, the generic use of the term Ethernet might refer to IEEE 802.3 or DIX Ethernet.

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

4.1.1 Shared media

Because all communications happen on the same wire, any information that is sent by one computer is received by all computers, 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 that is not addressed to it. Use of a single cable also means that the bandwidth is shared, so that network traffic can be extremely slow when many stations are simultaneously active.

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

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

4.1.2 Ethernet frame

Figure 4-1 shows an Ethernet frame.

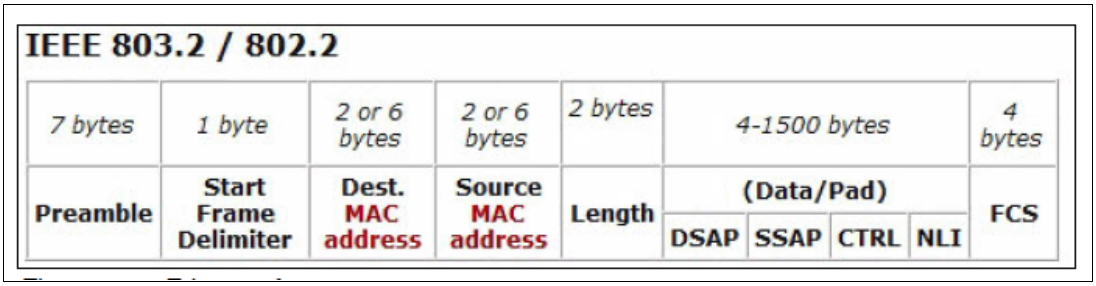


Figure 4-1 Ethernet frame

The Ethernet frame that is shown in Figure 4-1 contains the following components:

- Preamble: A *preamble* is a stream of bits that are used to allow the transmitter and receiver to synchronize their communication. The preamble is an alternating pattern of 56 binary ones and zeros. The preamble is immediately followed by the Start Frame Delimiter.
- Start Frame Delimiter: The *Start Frame Delimiter* is always 10101011. It is used to indicate the beginning of the frame information.
- Destination Media Access Control: The *destination Media Access Control* (MAC) is the address of the system that is receiving data. When a NIC is listening to the wire, the NIC is checking this field for its own MAC address.
- Source Media Access Control: The *source Media Access Control* (MAC) is the MAC address of the system that is transmitting data.
- Length: This field is the *length* of the entire Ethernet frame, in bytes. Although this field can hold any value 0 - 65,534, it is rarely larger than 1500. This smaller value is typically the maximum transmission frame size for most serial connections. Ethernet networks tend to use serial devices to access the Internet.
- Data/pad, which is also known as *payload*: The data is inserted in the *data/pad* or payload. This location is where the IP header and data are placed if you are running IP over Ethernet. This field contains Internetwork Packet Exchange (IPX) information if you are running IPX/Sequenced Packet Exchange (SPX) protocol (Novell). The following specific fields are contained within the data/pad section of an IEEE 803.2 frame:
  - Destination service access point (DSAP)
  - Source service access point (SSAP)
  - Control bits for Ethernet communication (CTRL)
  - Network layer interface (NLI)
  - Frame check sequence (FCS)
- Frame check sequence: This field contains the frame check sequence (FCS), which is calculated by using a cyclic redundancy check (CRC). The FCS allows Ethernet to detect errors in the Ethernet frame and reject the frame if the frame appears damaged.

4.1.3 How Ethernet works

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

All machines in the network are free to use the network whenever they choose if no one else is transmitting. This setup is known as *Multiple Access*.

You are required to have a means of ensuring that when two machines start to transmit data simultaneously, the resultant corrupted data is discarded. Also, retransmissions must be generated at differing time intervals. This assurance is known as *Collision Detection*.

Figure 4-2 shows a bus Ethernet network.

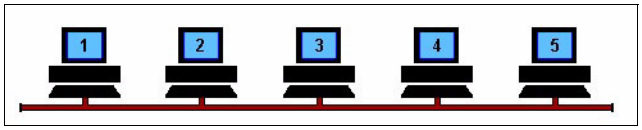


Figure 4-2 Bus Ethernet network

In Figure 4-2, assume that machine 2 wants to send a message to machine 4, but first it "listens" to make sure that no one else is using the network.

If the path is all clear, machine 2 starts to transmit its data on to the network. Each packet of data contains the destination address, the sender address, and the data to be transmitted.

The signal moves down the cable, and the signal is received by every machine in the network. But, because the signal it is only addressed to machine 4, the other machines ignore the signal. Machine 4 then sends a message back to machine 2 to acknowledge the 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 before it tries to transmit again. Figure 4-3 shows what happens when two machines transmit simultaneously.

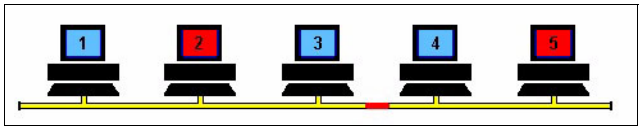


Figure 4-3 Machine 2 and machine 5 are both trying to transmit simultaneously

The resulting collision in Figure 4-3 destroys both signals. Each machine knows that this result happened because they do not "hear" their own transmission within a certain period. This time period is the *propagation delay*. The propagation delay is equivalent to the time that it takes for a signal to travel to the furthest part of the network and back again.

Both of the machines then wait for a random period before they try to transmit again. On small networks, this process all happens so quickly that it is virtually unnoticeable. However, as 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 on the applications that are used, but the general consideration is that 40 - 70 users are the limit before network speed is compromised.

Figure 4-4 shows two scenarios: hub and switch. The *hub* is where all of 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 use the media at a time.

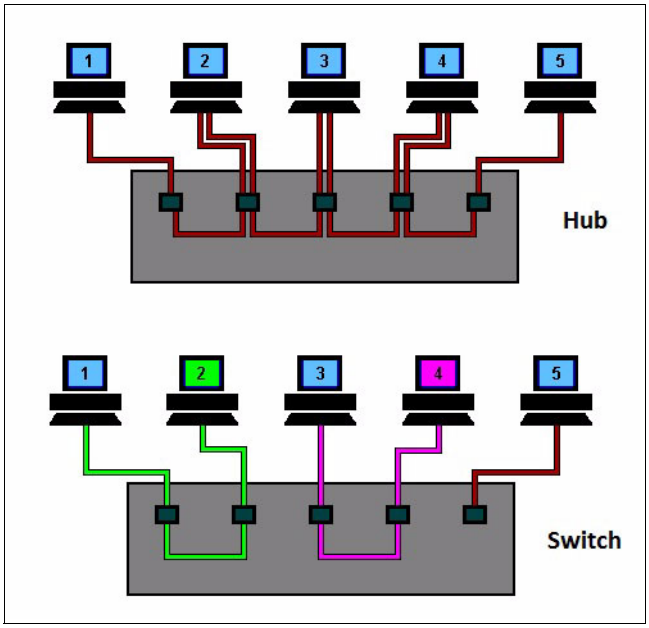


Figure 4-4 Hub and switch scenarios

An Ethernet hub changes the topology from a “bus” to a “star-wired bus”. As an example, assume 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.

Collisions can still occur but hubs have the advantage of centralized wiring, and they can automatically bypass any ports that are disconnected or have a cabling fault. This centralized wiring makes the network much more fault tolerant than a coax-based system where disconnecting a single connection shuts down the whole network.

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 in the network.

This *point-to-point switching* alleviates the problems that are associated with collisions and considerably improves the network speed.

4.1.4 Speed and bandwidth

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

The following numbers are simplex data rates, which might conflict with the duplex rates that 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:

- 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

These figures go against the traditional use of binary prefixes for memory size. These decimal prefixes are established in data communications.

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

Table 4-1 Technology rates and medium

Technology	Rate (Bit per second)	Rate (Byte per second)	Media
Fast Ethernet (100BASE-X)	100 Mbps	12.5 MBps	UTP Cat 5
Gigabit Ethernet (1000BASE-X)	1000 Mbps	125 MBps	UTP Cat 5e/6
10 Gigabit Ethernet (10GBASE-X)	10000 Mbps	1250 MBps	UTP Cat 7 - fiber

4.1.5 10 GbE

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

The IEEE 802.3ae 2002 (the 10 Gigabit Ethernet (10 GbE) standard) differs in certain respects from earlier Ethernet standards in that it operates only in full-duplex mode. (Collision-detection protocols are unnecessary.)

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

The 10 Gigabit Ethernet technology continues the evolution of Ethernet in terms of speed and distance, while it retains the same Ethernet architecture that is used in other Ethernet specifications. However, one key exception exists. Because 10 Gigabit Ethernet is a full-duplex-only technology, it does not need the CSMA/CD protocol that is used in other Ethernet technologies. In every other respect, 10 Gigabit Ethernet matches the original Ethernet model.

4.1.6 10 GbE copper versus fiber

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

Copper is the default standard for transmitting data between devices because of its low cost, easy installation, and flexibility. It also possesses distinct shortcomings. Copper is best when used in short lengths, typically 100 meters (328 feet) or less. When copper is employed over long distances, electromagnetic signal characteristics hinder performance. In addition, bundling copper cabling can cause interference, which makes it difficult to employ copper cabling as a comprehensive backbone. For these reasons, copper cabling is the principal data carrying technique for communication among personal computers and LANs, but not campus or long-distance transmission.

Conversely, fiber cabling is typically used for remote campus connectivity, crowded telecommunications closets, long-distance communications, and environments that need protection from interference. An example of this environment is a manufacturing area. Because fiber cabling is reliable and less susceptible to attenuation, fiber cabling is optimum for sending data beyond 100 meters (328 ft).

However, fiber costs more than copper. Therefore, the use of fiber cabling is typically limited to those applications that demand it.

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

The IEEE 802.3 Higher Speed Study Group formed in 1998, and the 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 10 times faster than the Gigabit Ethernet.

Subsequent standard updates ensued in relation to the first 10GigE version that was published in 2002. The IEEE 802.3ae-2002 fiber and the 802.3ak-2004 in 2004 were later consolidated into *IEEE 802.3-2005* in 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, which were known as *802.3aq-2006*. Finally, in 2007, the *802.3ap-2007* with copper backplane evolved.

As a result of these standards, two major types of 10 Gigabit Ethernet cabling, fiber and copper, are available.

The following standards apply to the 10 Gigabit Ethernet fiber cabling:

- 10GBASE-LX4: This standard supports ranges of 240 meters - 300 meters (790 ft - 980 ft) over traditional multi-mode cabling. This range is achieved by using four separate laser sources that operate at 3.125 Gbps in the range of 1300 nm on unique wavelengths. The 10GBASE-LX4 standard also supports 10 kilometers (6.2 miles) over System Management Facilities (SMF).
- 10GBASE-SR: Over obsolete 62.5 micron multi-mode fiber cabling (OM1), this standard has a maximum range of 26 meters - 82 meters (85 ft - 269 ft), depending on the cable type. Over standard 50 µm 2000 MHz·km OM3 multi-mode fiber (MMF), this standard has a maximum range of 300 meters (980 ft).
- 10GBASE-LR: This standard has a specified reach of 10 kilometers (6.2 miles), but 10GBASE-LR optical modules can often manage distances of up to 25 kilometers (16 miles) with no data loss.

- 10GBASE-LRM: This standard supports distances up to 220 meters (720 ft) on FDDI-grade 62.5 µm MMF. This fiber was originally installed in the early 1990s for Fiber Distributed Data Interface (FDDI), 100BaseFX networks, and for 260 meters (850 ft) on OM3. The reach of 10GBASE-LRM is not as far as the older 10GBASE-LX4 standard.
- 10GBASE-ER: This extended range has a reach of 40 kilometers (25 miles).
- 10GBASE-ZR: Several manufacturers introduced 80 km (49.7 miles) range ER pluggable interfaces under the name *10GBASE-ZR*. This 80 km (49.7 miles) PHY is not specified within the IEEE 802.3ae standard. Manufacturers created their own specifications that are based on the 80 km (49.7 miles) PHY that is described in the OC-192/STM-64 SDH/SONET specifications.

A 10G Ethernet connection can also run over Twinax cabling and twisted-pair cabling. The following standards apply to the 10 Gigabit Ethernet copper cabling:

- 10GBASE-CX4: This standard was the first 10G copper standard that was 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 gigabaud (Gbaud) of signaling bandwidth.
- 10GBASE-T or IEEE 802.3an-2006: This standard was released in 2006 to provide 10 Gbps connections over unshielded or shielded twisted-pair cables, over distances up to 100 meters (328 ft).

Cables that are needed to carry 10GBASE-T: Category 6A, or better, of balanced twisted-pair cables that are specified in ISO 11801 amendment 2 or ANSI/TIA-568-C.2 are needed to carry 10GBASE-T up to distances of 100 m (328 ft). Category 6 cables can carry 10GBASE-T for shorter distances when it is qualified, according to the guidelines in ISO TR 24750 or TIA-155-A.

The following standards refer to the 10 Gigabit Ethernet copper backplane cabling:

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

4.1.7 Virtual local area network

A virtual local area network (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 (Figure 4-5).

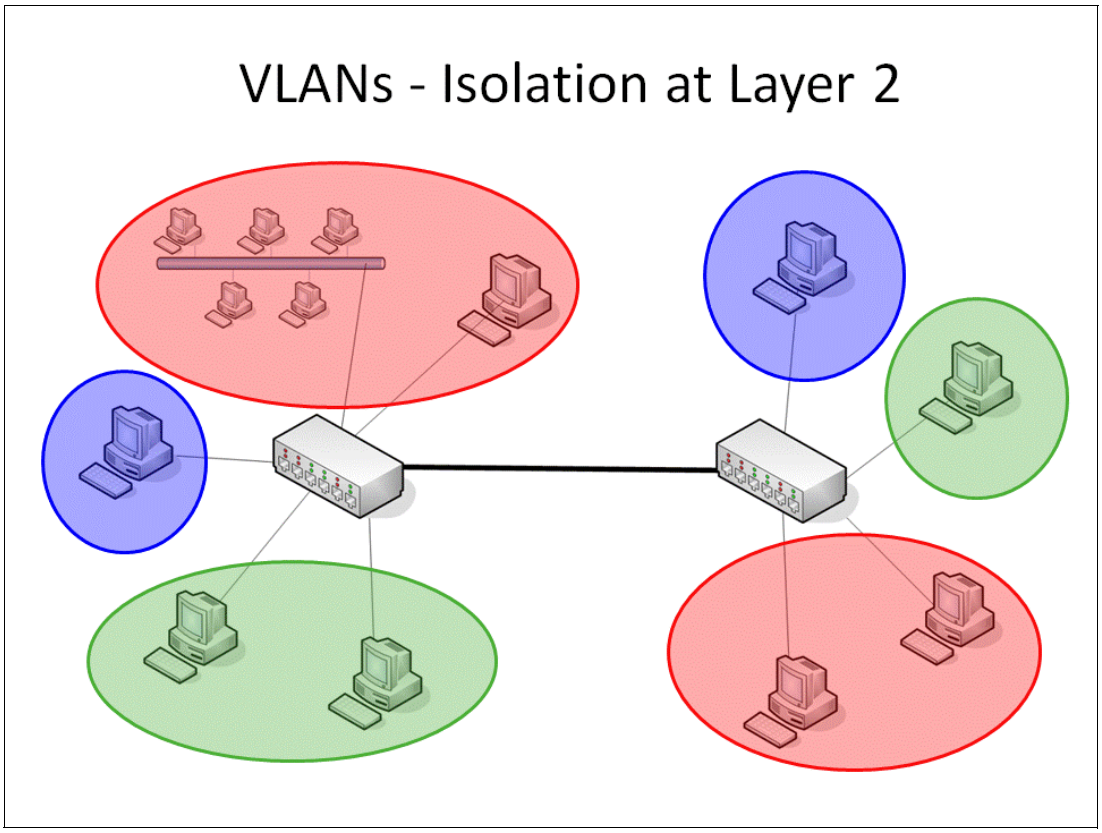


Figure 4-5 Isolation at Layer 2

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

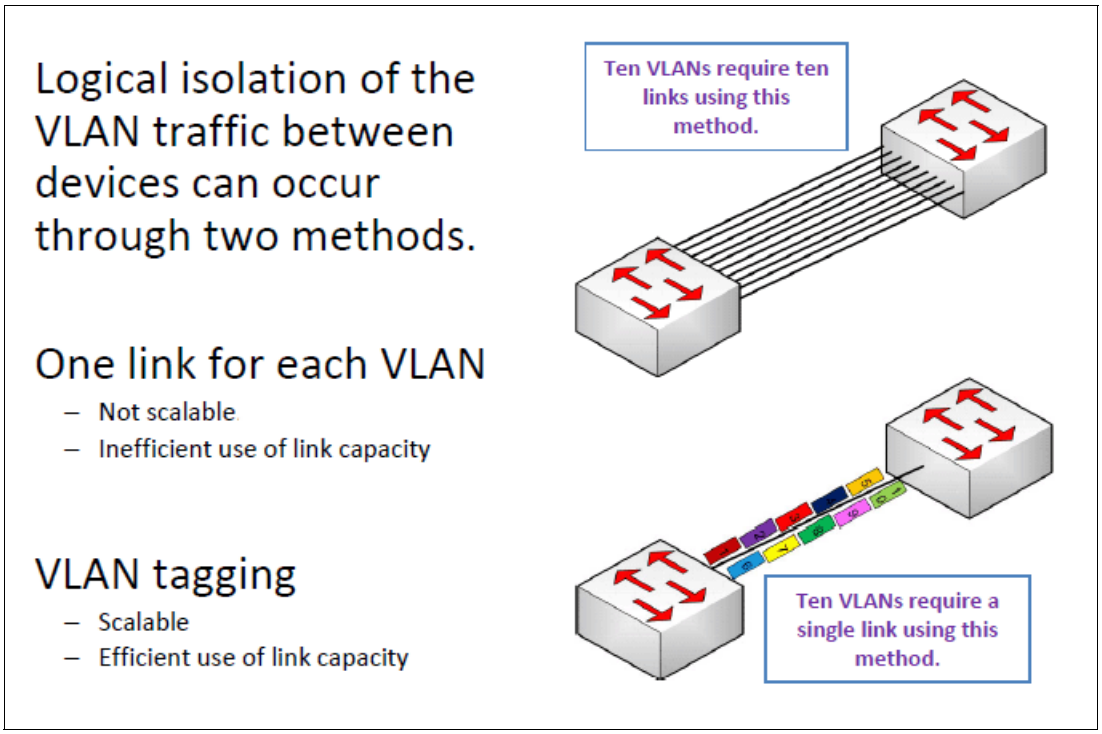


Figure 4-6 VLAN tagging

The first method uses a single link for each VLAN. This method does not scale well because it uses many ports in networks that have 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. This configuration 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 802.1p/q* standard (priority tagging and VLAN identifier tagging).

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 information, such as VLAN membership and priority, that is added to the frame (Figure 4-7).

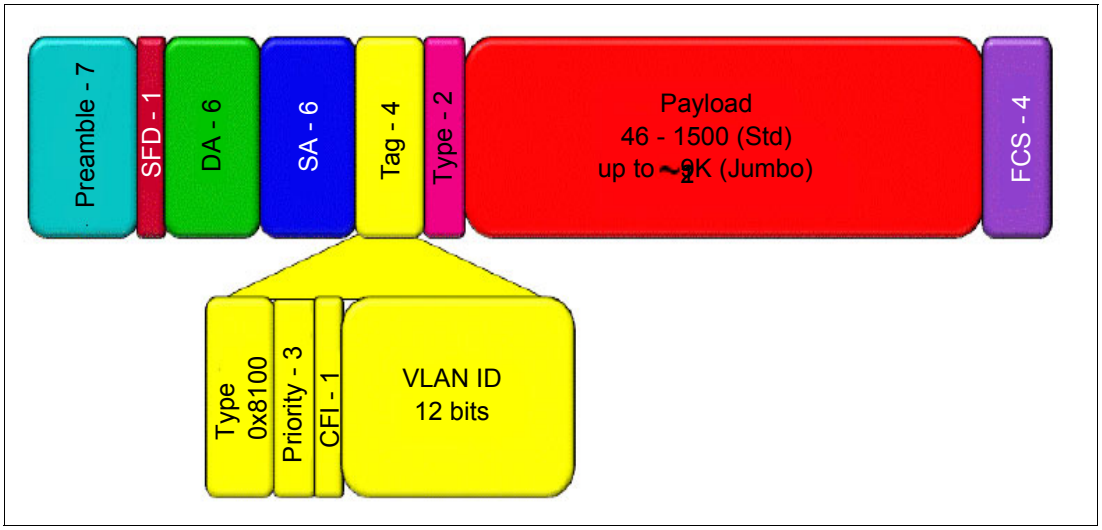


Figure 4-7 IEEE 802.1p/q tagged Ethernet frame

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

The 4-byte tag has four component fields:

- The *type field* is 2 bytes and has the hexadecimal value of x8100 to identify the frame as an 802.1p/q tagged frame.
- The *priority field* is 3 bits and allows a priority value of eight different values to be included in the tag. This field has the “P” portion of the 802.1p/q standard.
- The *Canonical Format Indicator field* is 1 bit and identifies when the contents of the payload field are in canonical format.
- The *VLAN ID field* is 12 bits and identifies the VLAN that the frame is a member of, with 4,096 different VLANs that are possible.

4.1.8 Interface virtual local area network operation modes

Interfaces on a switch can operate in two virtual local area network (VLAN) modes: single VLAN mode or multiple VLAN mode.

Single virtual local area network mode

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

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

Multiple virtual local area network mode

The *multiple VLAN mode* operation is also referred to as *trunk mode*. A port that is 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 that is configured in this manner is transmitted with no tag. This method allows the receiving device to receive the frame in an untagged format.

This method provides compatibility with existing devices. Compatibility is also provided for devices that are configured in the single VLAN mode and that are attached to a port that is configured as a multi-VLAN port.

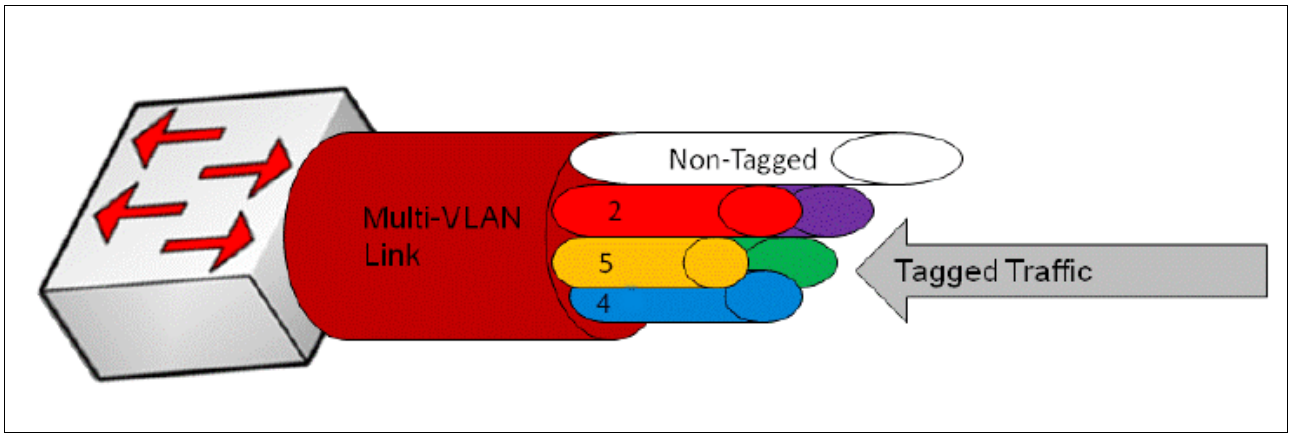


Figure 4-8 Multiple VLAN mode link

Variations in the meaning of trunk: The term *trunk* is used to express different ideas in the networking industry. When you use this term, remember that other individuals might use the term in a different manner. Trunk can mean that a port is 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).

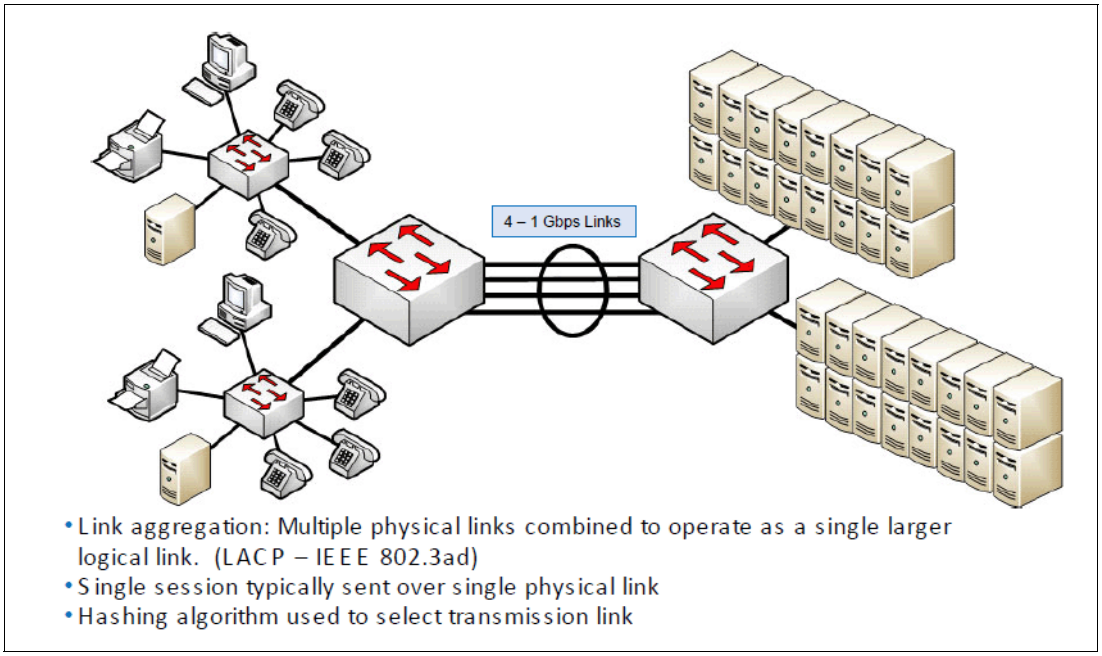


Figure 4-9 Link aggregation

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 the devices that is interconnected by the aggregated link uses a hashing algorithm to determine on which of the member links the frames will be transmitted. The hashing algorithm might use various 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. STP is available in different forms, such as existing STP, Rapid STP (RSTP), Multiple STP (MSTP), and VLAN STP (VSTP). RSTP is a common default STP. This form provides faster convergence times than STP. However, certain existing networks require the slower convergence times that are provided by basic STP.

The operation of Spanning Tree Protocol

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.

The following 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 TCN for its own root port. The process continues until the TCN reaches the root bridge.

STP uses the information that is provided by the BPDUs in several ways:

- To elect a root bridge
- To identify root ports for each switch
- To identify designated ports for each physical LAN segment
- To prune specific redundant links to create a loop-free tree topology

All leaf devices calculate the best path to the root device. The devices place their ports in blocking or forwarding states that are 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 reconvergence 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:

- The root port is the “best path” to the root device.
- The designated port indicates that the switch is the designated bridge for the other switch that is connecting to this port.
- The alternative port provides an alternative root port.
- The backup port provides an alternative 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 that is inherent in STP. This inherent issue is 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 (MSTIs). 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 1 - 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 that are 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.

Bridge Protocol Data Unit 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 that are 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 the interface stops forwarding the 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 a forwarding state in error if the interface stops receiving BPDUs from its designated port on the segment. This transition error can occur when a hardware error is on the switch or a software configuration error is 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 root ports and blocked ports 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 it transitions the interface 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 the most effective when it is enabled on 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 that is elected through the regular process has the possibility of being wrongly elected. A user bridge application that is running on a personal computer 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 must not receive superior BPDUs from the root bridge and must 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 enabled root protection, that port transitions to a root-prevented STP state (inconsistency state), and the interface is blocked. This blocking prevents a bridge that must 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. This state is 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 of 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 about 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 Power over Ethernet (PoE) policy. With the PoE Management TLVs, the switch ports can advertise the power level and power priority that is needed. For example, the switch can compare the power that is needed by an IP telephone that is running on a PoE interface with available resources. If the switch cannot meet the resources that are 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, the 802.1p class of service (COS) and 802.1Q tag information can be sent to the IP telephone.

4.1.12 Link Layer Discovery Protocol Type Length Values (LLDP TLVs)

The basic TLVs include the following information:

- Chassis identifier: The MAC address that is 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. This 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 contains information about the software and the current image that are running on the system. This information is not configurable, but it is taken from the software.
- System capabilities: The primary function that is performed by the system. The capabilities that are supported by the system, for example, bridge or router. This information is not configurable, but it 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 by way of medium dependent interface (MDI): A TLV that advertises MDI power support, a 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, support, and multistation access unit (MAU) type. The information is not configurable, but it is based on the physical interface structure.
- Link aggregation: A TLV that advertises whether the port is aggregated and its aggregated port ID.
- Maximum frame size: A TLV that advertises the maximum transmission unit (MTU) of the interface that is sending LLDP frames.
- Port VLAN: A TLV that advertises the VLAN name that is configured on the interface.

LLDP-MED provides the following TLVs:

- LLDP MED capabilities: A TLV that advertises the primary function of the port. The capability values range 0 - 15. The device class values range 0 - 255:
  - 0: Capabilities
  - 1: Network policy
  - 2: Location identification
  - 3: Extended power by way of MDI-PSE
  - 4: Inventory
  - 5 - 15: Reserved
- LLDP-MED device class values:
  - 0: Class not defined
  - 1: Class 1 device
  - 2: Class 2 device
  - 3: Class 3 device
  - 4: Network connectivity device
  - 5 - 255: Reserved
- Network policy: A TLV that advertises the port VLAN configuration and associated Layer 2 and Layer 3 attributes. The following attributes are included:
  - Policy identifier
  - Application types, such as voice or streaming video
  - 802.1Q VLAN tagging
  - 802.1p priority bits
  - Diffserv code points
- Endpoint location: A TLV that advertises the physical location of the endpoint.
- Extended power by way of 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 Storage area network IP networking

Now that we 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 the benefits that these technologies can offer? The following list provides a few of the benefits that you can gain:

- 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 a large financial investment without knowing that they will get a return. The appeal of these protocols is that they immediately provide benefits. Because these protocols are standards-based protocols, they allow the use of both the existing TCP/IP and Fibre Channel Protocol (FCP) infrastructure, they support existing Fibre Channel devices, and they enable the simplification of the infrastructure by removing any SAN islands.

4.2.1 The multiprotocol environment

Any technology comes with its unique jargon and terminology. Typically, a term is borrowed from the networking world, but it might have a separate meaning. It is not our intent to cover every unique description. However, we make several distinctions that we think 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, 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.

Each separate fabric has its own addressing scheme. When fabrics are joined by a router, a way to translate the addresses between the two fabrics must exist. This mechanism is called *network address translation* (*NAT*), and it is inherent in all of the IBM System Storage multiprotocol switch/router products. NAT is sometimes referred to as *FC-NAT* to differentiate it from a similar mechanism that exists in IP routers.

4.2.4 Tunneling

Tunneling is a technique that allows one network to send its data through the connection of another network. Tunneling works by encapsulating a network protocol within packets that are 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 is common usage to broaden the term *router* to include these functions. FCIP is an example of tunneling. Internet Small Computer System Interface (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 an IP network. iSNS provides storage discovery and management services that are comparable to the services that are in Fibre Channel networks. Therefore, the IP network seems to operate in a similar capacity as a SAN. Coupling this capability with its ability to emulate Fibre Channel fabric services, iSNS allows for a transparent integration of IP and Fibre Channel networks. This integration is possible because it can manage both iSCSI and Fibre Channel devices.

4.3 Delving deeper into the protocols

We introduced all of the protocols at a high level. Now, in greater depth, we show the methods by which the protocols handle Fibre Channel traffic.

4.3.1 Fibre Channel over Internet Protocol (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.

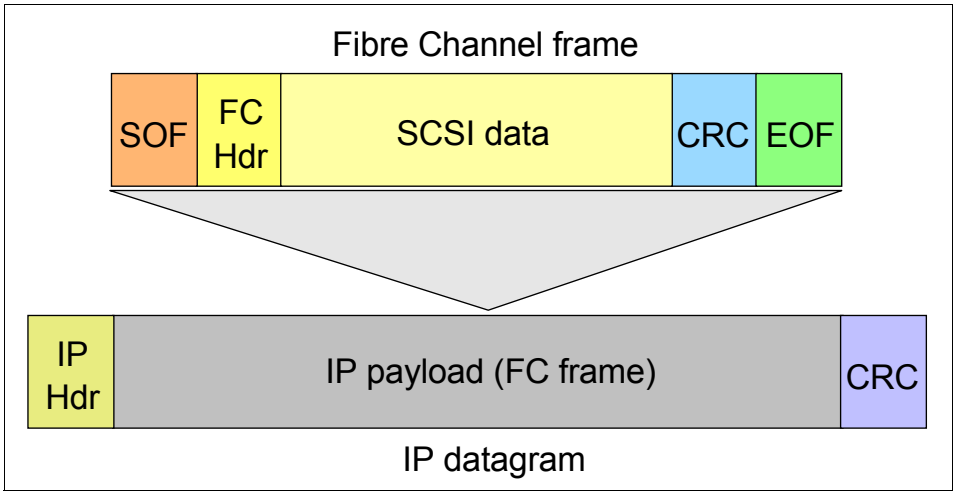


Figure 4-10 FCIP encapsulates the Fibre Channel frame into IP packets

The major advantage of FCIP is that it overcomes the distance limitations of basic Fibre Channel. It also enables geographically distributed devices to be linked by using the existing IP infrastructure, while it keeps the fabric services intact.

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

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

Because FCIP simply tunnels Fibre Channel, creating an FCIP link is similar to creating an ISL. And, the two fabrics at either end are merged into a single fabric. This merger 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 renegotiate and the fabric is reconfigured.

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

4.3.2 Internet Fibre Channel Protocol

Internet Fibre Channel Protocol (iFCP) is a gateway-to-gateway protocol. It provides Fibre Channel fabric services to Fibre Channel devices over an IP network. iFCP uses TCP to provide congestion control, error detection, and recovery. The primary purpose of iFCP is to allow interconnection and networking of existing Fibre Channel devices at wire speeds over an IP network.

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

To learn more about the architecture and specification of iFCP, see the document at this website:

http://tools.ietf.org/html/draft-ietf-ips-ifcp-14

A myth exists that iFCP does not use encapsulation. In fact, iFCP encapsulates the Fibre Channel packet in much the same way that FCIP encapsulates the Fibre Channel packet. In addition, iFCP maps the Fibre Channel header to the IP header and a TCP session (Figure 4-11).

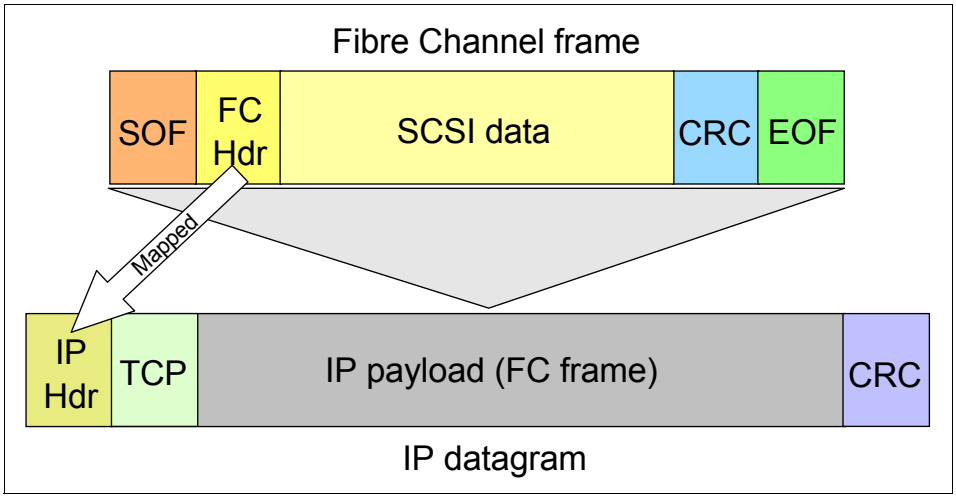


Figure 4-11 iFCP encapsulation and header mapping

iFCP uses the same iSNS mechanism that is used by iSCSI.

iFCP also allows data to fall across IP packets and share IP packets. Certain FCIP implementations can achieve a similar result when they run software compression, but not otherwise. FCIP typically breaks 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 for each *fabric login* (*FLOGI*). FCIP typically uses one connection for each router link (although more connections are possible). A FLOGI is the process by which a node port (N_port) registers its presence on the fabric; obtains fabric parameters, such as classes of service that are supported; and receives its N_port address. Because a separate TCP connection exists for each N_port to N_port couple under iFCP, you can manage each connection to have its own quality of service (QoS) identity. A single incidence of congestion does not have to drop the sending rate for all connections on the link.

Although all iFCP traffic between a specific remote and local N_port pair must use the same iFCP session, the iFCP session can be shared across multiple gateways or routers.

4.3.3 Internet Small Computer System Interface

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

The SCSI protocol is 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 the distance capabilities of each transport are limited.

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. This session ID is a combination of an initiator part (ISID) and a target part (Target Portal Group Tag).

The iSCSI transfer direction is defined in relation 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 might be shipped together from initiator to target, and data and responses might 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 NICs are used, they generally all present the same iSCSI initiator name to the targets because they are 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)”, at this website:

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

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

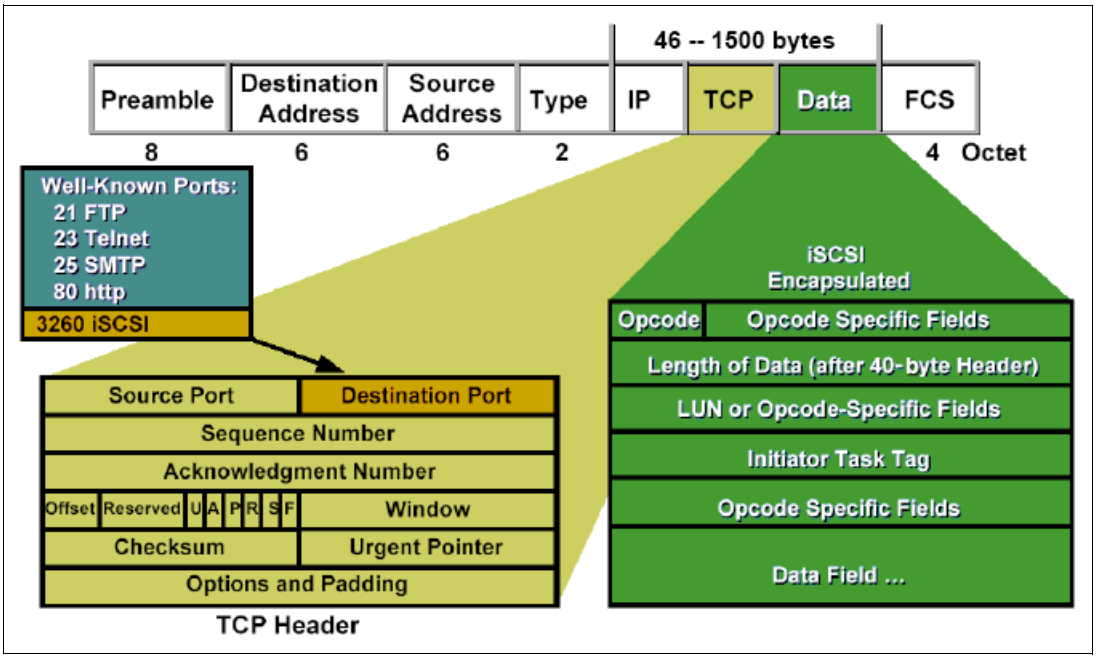


Figure 4-12 iSCSI packet format

Testing on iSCSI latency shows a difference of up to 1 ms of additional latency for each disk I/O as compared to Fibre Channel. This delay does not include factors, such as trying to perform iSCSI I/O over a shared, congested, or long-distance IP network, all of which might be tempting for certain clients. iSCSI generally uses a shared 1 Gbps network.

iSCSI naming and discovery

Although we do not go into an in-depth analysis of iSCSI in this book, an iSCSI initiator has methods to understand the devices that are in the network:

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

> **iSNS:** At time of writing this book, not all vendors can deliver iSNS.

For more information about drafts that describe iSCSI naming, discovery, and booting, see this website:

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

4.3.4 Routing considerations

As you might expect with any technology, you must consider a unique set of characteristics. The following topics briefly describe the issues or items to consider in a multiprotocol Fibre Channel environment.

4.3.5 Packet size

The standard size of a Fibre Channel packet is 2,148 bytes, and the standard IP packet size is 1,500 bytes (with a 1,460-byte payload). Because one packet is larger than the other packet, the larger packet must be accommodated.

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

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

Each technology and each vendor might implement this procedure differently. The key point is that they all try to avoid sending small, inefficient packets.

4.3.6 TCP congestion control

Sometimes, standard TCP congestion mechanisms might not be suitable for tunneling storage. Standard TCP congestion control is designed to react quickly and severely to network congestion, but to recover slowly. This design is suited to traditional IP networks that are variable and unreliable. But for storage applications, this approach is not always appropriate and this approach might 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, the sending rate performs a slow-start linear ramp-up again.

Certain vendors tweak the back-off and recovery algorithms. For example, the tweak causes the send rate to drop by 12.5% each time that congestion is encountered. And the algorithm is tweaked so that the network can recover rapidly to the full sending rate by doubling each time until the full rate is regained.

Other vendors take a simpler approach to achieve a similar outcome.

If you are sharing your IP link between storage and other IP applications, either of these storage-friendly congestion controls might affect your other applications.

For more information about the specification for TCP congestion control, see this website:

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

4.3.7 Round-trip delay

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

Fibre Channel cable has an inherent latency of approximately 5 microseconds for each kilometer each way. Typical Fibre Channel devices, such as switches and routers, have inherent latencies of around 5 microseconds each way. IP routers might vary 5 - 100 microseconds in theory, but when they are tested with filters that are applied, the results are more likely to be measured in milliseconds.

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

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

So, how does this latency 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 specific 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 2,112 bytes. The Fibre Channel headers add 36 bytes to this measurement, for a total Fibre Channel frame size of 2,148 bytes. When you transfer data, Fibre Channel frames at or near the full size are typically 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. This addition creates a total Ethernet packet size of 2,202 bytes, or 17,616 bits.

For smaller frames, such as the Fibre Channel acknowledgment 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 lists the transmission times of this FCIP packet over various 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
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 must be divided into two Ethernet packets. This requirement 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 that move 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 subfabrics

Assume 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 and test environment, two of the switches are used to connect a joint-venture subsidiary company, and four of the switches are used to connect the main production environment.

The development and 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 and restore environment is shared among the three environments.

In summary, this environment requires a degree of isolation, and a degree of sharing. In the past, this requirement was accommodated through zoning. Certain fabric vendors might 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. These features are beneficial in creating many simple fabrics, rather than a few more complex fabrics.

4.4.2 Connecting a remote site over IP

Suppose that you want to replicate your disk system to a remote site, perhaps 50 km (31.06 miles) away synchronously, or 500 km (310.68 miles) away asynchronously. By 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 network devices, or switches and IP ports to connect to a standard IP WAN router. Standard IP networks are generally much lower in cost to provision than traditional high-quality dedicated *dense wavelength division multiplexing* (*DWDM*) networks. Standard IP networks also often have the advantage of being 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.

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

4.4.3 Connecting hosts by using Internet Small Computer System Interface

Many hosts do not require high-bandwidth, low-latency access to storage. For such hosts, Internet Small Computer System Interface (iSCSI) might be a more cost-effective connection method. iSCSI can be thought of as an IP SAN. You are not required to provide a Fibre Channel switch port for every server. You do not need to purchase Fibre Channel host bus adapters (HBAs), or to lay Fibre Channel cable between the storage and servers.

The iSCSI router has both Fibre Channel and Ethernet ports to connect to servers that are located either locally on the Ethernet, or remotely, over a standard IP WAN connection.

The iSCSI connection delivers block I/O access to the server so that it is application independent. That is, an application cannot really tell the difference between direct SCSI, iSCSI, or Fibre Channel, because all three I/Os 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 workload can be mitigated by using the TCP/IP offload engine (TOE) Ethernet cards. However, because these cards can be expensive, they somewhat undermine the low-cost advantage of iSCSI.

iSCSI provides low-cost connections: 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 introduce Fibre Channel (FC) topologies and other fabric services that are commonly encountered in a storage area network (SAN).

We also provide an insight into the emerging converged topology and the option to merge FC to Fibre Channel over Ethernet (FCoE).

5.1 Fibre Channel topologies

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

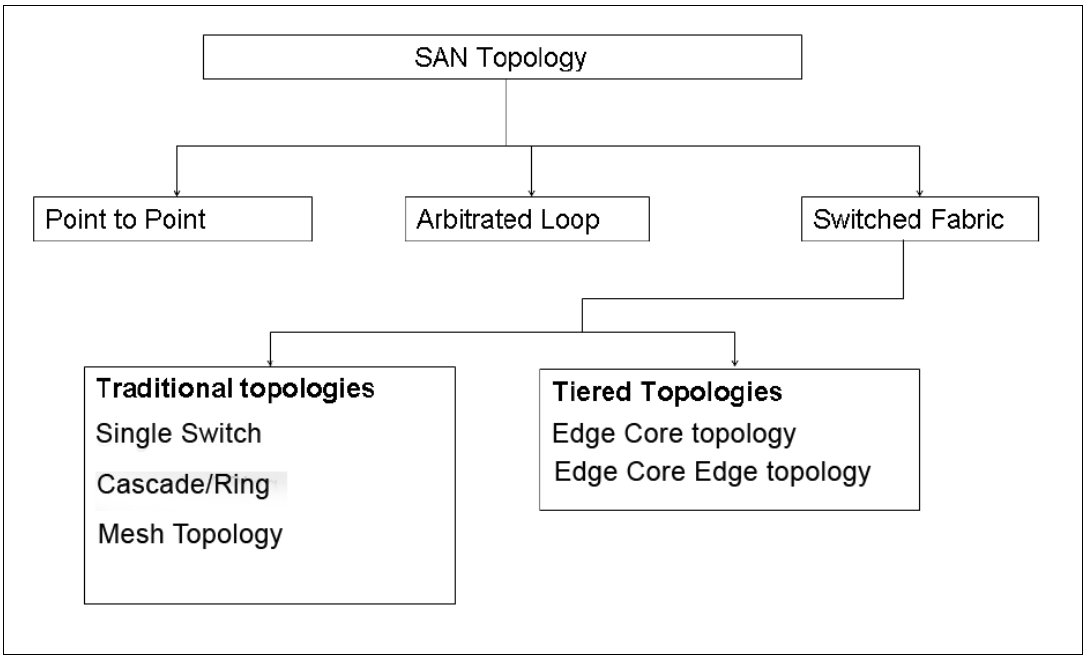


Figure 5-1 SAN topologies

5.1.1 Point-to-point topology

A *point-to-point connection* is the simplest topology. It is used when exactly two nodes exist, and future expansion is not predicted. Media is not shared, 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 that both paths transmit data simultaneously. For example, Fibre Channel connections that are based on the 1 Gbps standard can transmit at 100 Megabytes per second (MBps) and receive at 100 MBps simultaneously. For Fibre Channel connections that are based on the 2 gigabits per second (Gbps) standard, they can transmit at 200 MBps and receive at 200 MBps simultaneously. This speed also extends to 4 Gbps, 8 Gbps, and 16 Gigabytes per second (GBps) technologies.

Figure 5-2 shows a simple point-to-point connection.

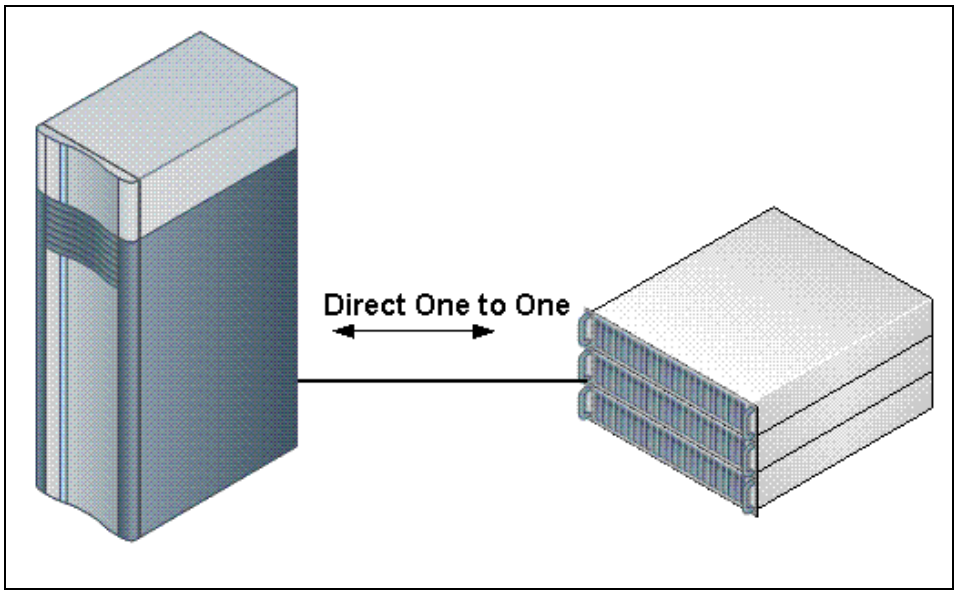


Figure 5-2 Point-to-point connection

5.1.2 Arbitrated loop topology

Arbitrated loop topology: Although this topology is rarely encountered anymore and considered a historical 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 node loop ports (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 topology 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 might be established. Loops can be configured with hubs to make connection management easier. A distance of up to 10 km (6.2 miles) 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, which is configured by using a hub (Figure 5-3).

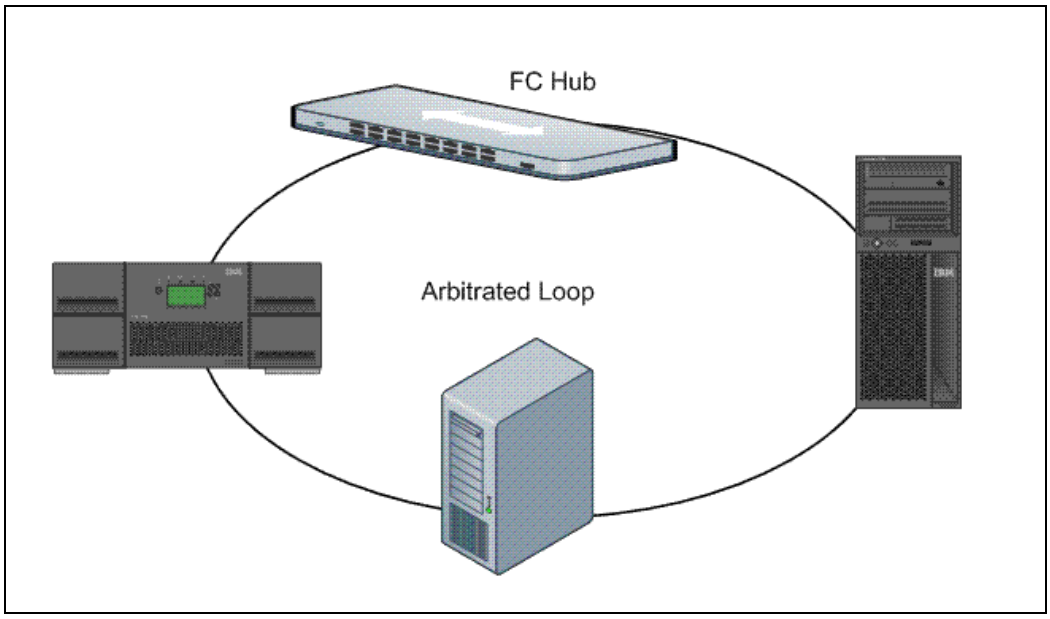


Figure 5-3 Arbitrated loop

We describe FC-AL in more depth in 5.4, “Fibre Channel Arbitrated Loop protocols” on page 114.

5.1.3 Switched fabric topology

Our third topology, and the most useful topology that is 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 for each port that is compared to the shared bandwidth for each port in arbitrated loop implementations.

One key differentiator is that if you add a device into the arbitrated loop, you further divide the shared bandwidth. However, in a switched fabric, adding a device or a new connection between existing devices or connections actually increases the bandwidth. For example, an 8-port switch (assume that 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. This total equates to 1,200 MBps if full-duplex applications are available.

Figure 5-4 shows a switched fabric configuration.

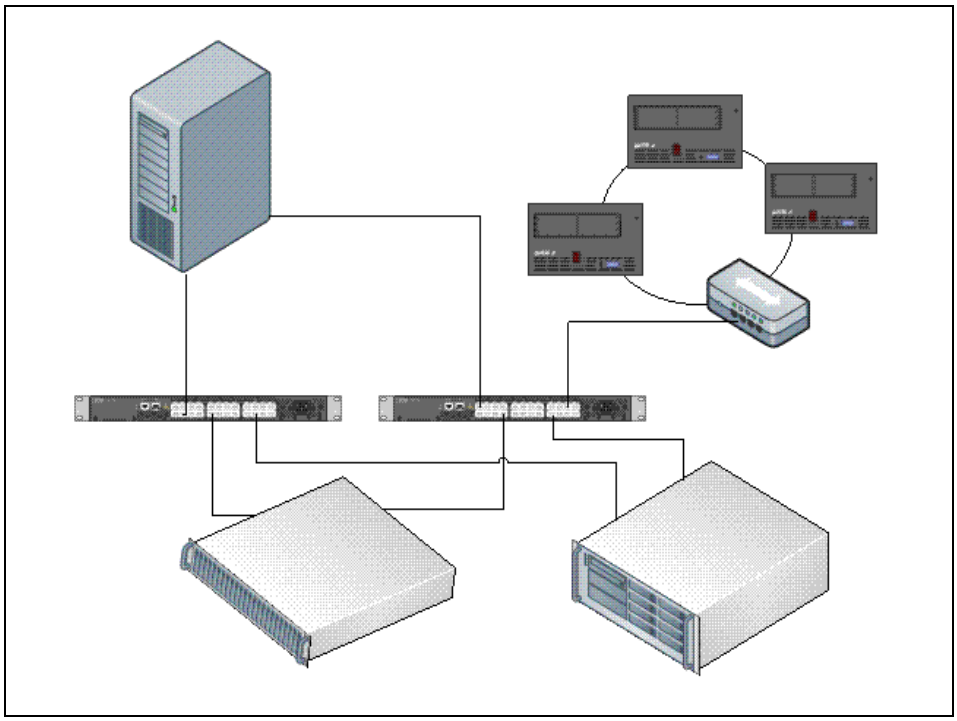


Figure 5-4 Sample switched fabric topology

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

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

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

5.1.4 Single switch topology

The *single switch topology* has only one switch and no *inter-switch links* (*ISLs*). It is the simplest design for infrastructures that do not need any redundancy. Because of the issues with this topology introducing a single point of failure (SPOF), this topology is rarely used.

Figure 5-5 indicates a single switch topology with all of the devices connected to same switch.

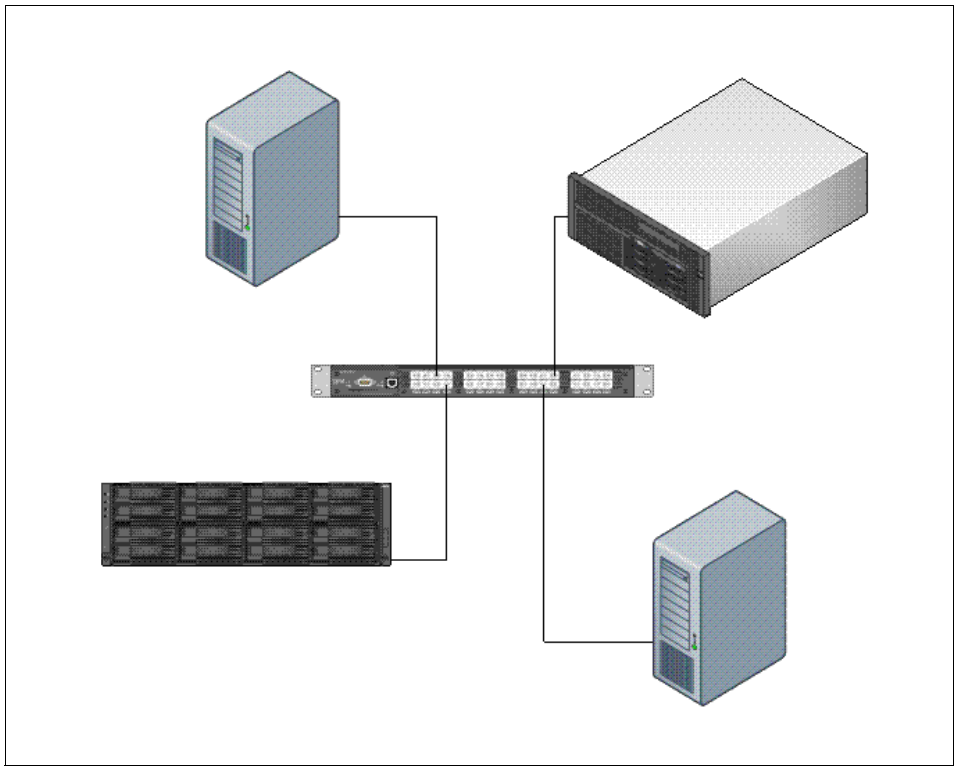


Figure 5-5 Single switch topology

5.1.5 Cascaded and ring topology

In a *cascaded topology*, switches are connected in a *queue fashion* (Figure 5-6).

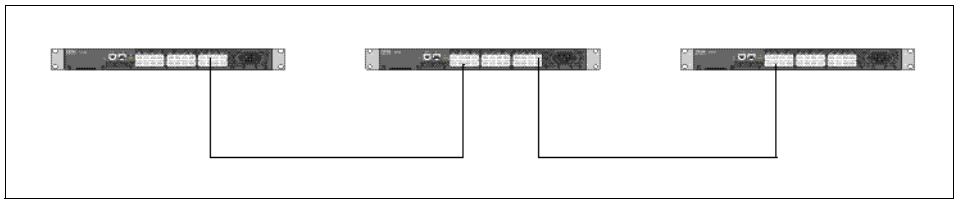


Figure 5-6 Cascade topology

Even in a *ring topology*, the switches connect in a queue fashion, but the ring topology forms a closed ring with an additional ISL (Figure 5-7).

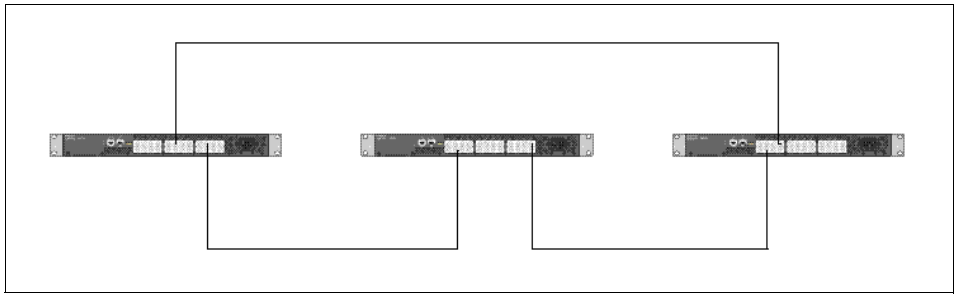


Figure 5-7 Ring topology

5.1.6 Mesh topology

In a full *mesh topology,* each switch is connected to every other switch in the fabric (Figure 5-8).

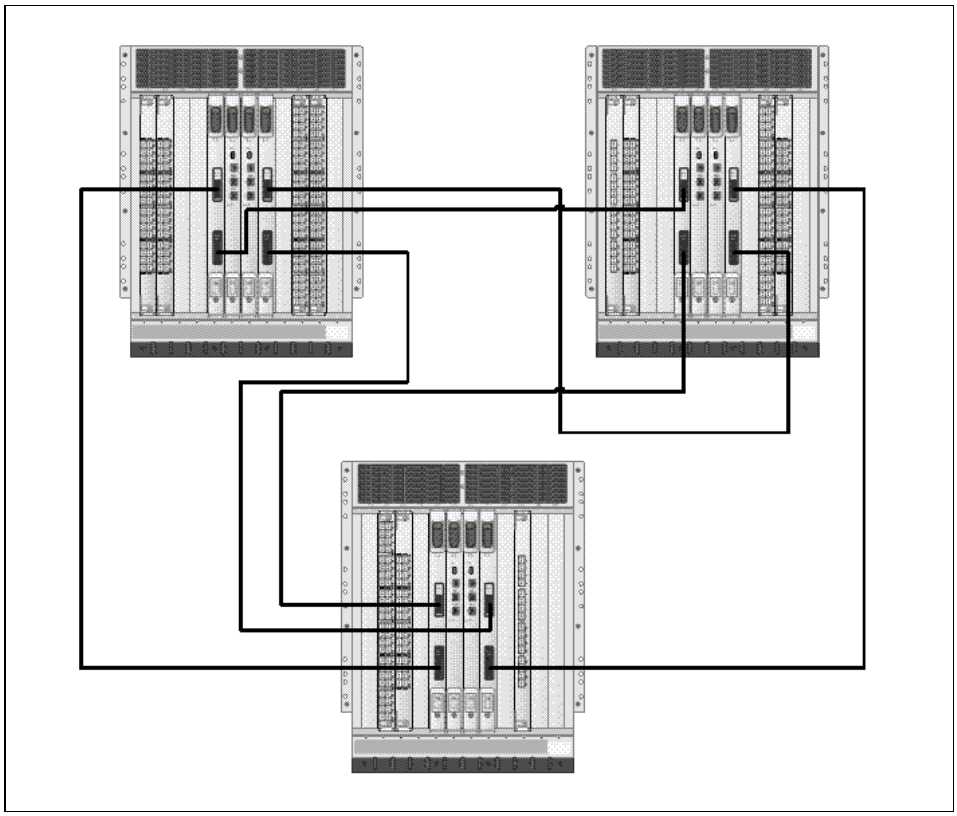


Figure 5-8 IBM SAN768B connected to form a mesh topology

In terms of a tiered approach, the switched fabric can be further classified with the following topologies:

- Core-edge topology
- Edge-core-edge topology

5.1.7 Core-edge topology

In *core-edge topology*, the servers are connected to the edge fabric and the storage is connected to core switches. Figure 5-9 shows the core-edge topology.

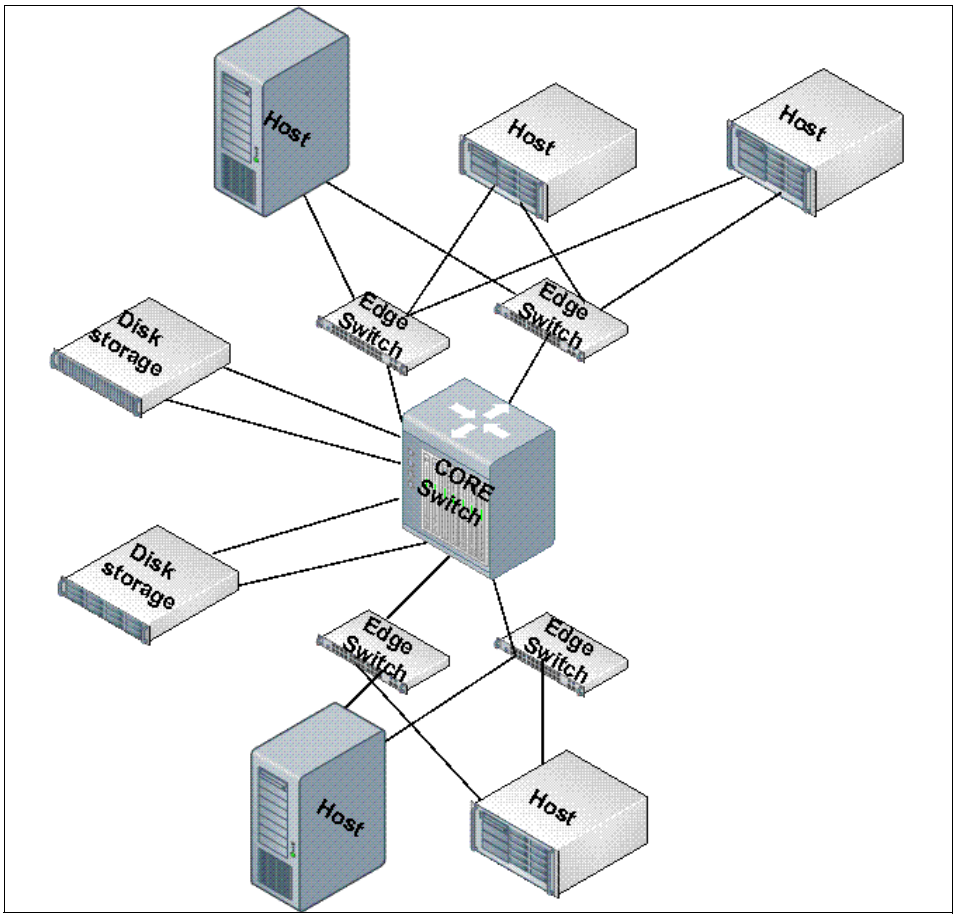


Figure 5-9 Core-edge topology

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. This configuration expands the SAN traffic flow to long distance by dense wavelength division multiplexing (DWDM), connecting to virtualization appliances, and encryption switches. Also, the servers might be isolated to one edge and storage can be at the other edge, which helps with management.

Figure 5-10 shows the edge-core-edge topology.

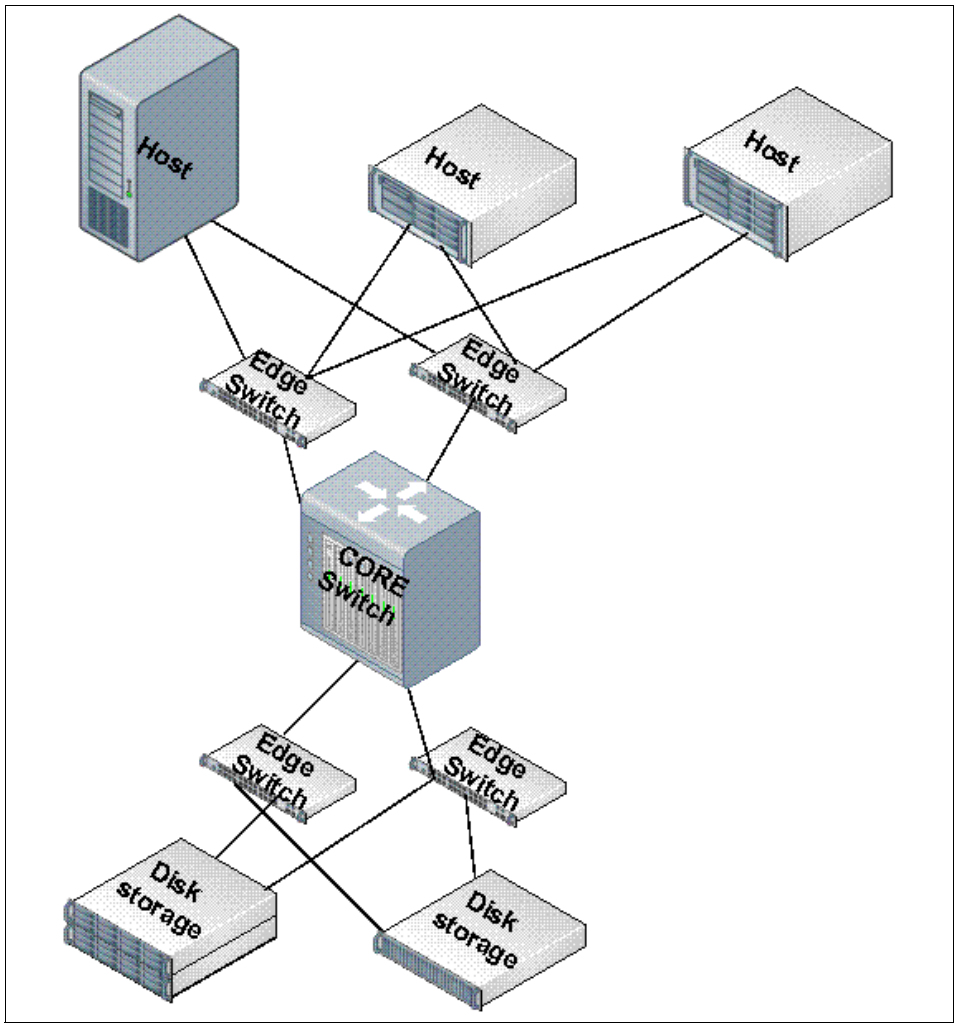


Figure 5-10 Edge-core-edge topology

5.2 Port types

The basic building block of the Fibre Channel is the *port*. Various kinds of Fibre Channel port types exist.

5.2.1 Common port types

The following list provides the various kinds of Fibre Channel port types and their purpose in switches, servers, and storage:

- F_port: This port type is a fabric port that is connected to a node port (N_port) in a point-point manner to a switch.
- FL_port: This port type is a fabric loop port that is connected to a loop device. It is used to connect a node loop port (NL_port) to the switch in a public loop configuration.
- TL port: A Cisco-specific translative loop port type. It is a translative loop port that is connected with non-fabric-aware, private loop devices.

- G_port: This generic port type can operate as either an expansion port (E_port) or an F_port. A port is defined as a G_port after it connects, but it did not receive a response to *loop initialization,* or it did not yet complete the link initialization procedure with the adjacent Fibre Channel device.
- L_port: This loop port type is a loop-capable node or switch port.
- U_port: This type 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 it did not yet assume a specific function in the fabric.
- N_port: This port type is a node port that is not loop capable. It is a host end port that is used to connect to the fabric switch.
- NL_port: This port type 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 shows the common port types of the switch and nodes.

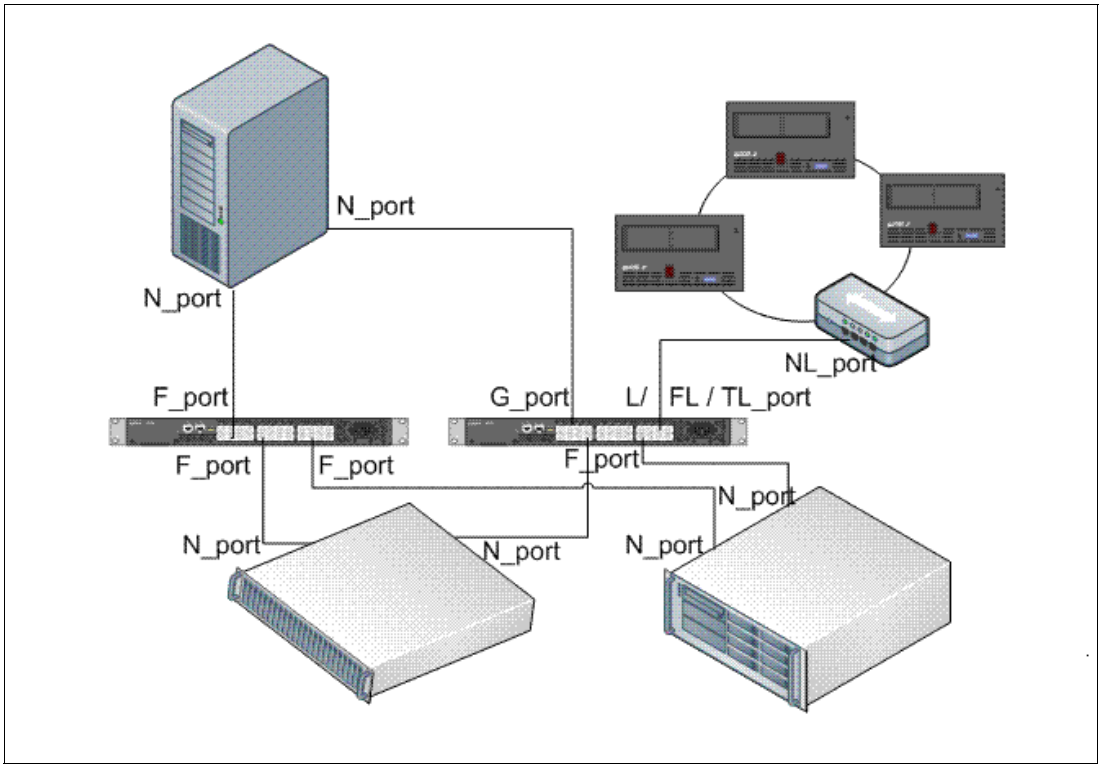


Figure 5-11 Common port types

5.2.2 Expansion port types

The following ports are found in a multi-switch fabric where switches are interconnected through an FC link:

- E_Port: This type is an expansion port. A port is designated as an E_Port when it is used as an ISL to connect to the E_Port of another switch to enlarge the switch fabric.
- EX_Port: This type of E_Port is used to connect a multiprotocol router to an edge fabric. An EX_Port follows standard E_Port protocols. An EX_Port supports Fibre Channel Network Address Translation (FC-NAT), but it does not allow fabric merging across EX_Ports.

- VE_Port: A virtual E_Port is a port that emulates an E_Port over a Fibre Channel over Internet Protocol (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 behave and function in the same way.
- TE_port: The TE_port provides standard E_port functions and it also allows for the routing of multiple virtual SANs (VSANs). This capability is accomplished by modifying the standard Fibre Channel frame (VSAN tagging) on ingress and egress of the VSAN environment. It is also known as a Trunking E_port.

Figure 5-12 shows a fabric with expansion ports.

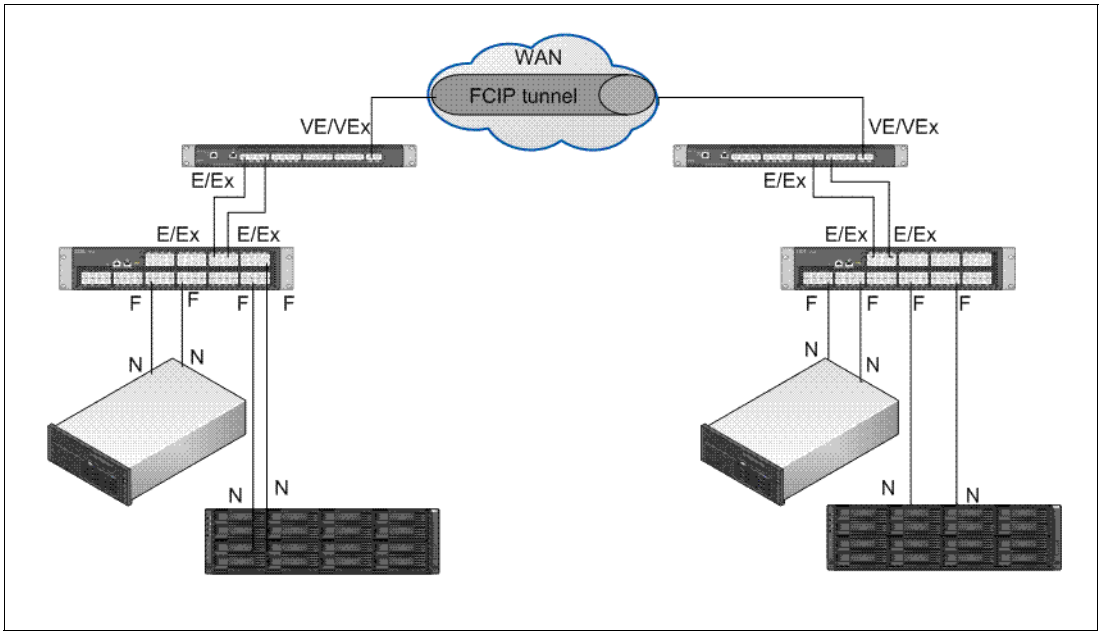


Figure 5-12 Fabric with expansion ports

5.2.3 Diagnostic port types

D_port is a diagnostic port type that can be enabled only on the 16 Gbps b-type switches with Fabric Operating System 7.0. This system uses the *Spinfab* test. It performs electrical loop back and optical loop back. It measures link distance and it also performs stress tests with a link saturation test.

Figure 5-13 shows the various test options. You can perform long-distance cable checks also with D_port diagnostic capabilities.

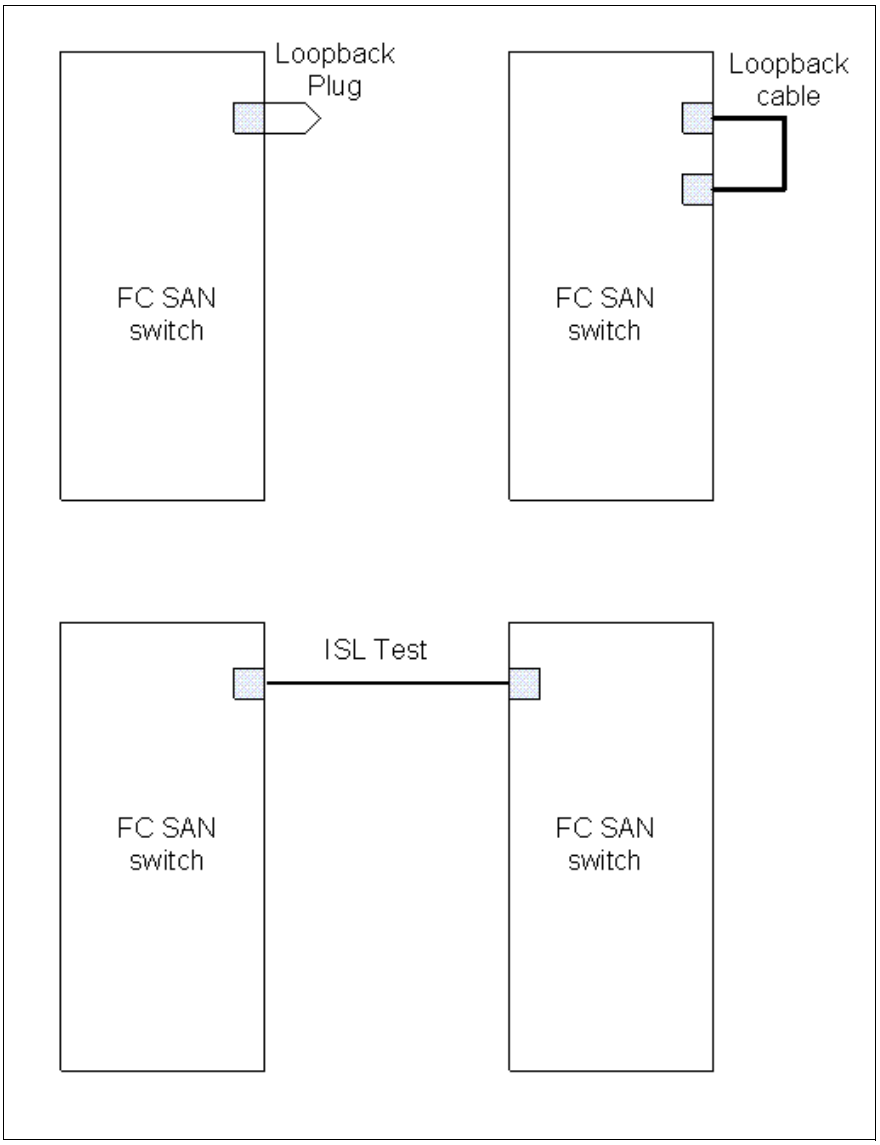


Figure 5-13 D_port type diagnostics

Additional diagnostic port types are shown:

- MTx_port is a CNT port that is used as a mirror for viewing the transmit stream of the port to be diagnosed.
- MRx_port is a CNT port that is used as a mirror for viewing the receive stream of the port to be diagnosed.
- SD_port is a Cisco switched port analyzer (SPAN) destination diagnostic port that is used for diagnostic capture with a connection to a SPAN.
- ST_port is the Cisco port type for Remote Strategic Position Analysis (RSPAN) monitoring in a source switch. This switch is an undedicated port that is used for RSPAN analysis, and it is not connected to any other device.

Figure 5-14 shows the Fibre Channel port types that are specific to Cisco.

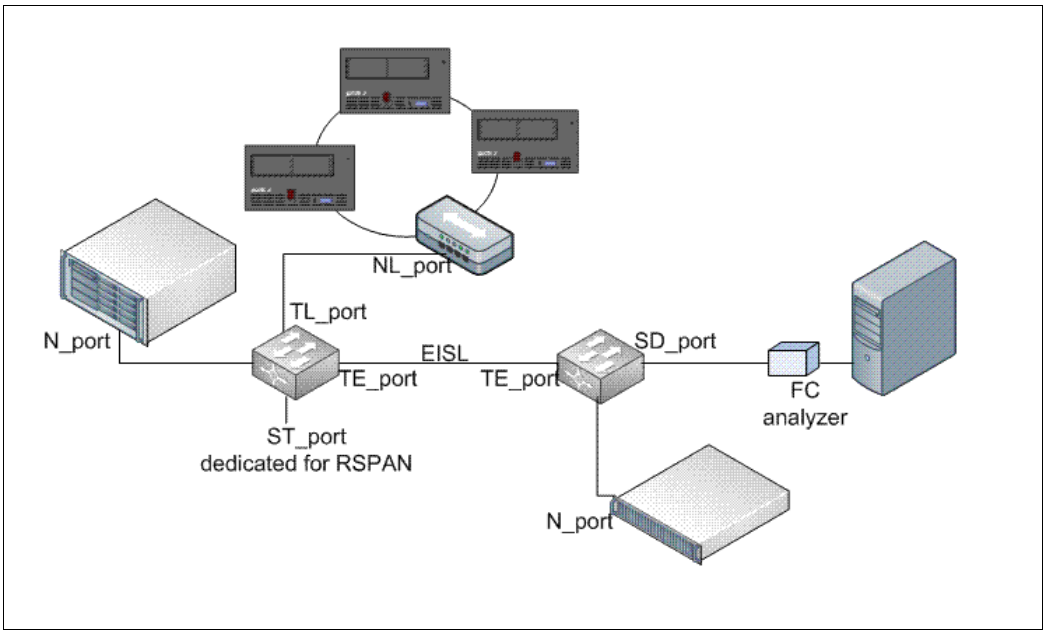


Figure 5-14 Cisco Fibre Channel port types

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, addressing is performed differently in an arbitrated loop and in a fabric.

5.3.1 Worldwide name

All Fibre Channel devices have a unique identity that is called a *worldwide name* (*WWN*). This identification is similar to the way that all Ethernet cards have a unique *Media Access Control* (*MAC*) address.

Each N_port has its own WWN, but it is also possible for a device with more than one Fibre Channel adapter to have its own WWN. Therefore, for example, a storage server can have its own WWN and incorporate the WWNs of the adapter within it. Therefore, a soft zone can be created by using the entire array, or individual zones can be created by using particular adapters. In the future, this ability will be available for the servers.

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

Each device in the SAN is identified by a unique WWN. The WWN contains a vendor identifier field, which is defined and maintained by the Institute of Electrical and Electronics Engineers (IEEE), and a vendor-specific information field.

Currently, two formats of the WWN are defined by the IEEE. The original format contains either a hex 10 or hex 20 in the first 2 bytes of the address. This address is then followed by the vendor-specific information.

Figure 5-15 shows both the old and new WWN formats.

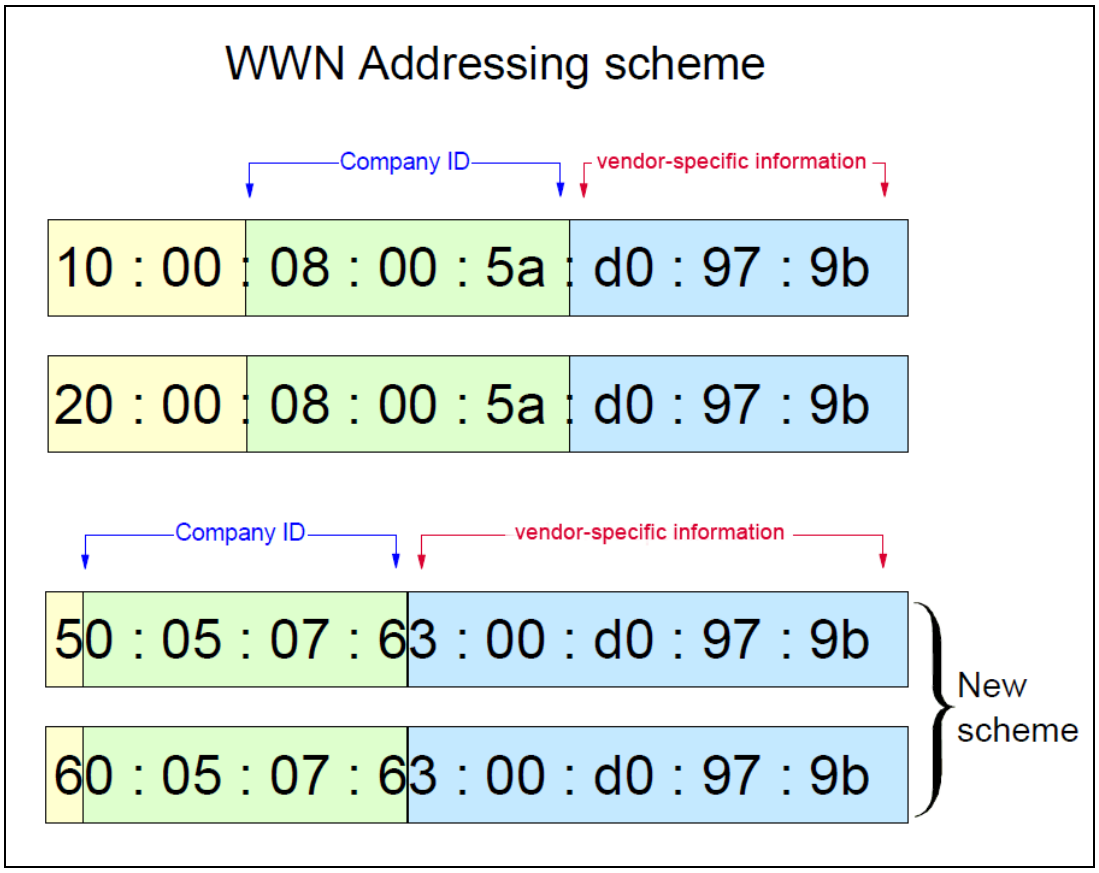


Figure 5-15 Worldwide name (WWN) addressing scheme

The new addressing scheme starts with a hex 5 or 6 in the first half-byte, which is followed by the vendor identifier in the next 3 bytes. The vendor-specific information is then contained in the following fields. Both of these formats are currently in use and they depend on the hardware manufacturer standards to follow either of the formats. However, the vendor ID and company ID are assigned uniquely by the IEEE standards, and each vendor and its identifier are in the following text file:

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

A *worldwide node name* (*WWNN*) is a globally unique 64-bit identifier that is assigned to each Fibre Channel *node* or *device*. For servers and hosts, the WWNN is unique for each *host bus adapter* (*HBA*). For a server with two HBAs, each HBA has a WWNN (two WWNNs total for the server). For a SAN switch, the WWNN is common for the chassis. For storage, the WWNN is common for each controller unit of midrange storage. And, in 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 a 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 worldwide node name and worldwide port name

Figure 5-16 shows a WWNN for each HBA. Every port in the HBA has an individual WWPN.

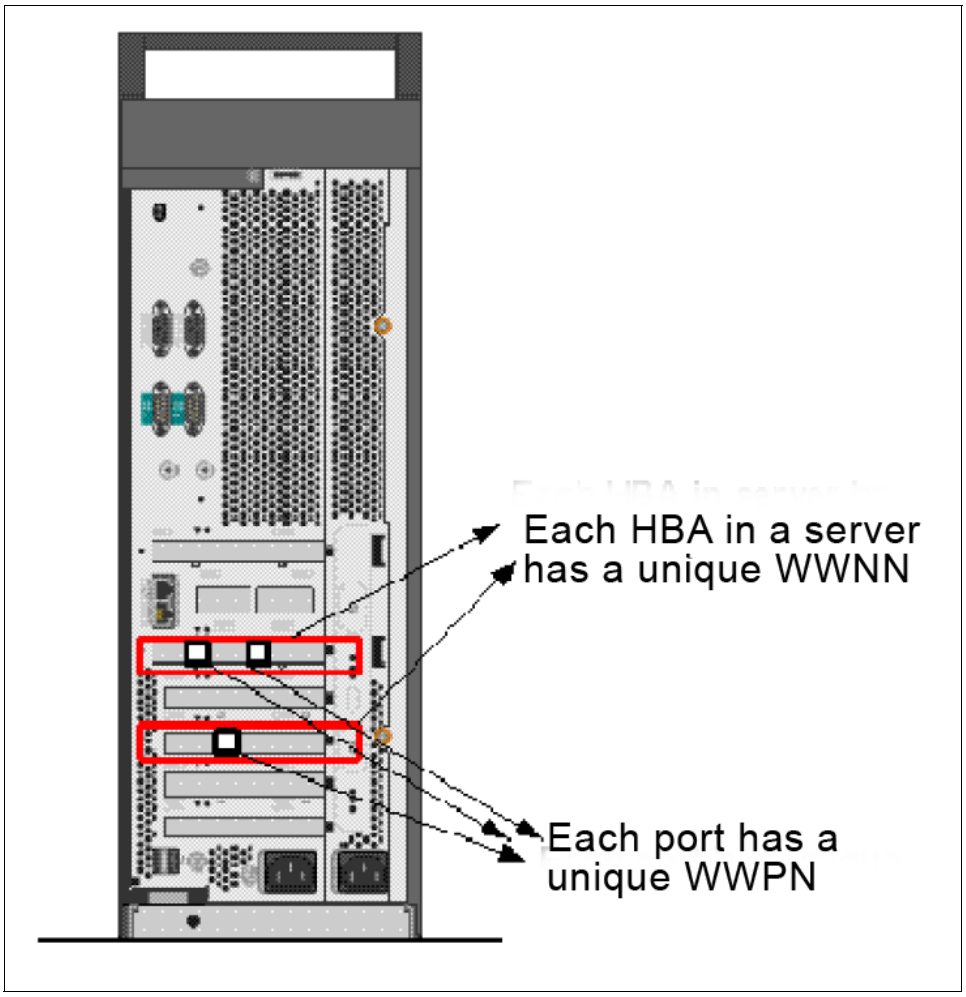


Figure 5-16 Server worldwide node name and worldwide port name

Storage area network worldwide node name and worldwide port name

Figure 5-17 shows that the WWNN is for the entire SAN switch chassis and the WWPN is for each FC port in the SAN switch chassis.

Fabric-assigned PWWNs: The new 16 Gbps b-type switches with Brocade Fabric OS (FOS) 7.0 can also have a virtual WWPN that is defined by switches that are called fabric-assigned PWWNs (FAPWWNs). These FAPWWNs can be used for pre-configuring zoning before the physical servers are connected. This feature helps to simplify and accelerate server deployment and improve operational efficiency by avoiding the wait time for setting up physical connectivity. This feature also requires that servers use Brocade HBAs/adapters with an HBA driver version 3.0.0.0 or higher, which can be configured to use FAPWWN.

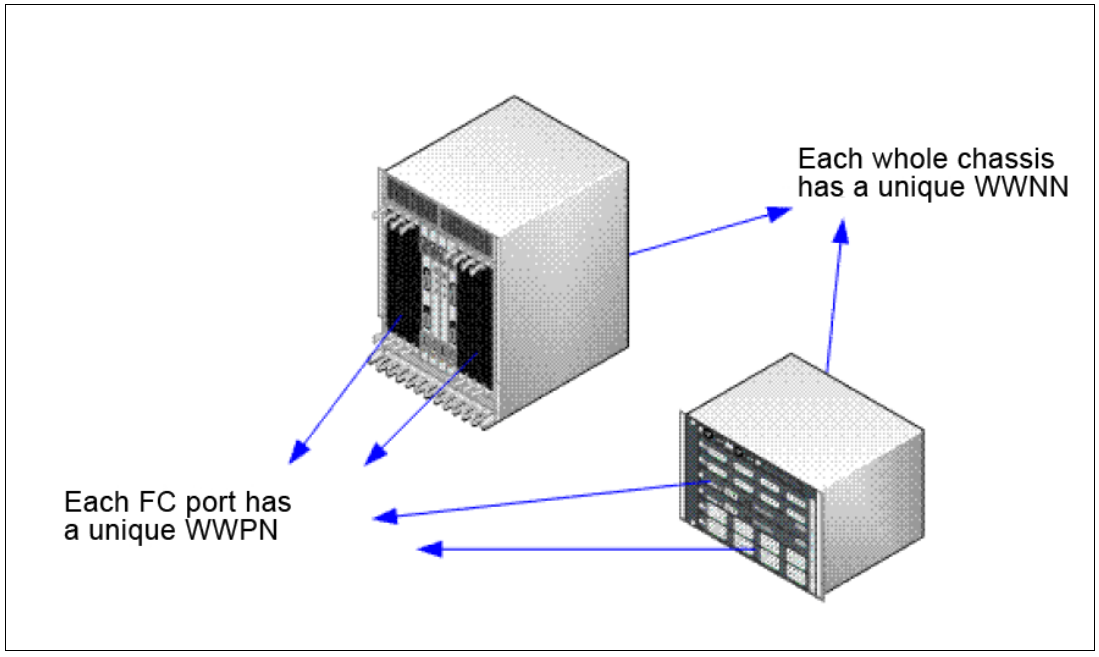


Figure 5-17 SAN switch worldwide node name and worldwide port name

Storage worldwide node name and worldwide port name

Disk storage has an individual WWNN for the entire storage system, and the individual FC host port have unique WWPNs (Figure 5-18). The diagram shows a dual-controller module.

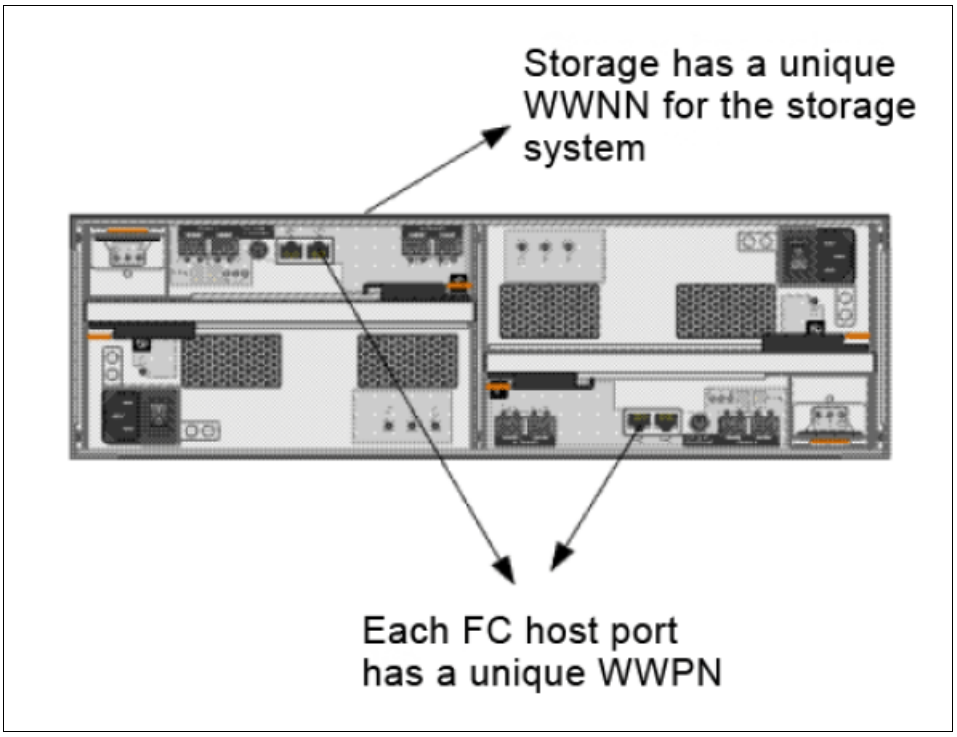


Figure 5-18 Storage worldwide node name and worldwide port name

Worldwide node name (WWNN): The IBM virtualization storage systems use WWNN differently. For example, each node in an IBM SAN Volume Controller or the IBM Storwize® V7000 has an individual and unique WWNN.

For the IBM DS8000®, each Storage Facility Image has a unique individual WWNN.

5.3.2 Tape device worldwide node name and worldwide port name

For tape devices, each drive inside the tape library has an individual WWPN and WWNN. Figure 5-19 shows that multiple drive libraries have an individual WWNN and WWPN for each drive.

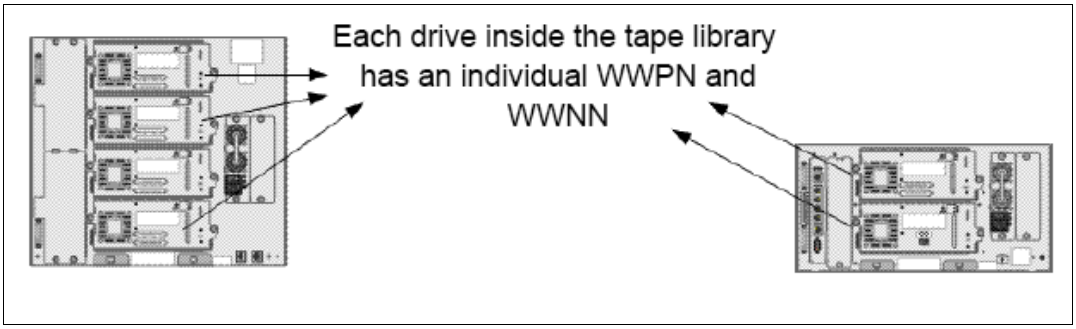


Figure 5-19 Tape device worldwide node name and worldwide port name

5.3.3 Port address

Because of the potential effect on routing performance by using 64-bit addressing, another addressing scheme is 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, a smaller frame header exists, and this configuration can speed up the routing process. With this frame header and routing logic, the Fibre Channel is optimized for high-speed frame switching.

With a 24-bit addressing scheme, this configuration allows for up to 16 million addresses. This address space is larger than any practical SAN design that is available today. A relationship must exist between this 24-bit address and the 64-bit address that is associated with the WWNs. We explain this relationship in the following section.

5.3.4 The 24-bit port address

The 24-bit address scheme removes the performance overhead of the manual administration of addresses by allowing the topology itself to assign addresses. This configuration 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 assigns the 24-bit addresses, another device must be responsible for maintaining the addressing scheme from WWN addressing to port addressing.

In the switched fabric environment, the switch is responsible for assigning and maintaining the port addresses. When a device with a WWN logs in to the switch on a specific port, the switch assigns the port address to that port. The switch also maintains 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 that runs inside the switch. The name server is essentially a database of objects in which a 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 the following parts:

- Domain (bits 23 - 16)
- Area (bits 15 - 08)
- Port or arbitrated loop physical address: AL_PA (bits 07 - 00)

Figure 5-20 shows how the address is built.

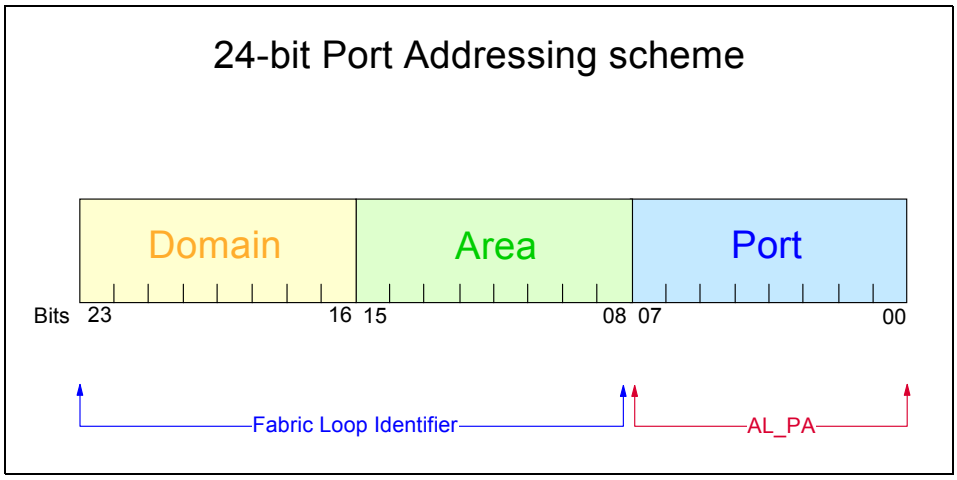


Figure 5-20 Fabric port address

The following functions provide the significance of several of the bits that make up the port address:

- Domain

  The most significant byte of the port address is the *domain*. This byte is the address of the switch. 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 Fibre Channel connection (FICON). Each manufacturer has 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 many of these addresses are reserved, such as the address for broadcast, only 239 addresses are available. This number 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. Therefore, to have more than 256 ports in one switch in a director class of switches, you must follow the shared area addressing.

- Port

  The final part of the address provides 256 addresses for identifying attached N_ports and NL_ports.

A simple calculation is used to arrive at the number of available addresses:

```
Domain x area x ports
```

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

Depending on the fabric topology, the fabric addressing format of the 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. This switch is connected to port 03 and the ALPA address is 00, which indicates that this device is not a loop fabric device. That is, it is a switched fabric device. For any switched fabric device, the ALPA ID is always *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 zeros (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 2 bytes are assigned a positive value by the switch. We call this mode a *public loop*.

Because 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 2 bytes represents the loop identifier, and this ID is common to all NL_ports on the same loop that logged in 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 the initialization of the loop and might, in a fabric-capable loop device, be modified by the switch during login.

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

5.3.6 The b-type addressing modes

IBM b-type (the IBM original equipment manufacturer (OEM) agreement with Brocade is referred to as *b-type*) has three addressing modes: native mode, core PID mode, and shared area addressing mode.

Native mode is used in traditional switches that support a maximum of 16 ports. This number is used because in native mode, the fabric addressing format that is used is *DD1A00*. The area part of the fabric address always has a prefix of 1. Therefore, it supports a port count from hexadecimal 10 to 1F (a maximum of 16 ports).

Core PID mode is used to support a maximum of 256 ports for each domain or switch. This number is used because in core PID mode, the area part of the fabric address supports addresses from hexadecimal 00 to FF (a maximum of 256 ports). The fabric addressing format that is used for this mode is *DDAA00*.

Figure 5-21 shows the native and core PID modes with the example FC address of two devices.

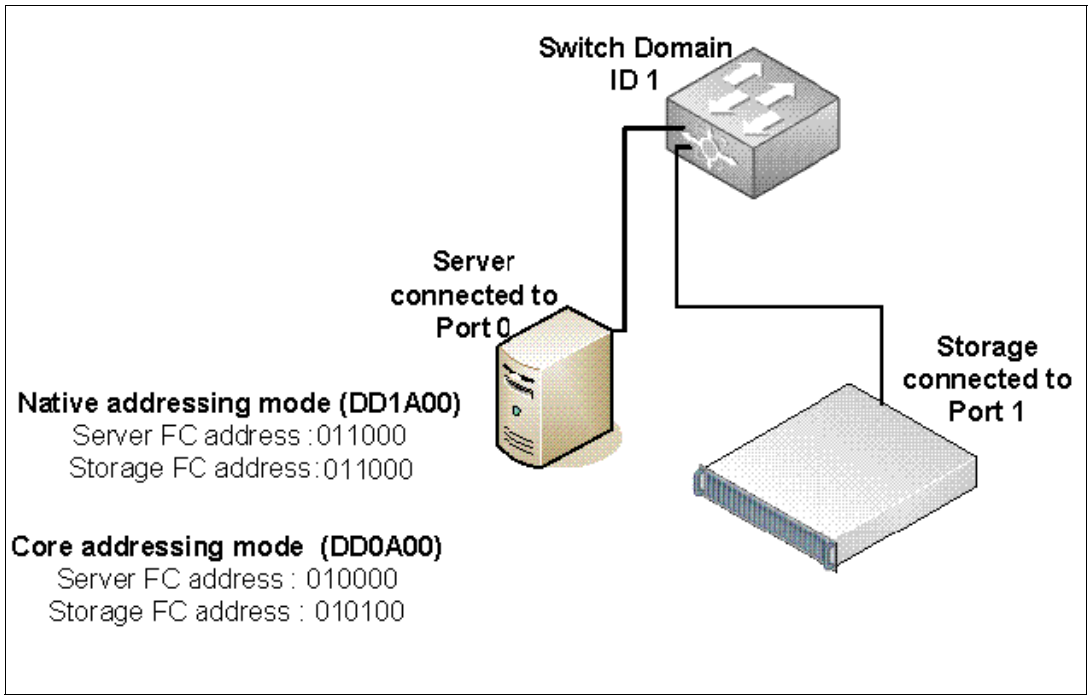


Figure 5-21 Native versus core addressing mode

Shared addressing mode is used when more than 256 ports are used in the same domain or switch. This mode is used in directors with high port density. The port addressing in these directors uses 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 using the area ID more than one time, this mode enables more than 256 ports to exist in a single domain.

Figure 5-22 shows port 24. Port 25 shares the area ID with port 32, and port 33 of the 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 control unit (CU) port, the FICON channel (by using FC-SB-2 and FC-FS protocol information) provides the address of its port, the source port address identifier (S_ID), and the address of the CU port. This CU port address is 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 determination depends on the node and N_port implementation. A server can determine the address of the N_port with which it wants to communicate in two ways:

- The *discovery method.* The address is determined by knowing the WWN of the target node N_port, and then by requesting a WWN for the N_port port address from a Fibre Channel Fabric Service. This service is called the *fabric name server*.
- The *defined method*. The address is determined 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 *port address definition.* It is the approach that is implemented for the FICON channel in the FICON native (FC) mode. This method is performed by using either the z/OS hardware configuration definition (HCD) function or an input/output configuration program (IOCP). These functions are used to define a 1-byte switch port, which is a 1-byte FC area field of the 3-byte Fibre Channel N_port port address.

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

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

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

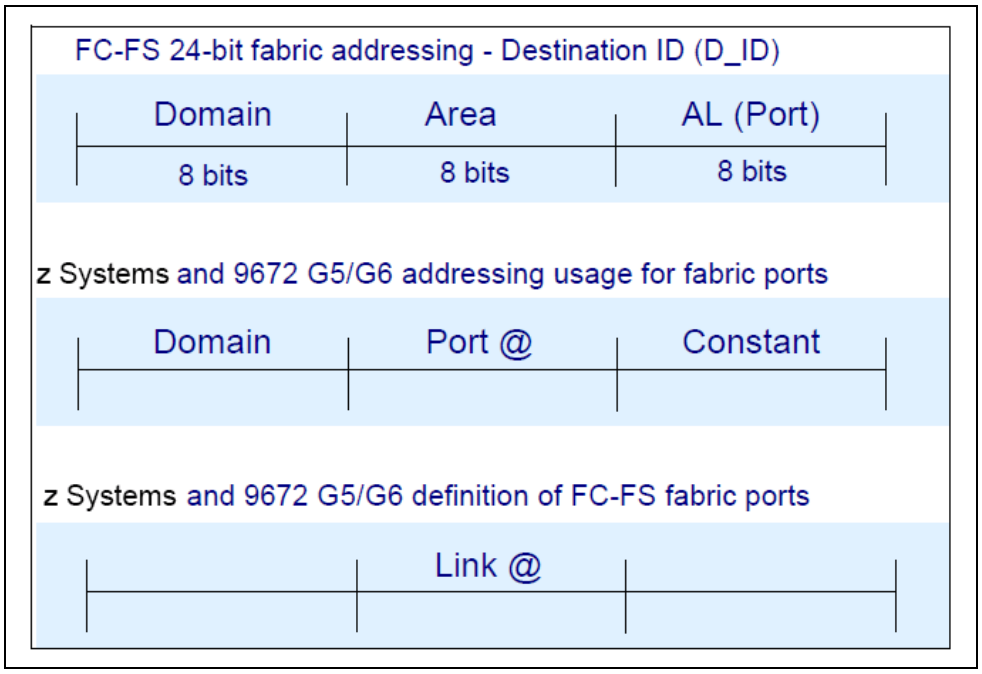


Figure 5-23 FICON port addressing

Figure 5-23 shows the FC-FS 24-bit port address and the definition of usage of that 24-bit address in an IBM z Systems server and 9672 G5/G6 environment. Only the 8 bits that make up the FC port address are defined for the z Systems server and 9672 G5/G6 to access a FICON CU.

The FICON channel in FICON native (FC) mode that works with a switched point-to-point FC topology (single switch) provides the other 2 bytes that make up the 3-byte FC port address of the CU to be accessed.

The z Systems and 9672 G5/G6 processors, when they work with a switched point-to-point topology, require that the *Domain* and the *AL_port* (*Arbitrated Loop*) field values are the same for all of the FC F_ports in the switch. Only the area field value differs for each switch F_port.

For the z Systems server and 9672 G5/G6, the *area* field is referred to as the *port address field* of the F_port. This field is only a 1-byte value. When the access to a CU that is attached to this port is defined, by using the z Systems HCD or IOCP, the port address is referred to as the *link address*.

The 8 bits for the domain address and the 8-bit constant field are provided from the *Fabric Login* initialization result (Figure 5-24). The 8 bits and the 1 byte for the port address (1-byte link address) are provided from the z Systems or 9672 G5/G6 CU link definition (by using HCD and IOCP).

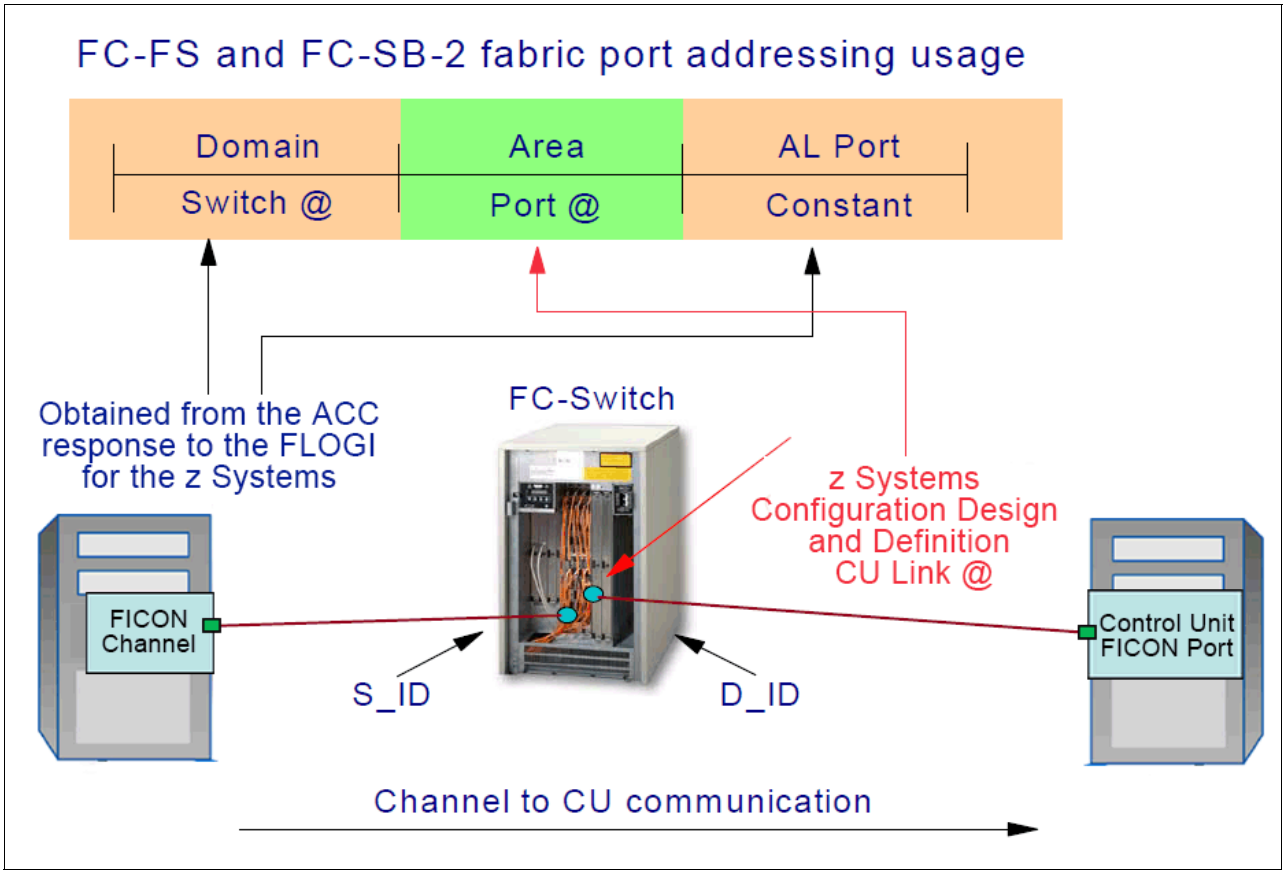


Figure 5-24 FICON single switch: Switched point-to-point link address

FICON address support for cascaded switches

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

For the FICON channel in FICON native (FC) mode, the Accept (ACC ELS) response to the Fabric Login (FLOGI) in a 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 also performs a fabric login to obtain its 24-bit FC port address.

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

- Domain
- Area
- AL (Port)

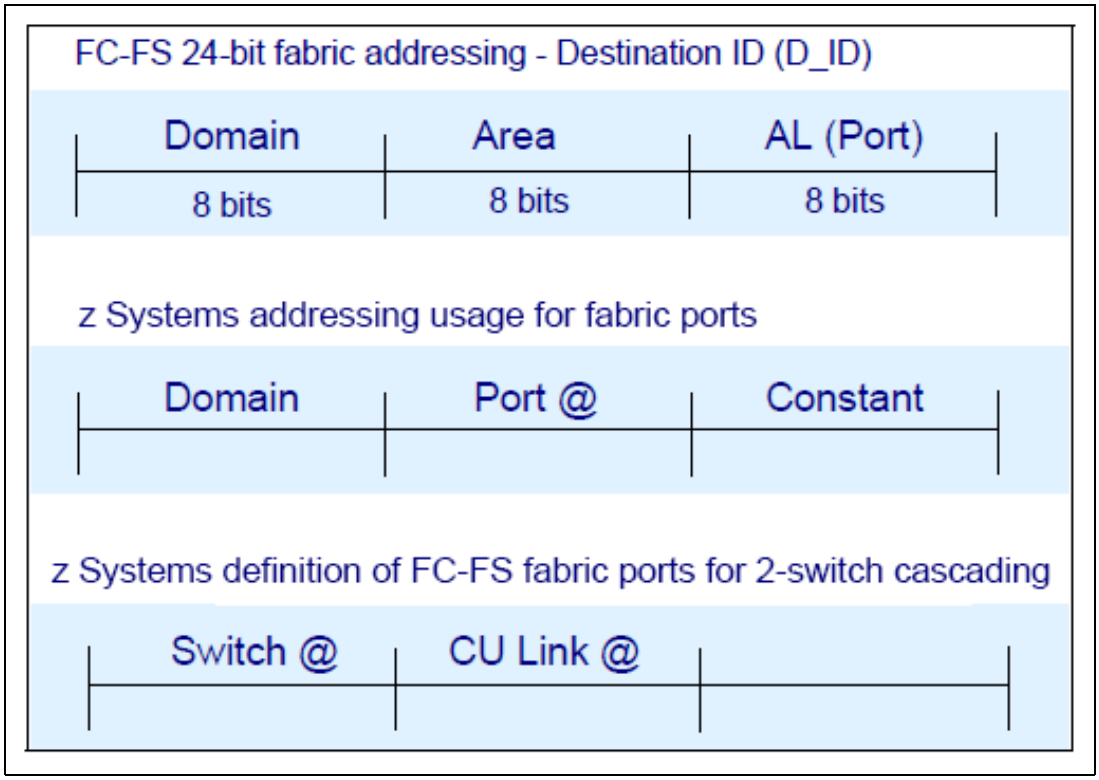


Figure 5-25 FICON addressing for cascaded directors

Figure 5-25 shows the FC-FS 24-bit port address and the definition usage of that 24-bit address in a z Systems environment. The 16 bits that make up the FC port address must be defined for the z Systems to access a FICON CU in a cascaded environment. The FICON channel in the FICON native (FC) mode that works with a two-switch cascaded FC topology provides the remaining byte that makes up the full 3-byte FC port address of the CU to be accessed.

It is required that the Domain, Switch @, AL_Port, and the Arbitrated Loop field values are the same for all of the FC F_ports in the switch. Only the Area field value differs for each switch F_port.

The z Systems Domain and Area fields are referred to as the *port address field* of the F_port. This field is a 2-byte value. When access is defined to a CU that is attached to this port (by using the z Systems HCD or IOCP), the port address is referred to as the *link address*.

The 8 bits for the constant field are provided from the Fabric Login initialization result (Figure 5-26). The 16 bits for the port address and 2-byte link address are provided from the z Systems CU link definition by using HCD and IOCP.

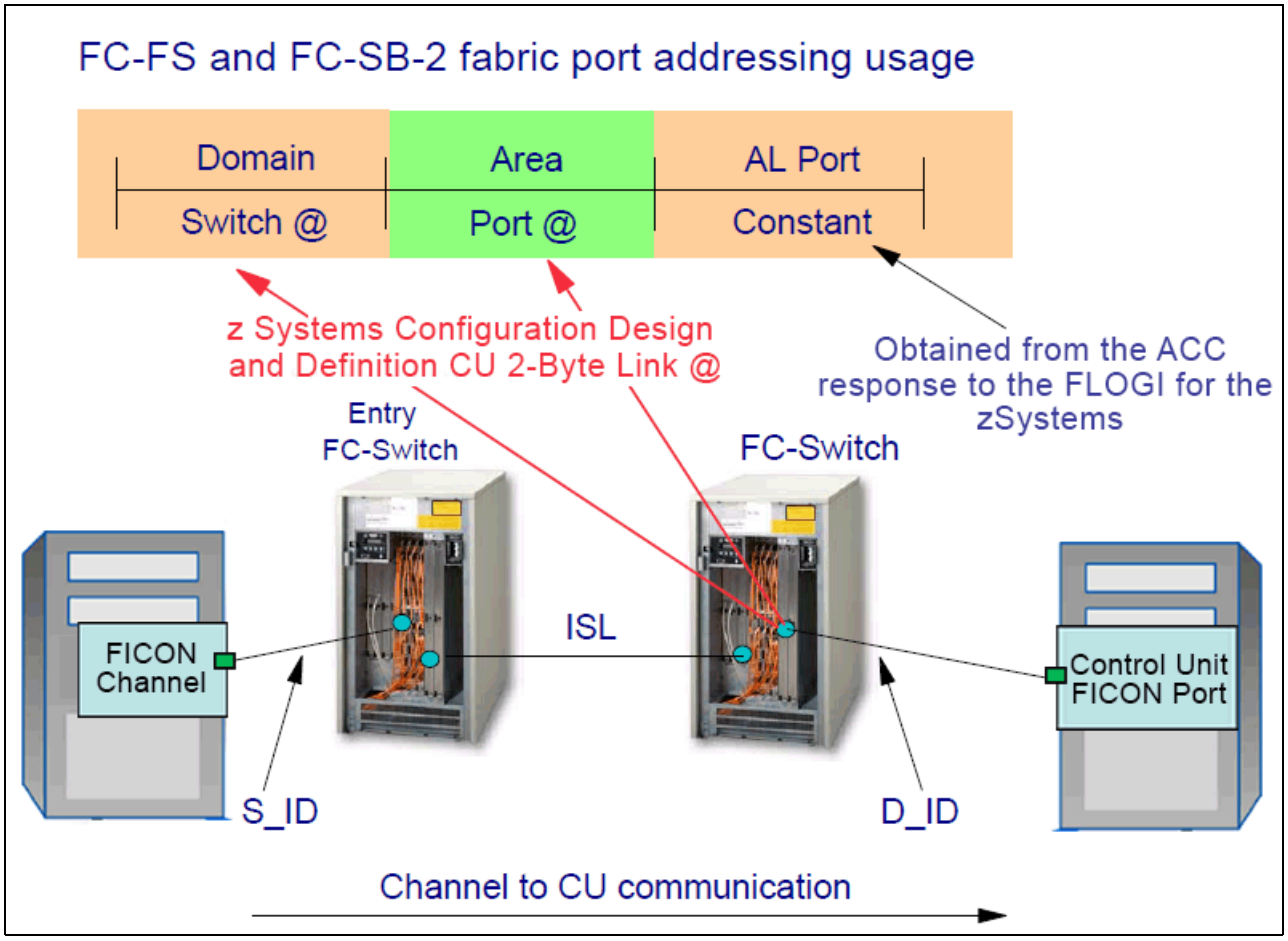


Figure 5-26 Two cascaded directors and FICON addressing

As a footnote, FCP connectivity is device-centric and defined in the fabric by 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 (that is, the *zoning information*). FICON devices do not query the name server for accessible devices because the allowable port and device relationships are defined on the host. Therefore, 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*), many loop-specific protocols are used. These protocols are used in the following ways:

- 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 complete their current use of the loop.
- Implement the access fairness mechanism to ensure that each port has an opportunity to access the loop.

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 performed. 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 *arbitrate primitive signal* (*ARB*) ordered set. If a higher-priority device arbitrates before the first NL_port sees its own ARB come by, it loses the arbitration. But, if it sees that its own ARB went all the way around the loop, it won arbitration. It can then open a communication to another NL_port. When it finishes, it can close the connection and either arbitrate for the loop or send one or more IDLEs. If it complies with the fairness algorithm, it takes the option of sending IDLEs. That action forces lower-priority NL_ports to successfully arbitrate for sending IDLEs, which allows lower-priority NL_ports to successfully arbitrate for the loop. However, no 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 zeros (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 the NL_port supports a fabric login, the upper 2 bytes are assigned a positive value by the switch. We call this mode a *public loop*.

Because 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. If a public loop assignment exists, the value of the upper 2 bytes represents the loop identifier. This identifier is common to all NL_ports on the same loop that logged in 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 might, in fabric-capable loop devices, be modified by the switch during login.

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

As a frame terminates with an *end-of-frame* (*EOF*) character, the current running disparity is forced to be *negative*. In the Fibre Channel standard, each transmission word between the end of one frame and the beginning of another frame also leaves 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 that left are assigned as AL_PAs. Stated another way, the 127 AL_PA limit is the maximum number, minus the reserved values, of neutral disparity addresses that can be assigned for use by the loop. This number does not imply that we recommend this amount, or load, but only that it is possible.

Arbitrated loop assigns priority to AL_PAs based on numeric value. The lower the numeric value, the higher the priority.

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

IBM z Systems: IBM System z9® and z Systems servers do not support the arbitrated loop topology.

5.5 Fibre Channel port initialization and fabric services

You learned that different port types exist. At a high level, port initialization starts with *port type detection*. Then, the *speed and active state detection* occurs 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 and FL_port that has an N_port or NL_port that is connected, as well as the Extended Link Service (ELS) and the Fibre Channel Common Transport (FCCT) protocol, are used for further *switch-port to node-port* communication. Data flow can happen only after this initialization completes. We now review the services that are responsible for the port initialization in a fabric switch.

The following three login types are available for fabric devices:

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

In addition to these login types, we also describe the roles of other fabric services, such as the fabric controller, management server, and time server.

5.5.1 Fabric login (FLOGI)

After the fabric-capable Fibre Channel device is attached to a fabric switch, it carries 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 sends 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 several of the service parameters that are requested by the N_port or NL_port are not supported, the switch sets the appropriate bits in the ACC frame to indicate this status.

NL_ports derive their AL_PA during the *loop initialization process* (*LIP*). The switch then decides whether to accept this AL_PA (if this AL_PA 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 shows nodes that are performing FLOGI.

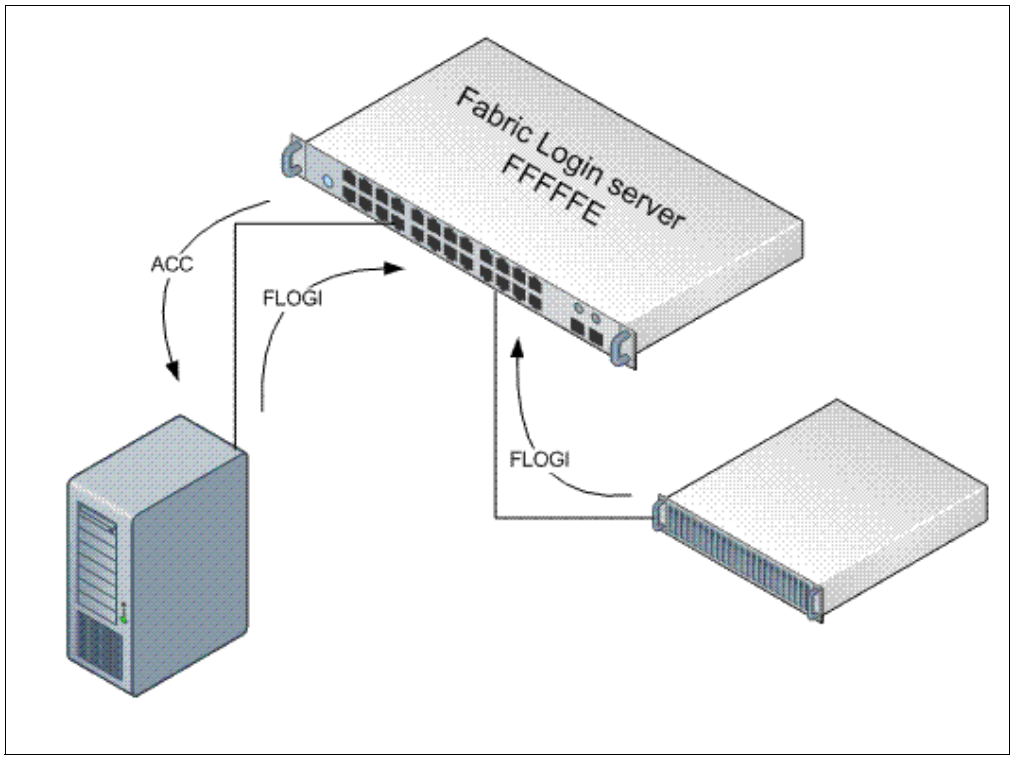


Figure 5-27 FLOGI of nodes

5.5.2 Port login (PLOGI)

Port login (*PLOGI)* is used to establish a session between two N_ports. PLOGI 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 the well-known address of $0xFFFFFC$. The device might register values for all or part of its objects, but the most useful values include the following objects:

- Twenty-four-bit port address
- Sixty-four-bit port name
- Sixty-four-bit node name
- Buffer-to-buffer credit capability
- Maximum frame size
- Class-of-service parameters
- FC-4 protocols that are supported
- Port type

When 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 shows the PLOGI of a host.

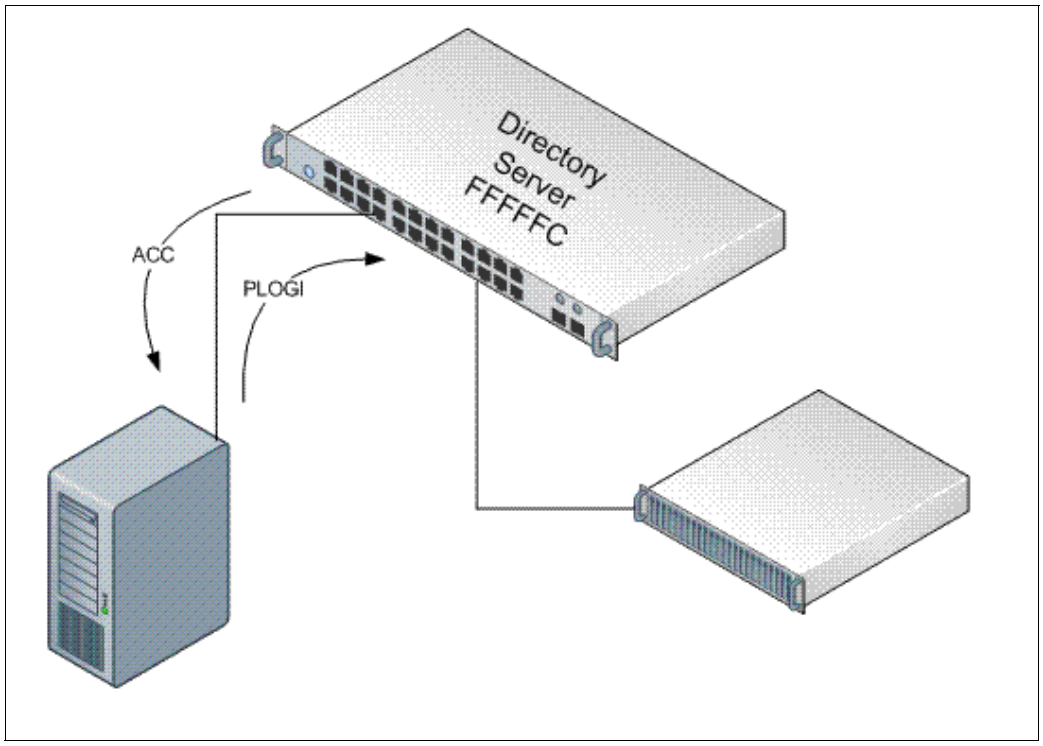


Figure 5-28 Node PLOGI to probe other nodes in the fabric

5.5.3 Process login (PRLI)

The *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 that are 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. However, PRLI might be required by a specific upper-level protocol, such as SCSI-FCP mapping.

Figure 5-29 shows the PRLI from server to storage.

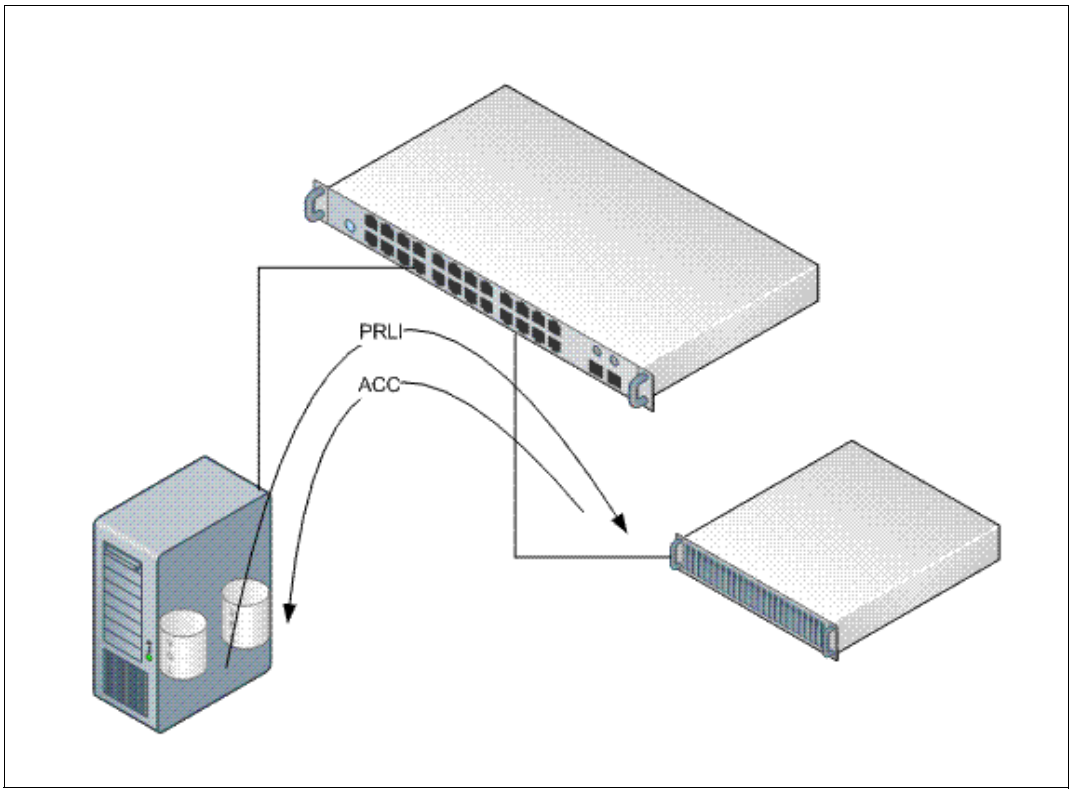


Figure 5-29 PRLI request from the initiator to the target

5.6 Fabric services

Fabric services are a set of services that are available to all devices that participate in a Fibre Channel fabric. Fabric services include the following functions:

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

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

All of these services are addressed by FC-2 frames, and they are accessed by *well-known addresses*.

5.6.1 Management server

Management server is an in-band fabric service that allows data to be passed from the device to management platforms. This data includes information, such as the topology of the SAN. A critical feature of this service is that it allows management software access to the *simple name server* (*SNS*), bypassing any potential block that is caused by zoning. Therefore, a management suite can have a view of the entire SAN. The well-known port that is 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. This entity represents a user that is accessing the time service.
- Time Server. This entity provides the time information through the time service.

More than one distributed time server instance can exist within the Fibre Channel network. However, from the user's perspective, the time service seems to come from the entity that is accessible at the time service well-known address identifier. If the time service is distributed, it is transparent to the application.

5.6.3 Simple name server

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

5.6.4 Fabric login server

To perform 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

The service, *registered state change notification* (*RSCN*), is critical because it propagates information about a change in the state of one node to all other nodes in the fabric. This communication means that if, for example, a node is shut down, that the other nodes on the SAN are informed and can take the necessary steps to stop communicating with the shutdown node. This notification prevents the other nodes from trying to communicate with the shutdown node that is timing out and trying again.

The nodes register to the *fabric controller* with a *state change registration* (*SCR*) frame. The fabric controller, which maintains the fabric state with all of the registered device details, alerts registered devices with an RSCN. This alert is sent whenever any device is added or removed, a zone changes, a switch IP changes, or a name changes. 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, even spanning a LAN or wide area network (WAN) connection. The challenge is to route the traffic with a minimum of performance overhead and latency and to prevent an out-of-order delivery of frames, while the system remains reliable. The following sections describe several of the mechanisms.

5.7.1 Spanning tree

If a failure occurs, it is important to consider an available alternative path between the source and destination so that data can still reach its destination. However, the availability of different paths might lead to the delivery of frames that are out of order. This order might happen because a frame takes a different path and arrives earlier than one of its predecessors.

A solution, which can be incorporated into the meshed fabric, is called a *spanning tree*. A spanning tree is an IEEE 802.1 standard. This concept means that switches stay on certain paths because the spanning tree protocol blocks 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. Therefore, 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 typically performed at the time of booting, and no configuration is needed. All paths are established at start time, and reconfiguration takes place only if no ISLs are broken or added.

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 that are used in FSPF were first proposed by Brocade, and they are incorporated into the FC-SW-2 standard. FSPF is used by most, if not all, manufacturers.

Fabric shortest path first

FSPF tracks the links on all switches in the fabric and associates a cost with each link. The cost is always calculated as 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 that are traversed by the path, and by choosing the path that minimizes the cost.

How fabric shortest path first 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. An initial database synchronization and an update mechanism exist. The initial database synchronization is used when a switch is initialized, or when an ISL comes up. The update mechanism is used when a link state changes. This mechanism ensures consistency among all switches in the fabric.

How fabric shortest path first helps

Where there are multiple routes, FSPF ensures that the route with the lowest number of hops is used. If all of the hops have the same latency, operate at the same speed, and have no congestion, FSPF ensures 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; all other attempts from outside are rejected.

For example, you might want to separate a Microsoft Windows environment from a UNIX environment because of the manner in which Windows attempts to claim all available storage for itself. Because not all storage devices can protect their resources from any host that seeks available resources, a preferred practice is to protect the environment in another manner. We show an example of zoning in Figure 5-30 where we separate AIX from Windows and create Zone 1 and Zone 2. This diagram also shows how a device can be in more than one zone.

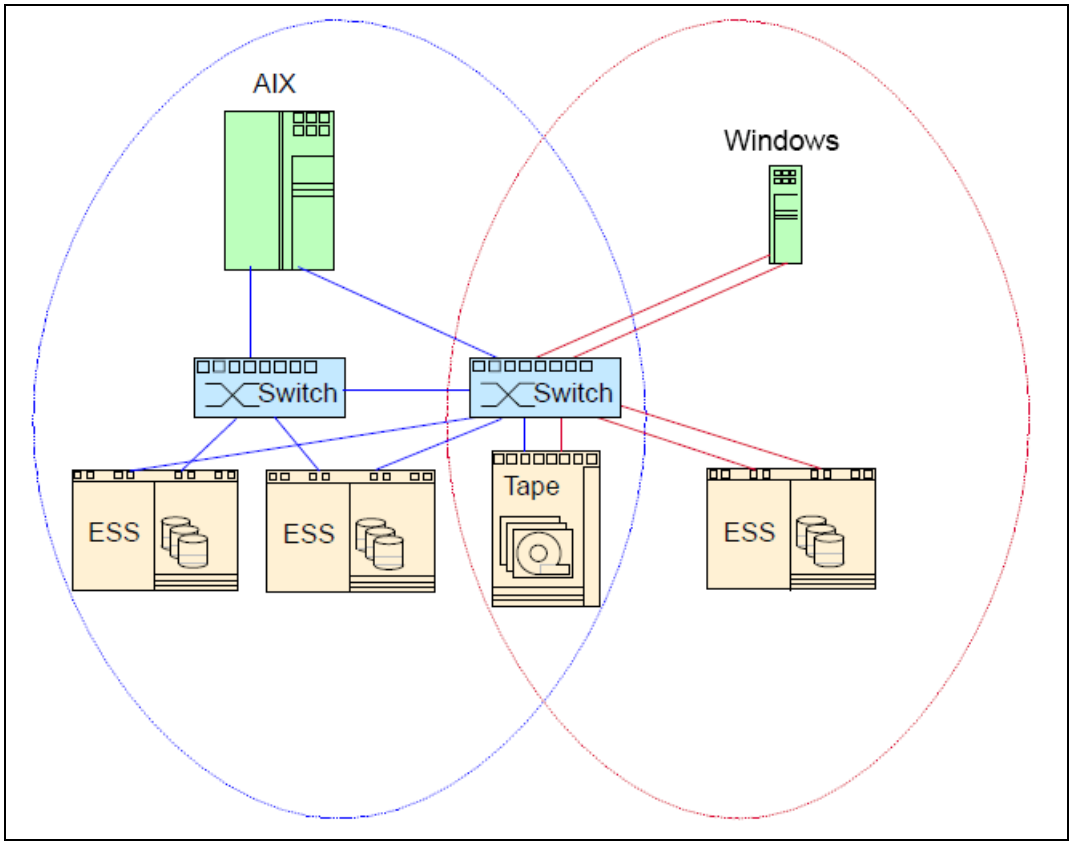


Figure 5-30 Zoning

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

Figure 5-31 shows an example of zoning:

- 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.

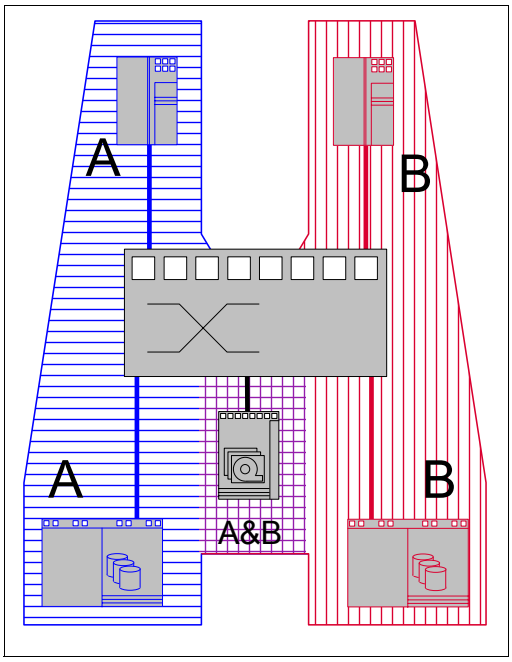


Figure 5-31 Zoning example

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

Zoning can be implemented in the following ways:

- Hardware zoning
- Software zoning

These forms of zoning are different, but they are not necessarily mutually exclusive. Depending on the particular manufacturer of the SAN hardware, hardware zones and software zones can overlap. Although this ability adds to the flexibility, it can complicate the solution, increasing the need for good management software and SAN documentation.

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. Hardware zoning can be implemented in the following configurations:

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

Figure 5-32 shows an example of zoning that is based on the switch port numbers.

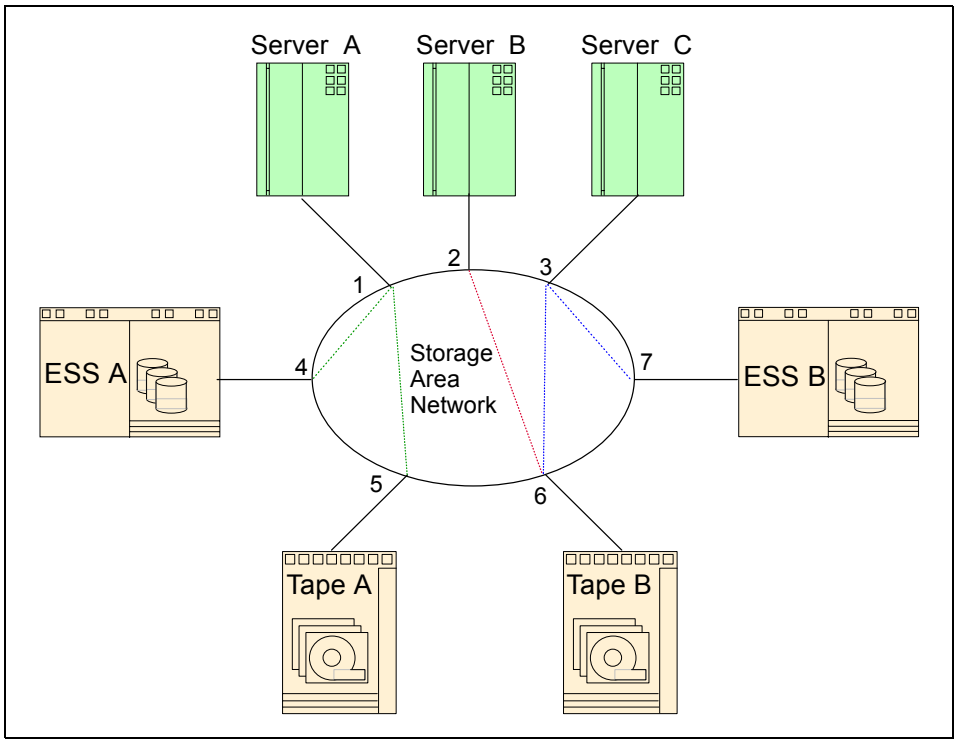


Figure 5-32 Zoning that is based on the switch port number

In Figure 5-32, zoning is based on the switch port number:

- Server A is restricted to see only storage devices that are zoned to port 1: ports 4 and 5.
- Server B is also zoned so that it can see only 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.

Figure 5-33 shows an example of hardware zoning. This example illustrates another way to consider the hardware zoning as an array of connections.

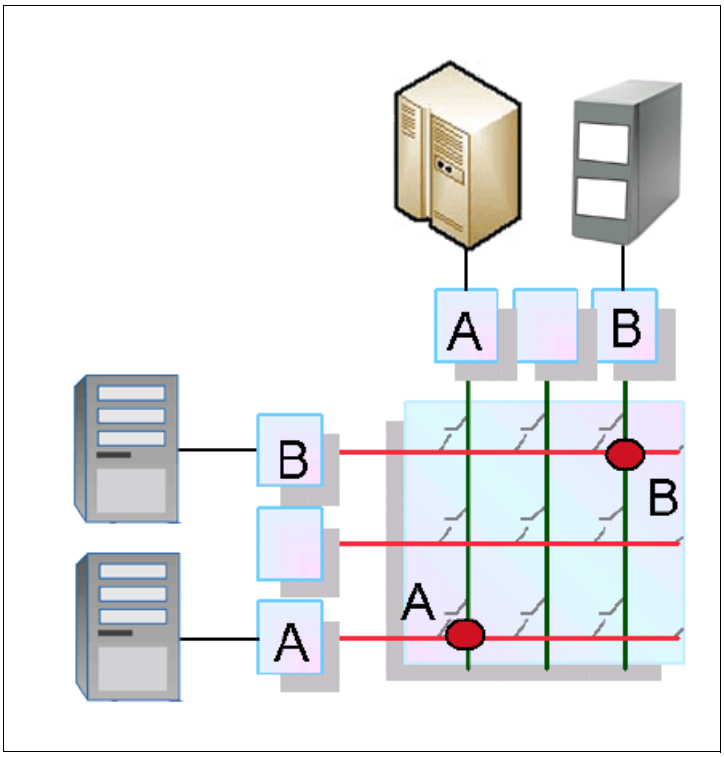


Figure 5-33 Hardware zoning

In Figure 5-33, device A can access only storage device A through connection A. Device B can access only storage device B through connection B.

In a hardware-enforced zone, *switch hardware*, usually at the *application-specific integrated circuit* (*ASIC*) level, ensures that no data is transferred between unauthorized zone members. However, devices can transfer data between ports within the same zone. Therefore, hard zoning provides the highest level of security. The availability of hardware-enforced zoning and the methods to create hardware-enforced zones depend on the switch hardware.

One disadvantage of hardware zoning is that devices must be connected to a specific port, and the whole zoning configuration can 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 simplify your configuration.

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

If possible, the designer can include several unused ports in a hardware zone. Therefore, if a particular port fails, perhaps because of a gigabit interface converter (GBIC) or transceiver problem, the cable can be moved to a different port in the same zone. Therefore, the zone does not need to be reconfigured.

5.8.2 Software zoning

Software zoning is implemented by the fabric operating systems within the fabric switches. Software zoning is 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 replies only with information about the ports in the same zone as the requesting port. A *soft zone*, or *software zone*, is not enforced by hardware. Therefore, if a frame is incorrectly delivered (addressed) to a port that it was not intended for, the frame is delivered to that port. This type of zoning is in contrast to hard zones.

When you use software zoning, the members of the zone can be defined by using their WWNs:

- Node WWN
- Port WWN

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

The number of possible members 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 that you want, without needing to power down the switch.

With software zoning, you do not 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 remains in the same zoning definition because the WWN of the device remains the same. The zone follows the WWN.

Important: Do not worry about your physical connections to the switch when you use software zoning. However, this statement does not mean that if you unplug a device, such as a disk subsystem, and you plug it into another switch port, that your host is still able to communicate with your disks. That is, you cannot assume that your host is still able to communicate until you either reboot or unload, and load your operating system device definitions, even if the device remains a member of that particular zone. The connection depends on the components that you use in your environment, such as the operating system and multipath software.

Figure 5-34 shows 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 that were shown in Figure 5-32 on page 124 for port zoning:

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

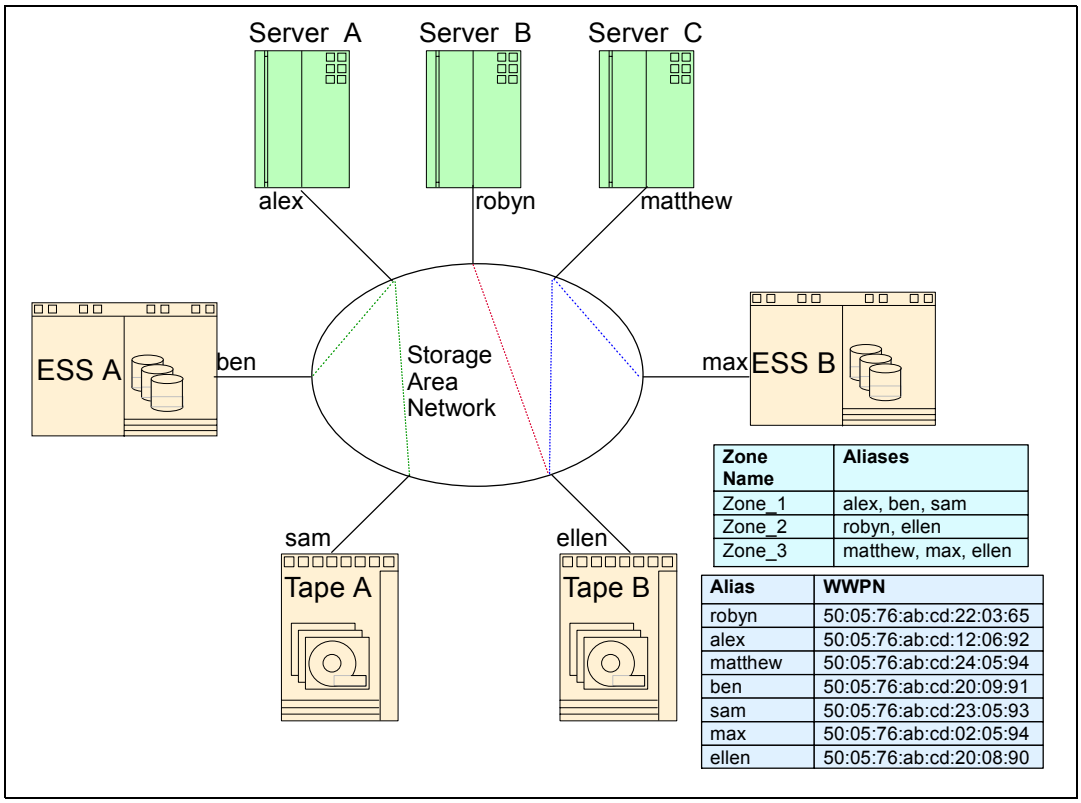


Figure 5-34 Zoning that is based on the WWNs of the devices

The following security leaks are possible with software zoning:

- When a specific host logs in to the fabric and asks for available storage devices, the simple name server (SNS) looks in the software zoning table to see the allowable devices. The host sees only the storage devices that are defined in the software zoning table. But, the host can connect directly to the storage device, by using device discovery, without asking SNS for the information.
- A device can define the WWN that it uses, rather than using the WWN that is designated by the manufacturer of the HBA. This concept is known as *WWN spoofing*. An unknown server can masquerade as a trusted server and therefore gain access to data on a particular storage device. Certain 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, nothing stops a person from dialing that number, whether by design or accident. The same is true for the WWN. Certain devices randomly probe for WWNs to see whether they can start a conversation with them.

Many 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.

Software zoning: 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 you use software-enforced zones, the switch does not control data transfer, and no guarantee exists for data that is 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. This service can also enable the partitioning of resources for management and access control. *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, WWN zoning, device-level zoning, protocol-level zoning, and *logical unit number* (*LUN*)-level zoning. Frame filtering is commonly performed by an application-specific integrated circuit (ASIC). Use this configuration, after the filter is set up, to achieve the complicated function of zoning and filtering at wire speed.

5.8.3 Logical unit number masking

The term logical unit number (LUN) was originally used to represent the entity within a SCSI target that runs I/Os. A single SCSI device typically has only a single LUN, but certain devices, such as tape libraries, might have more than one LUN.

With storage arrays, the array makes virtual disks available to the servers. These virtual disks are identified by LUNs.

More than one host can see the same storage device or LUN. This capability is potentially a problem, both from a practical and a security perspective. Another approach to securing storage devices from hosts that want to take over already assigned resources is *LUN masking*. Every storage device offers its resources to the hosts with LUNs.

For example, each partition in the storage server has its own LUN. If the host server wants to access the storage, it must 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 the hosts can access specific LUN with the storage device control program. Whenever the host accesses a particular LUN, the storage device checks its access list for that LUN. And, the storage device allows or prevents access to the LUN.

6

Storage area network as a service for cloud computing

Although 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 in the data center.

Cloud computing offers a new way to achieve solutions with significant cost savings and high reliability.

6.1 The cloud

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (for example: networks, servers, storage, applications, and services). These resources can be rapidly provisioned and released with minimal management effort or service provider interaction. Figure 6-1 shows an overview of cloud computing.

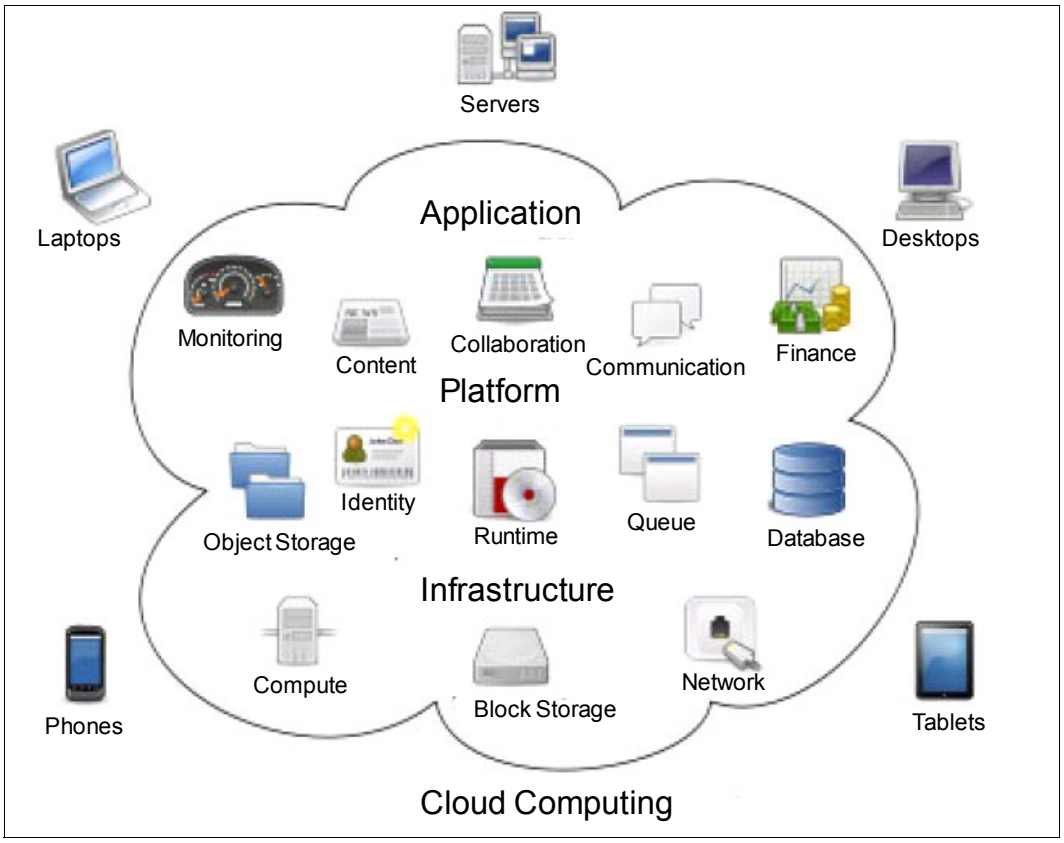


Figure 6-1 Cloud computing overview

Cloud computing provides computation, software, data access, and storage services that do not require 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 users use power without needing to understand the component devices or the infrastructure that is 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 savings, service reliability, and infrastructure flexibility.

To cater to the increasing, on-demand needs of business, IT services and infrastructures are moving rapidly toward 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 use Internet offerings and services by using this self-service, on-demand model.

Cloud computing can potentially affect your business dramatically 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 user IT support costs

From a technical perspective, cloud computing enables these capabilities, among others:

- 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*. Whether 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 (Figure 6-2).

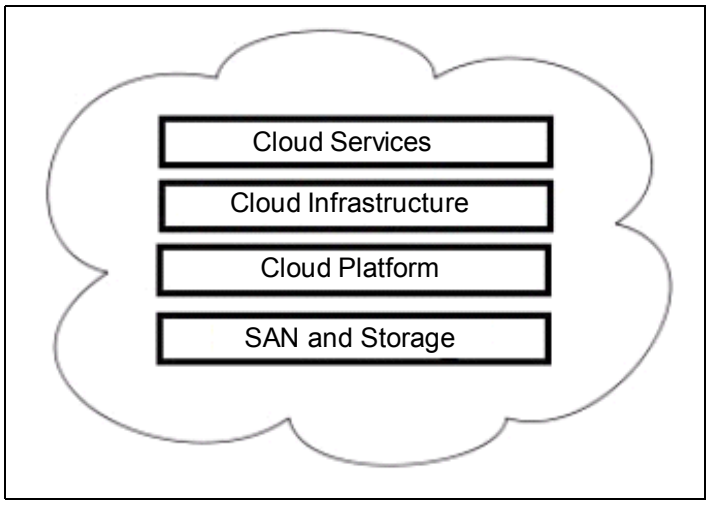


Figure 6-2 Cloud computing components

6.1.2 Cloud computing components

We describe the cloud computing components, or layers, in our model.

Cloud Services

This layer is the service that is delivered to the client. It can be an application, a desktop, a server, or disk storage space. The client does not need to know where or how their service is running; they just use it.

Cloud Infrastructure

This layer can be difficult to visualize depending on the final delivered service. If the final service is a chat application, the cloud infrastructure is the servers on which the chat application is running. In the other case, if the final service is a virtualized server, the cloud infrastructure is all of the other servers that are required to provide “a server” as a service to the client. Examples of these types of servers include a domain name server (DNS), security services, and management.

Cloud Platform

This layer consists of the selected platform to build the cloud. Many vendors exist, such as IBM Smart Business Storage Cloud, VMware vSphere, Microsoft Hyper V, and Citrix Xen Server, which are well-known cloud solutions in the market.

SAN and storage

This layer is where information flows and lives. Without it, nothing can happen. Depending on the cloud design, the storage can be any of the previously presented solutions, for example:

- Direct-attached storage (DAS)
- Network-attached storage (NAS)
- Internet Small Computer System Interface (iSCSI)
- Storage area network (SAN)
- Fibre Channel over Ethernet (FCoE)

For this book, we describe Fibre Channel or FCoE for networking and compatible storage devices.

6.1.3 Cloud computing models

Although cloud computing is still a relatively new technology, three cloud service models exist, each with a unique focus. The American National Institute of Standards and Technology (NIST) defined the following cloud service models:

- Software as a service (SaaS)

  This capability is provided to the consumer to use the applications that a provider runs 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 the network, servers, operating systems (OS), storage, or even individual application capabilities. One possible exception is for the consumer to continue to control limited user-specific application configuration settings.

- Platform as a service (PaaS)

  This capability is provided to the consumer to deploy consumer-created or acquired applications onto the cloud infrastructure. Examples of these types of applications include those applications that are created by using programming languages and tools that are supported by the provider. The consumer does not manage or control the underlying cloud infrastructure, including the network, servers, operating systems, or storage. But, the consumer controls the deployed applications and possibly application-hosting environment configurations.

- Infrastructure as a service (IaaS)

  This capability is provided to the consumer to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software. These resources can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure, but the consumer controls the operating systems, storage, and deployed applications. The consumer might also have limited control of select networking components (for example, hosts).

In addition, NIST also defined the following models for deploying cloud services:

- Private cloud

  The cloud infrastructure is owned or leased by a single organization and operated solely for that organization.

- Community cloud

  The cloud infrastructure is shared by several organizations and supports a specific community that shares, for example, mission, security requirements, policy, and compliance considerations.

- Public cloud

  The cloud infrastructure is owned by an organization that sells 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. However, these entities are bound together by standardized or proprietary technology that enables data and application portability, for example, cloud bursting.

Figure 6-3 shows cloud computing deployment models.

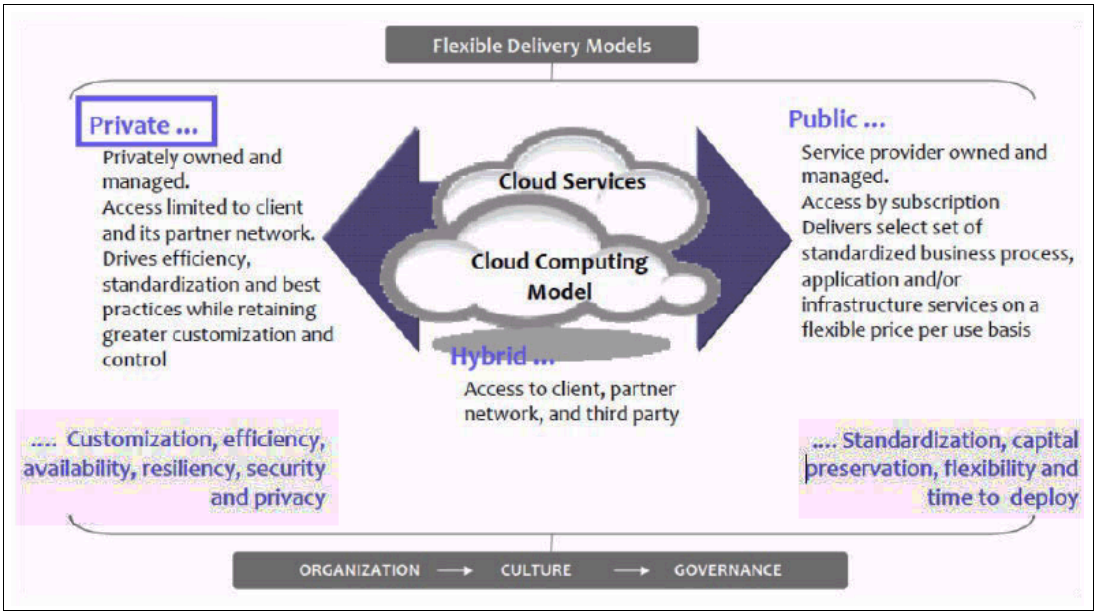


Figure 6-3 Cloud computing deployment models

From a storage perspective, IBM clients, based on their business requirements, can choose to adopt either a public or private storage cloud. The following definitions describe these types of storage clouds:

- Public storage cloud

  This type is designed for clients who do not want to own, manage, or maintain the storage environment, therefore 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 client level, and security measures to isolate client data. The public storage cloud allows for variable billing options and shared tenancy of the storage cloud, giving clients the flexibility to manage the use and growth of their storage needs. This type is the industry-standard view of a storage cloud offering and comparable to storage cloud offerings by other vendors.

- Private storage cloud

  With a private storage cloud, clients have the choice of technology and location on a dedicated infrastructure with fixed monthly charges and a physical capacity that is manageable by the client. Each application can use 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, see this website:

`http://www.ibm.com/cloud-computing/us/en/`

6.2 Virtualization and the cloud

When people talk about virtualization, they are typically 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 (VMs) can run different operating systems and multiple applications while they share the resources of a single physical computer. And, because each virtual machine is isolated from other virtualized machines, if one virtual machine crashes, it does not affect the other virtual machines.

Hypervisor software is the secret that makes virtualization possible. This software sits between the hardware and the operating system, and decouples 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 (I/O).

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 solution is *storage virtualization*. This type of virtualization is where multiple network storage resources are pooled into what is displayed as a single storage device for easier and more efficient management of these resources.

Other types of virtualization you might hear about include the following examples:

- *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 type of virtualization gives users a full client experience. IT staff can provision, manage, upgrade, and patch desktops virtually, instead of physically.

Virtualization was first introduced in the 1960s by IBM. It was designed to boost the utilization of large, expensive mainframe systems by partitioning them into logical, separate virtual machines to 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 type consists of virtualizing three key parts: servers, desktops, or applications. The virtualization concept that is used for servers and desktops is almost the same, but for applications, the concept is different.

Virtualizing servers and desktops basically takes physical computers and makes them virtual. To make virtualization possible, a cloud platform is required. We show the traditional physical environment in Figure 6-4. This model shows where one application maps to one operating system (OS), and one OS to one physical server, and one physical server to one storage system.

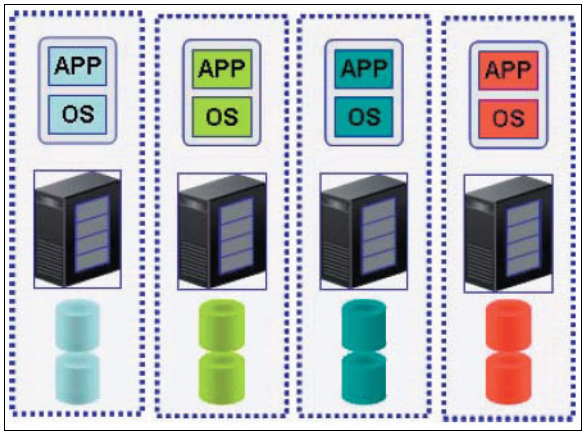


Figure 6-4 Traditional physical environment model

6.2.2 Cloud platforms

A platform places multiple virtual servers in a single physical computer. This platform is called the *hypervisor*. This platform is a layer in the computer stack between the virtual and physical components.

Four core concepts exist in virtualization:

- 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 that you manage other files.

- Isolation

  Virtual machines that run on a hardware platform cannot see or affect each other, so multiple applications can 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. This layer allows hardware from multiple vendors to run on the same physical resource, if the server is on the vendor's hardware compatibility list (or a similar list). For more information, see the specific vendor's website.

Figure 6-5 shows the virtualized environment.

Figure 6-5 Virtualized environment model

Server virtualization

Three popular approaches to server virtualization are available:

- Virtual machine model
- Paravirtual machine model
- Virtualization at the operating system layer

Virtual machines (VMs) are based on the host/guest design. Each guest runs on a virtual implementation of the hardware layer. This approach allows the guest operating system to run without modifications. This way, the administrator can create guests that use different operating systems. The guest has no knowledge of the host operating system because the guest is unaware that it is not running on real hardware. This model, however, requires real computing resources from the host so this model uses a hypervisor to coordinate instructions to the CPU.

The paravirtual machine (PVM) model is also based on the host/guest design, and it uses a virtual machine monitor (VMM). In the paravirtual machine model, however, the VMM actually modifies the code of the guest operating system. This modification is called *porting*. Porting supports the VMM so that the VMM can use privileged system calls sparingly. Paravirtual machines also can run multiple operating systems. Xen and Unified Modeling Language (UML) both use the paravirtual machine model.

Virtualization at the OS level works slightly differently. It is not based on the host/guest design. In the OS level model, the host runs a single OS kernel as its core and exports the 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, reducing CPU usage overhead. This model also requires each partition to remain strictly isolated from its neighbors so that a failure or security breach in one partition is unable to affect any of the other partitions. In this model, common binary files 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 virtual I/O (VIO) and Solaris Zones both use OS-level virtualization.

Desktop Virtualization

Desktop virtualization is sometimes referred to as *client virtualization.* Desktop virtualization 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 on the physical machine that is virtualized.

Desktop virtualization virtualizes desktop computers. These virtual desktop environments are “served” to users in the network. Users interact with a virtual desktop in the same way that they access and use a physical desktop. Another benefit of desktop virtualization is that you can log in remotely 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 are hosted on a data center server.

Also, options are available for using hosted virtual desktops, where the desktop virtualization services are provided to a business through a third party. The service provider provides the managed desktop configuration, security, and SAN.

Application Virtualization

Application virtualization is similar to desktop virtualization, where individual desktop sessions (OS and applications) are virtualized and run from a centralized server. However, *application virtualization* virtualizes the applications so that an application can either be run from a centralized server or streamed from a central server and run in an isolated environment on the desktop.

In the first type of application virtualization, the application image is loaded on to a central server. When a user requests the application, the application 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 because the application is isolated from other applications, conflicts are less likely. 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 the applications are run (executed) in the servers. Only the on-screen information that is required to be seen by the user is sent over the LAN. This type of application virtualization is closer to desktop virtualization, but only the application is virtualized instead of both the application and the operating system. The greatest advantage of this type of application virtualization is that it does not matter what the underlying OS is in the user's computer because the applications are processed in the server. Another advantage is the effectiveness of mobile devices, such as mobile phones and tablet computers, with less processing power when the user runs applications that require significant processing capabilities. These applications 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. Storage virtualization is a foundation for the implementation of other technologies, such as thin provisioning, tiering, and data protection, which are transparent to the server.

Storage virtualization offers several advantages:

- Improved physical resource usage: By consolidating and virtualizing storage systems, previously wasted "white" space can be used.
- Improved responsiveness and flexibility: By decoupling physical storage from virtual storage, you can reallocate resources dynamically, as required by the applications or storage subsystems.
- Lower total cost of ownership (TCO): Virtualized storage offers more capability with the same or less storage.

Several types of storage virtualization are available.

Block-level storage virtualization

Block-level storage virtualization 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 that are used by this type of storage virtualization.

Two types of block-level virtualization are available:

- Disk-level virtualization

  Disk-level virtualization is an abstraction process from a physical disk to a logical unit number (LUN) that is presented as 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). File-level storage virtualization is a file presentation in a single global namespace, regardless of the physical file location.

Tape virtualization

Tape virtualization refers to the virtualization of tapes and tape drives that use 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 available virtualization features in the IBM Storage portfolio. These features enable the SAN infrastructure to support the requirements of scalability and consolidation, combining them with a lower TCO and a higher return on investment (ROI):

- IBM b-type Virtual Fabrics
- CISCO Virtual SAN (VSAN)
- N_Port ID Virtualization (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 that enables the logical partitioning of SAN switches. When Virtual Fabric is enabled, a default logical switch that uses all of the ports is formed. This default logical switch can be then divided into multiple logical switches by grouping them together at a port level.

Figure 6-6 shows the flow of Virtual Fabric creation.

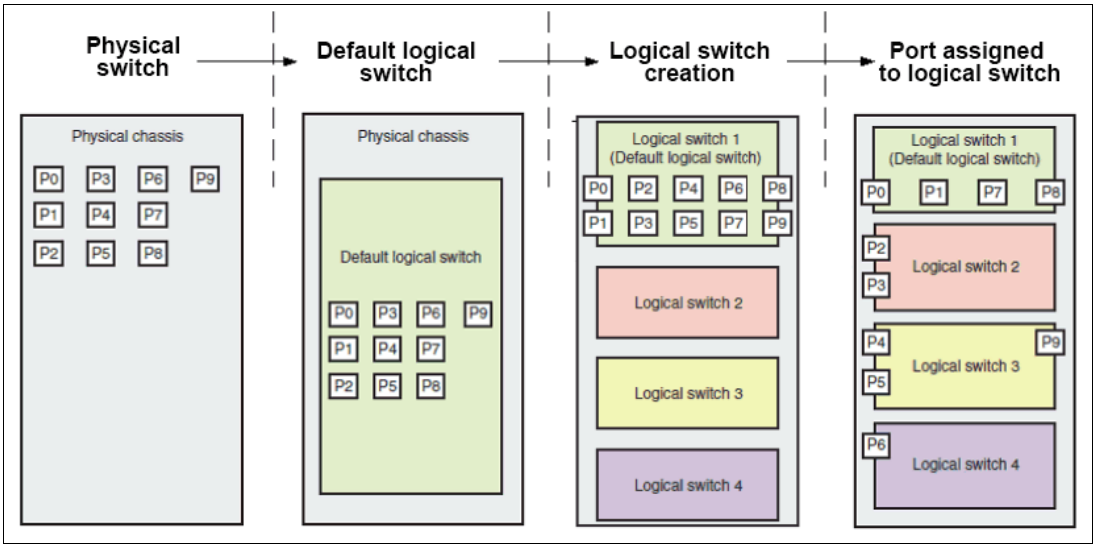


Figure 6-6 Virtual Fabric creation

Logical fabric

When the fabric is formed with at least one logical switch, the fabric is called a *logical fabric*. Two methods of fabric connectivity are available for logical fabrics:

- A logical fabric is connected with a dedicated inter-switch link (ISL) to another switch or a logical switch. Figure 6-7 shows a logical fabric that is formed between logical switches through a dedicated ISL for logical switches.

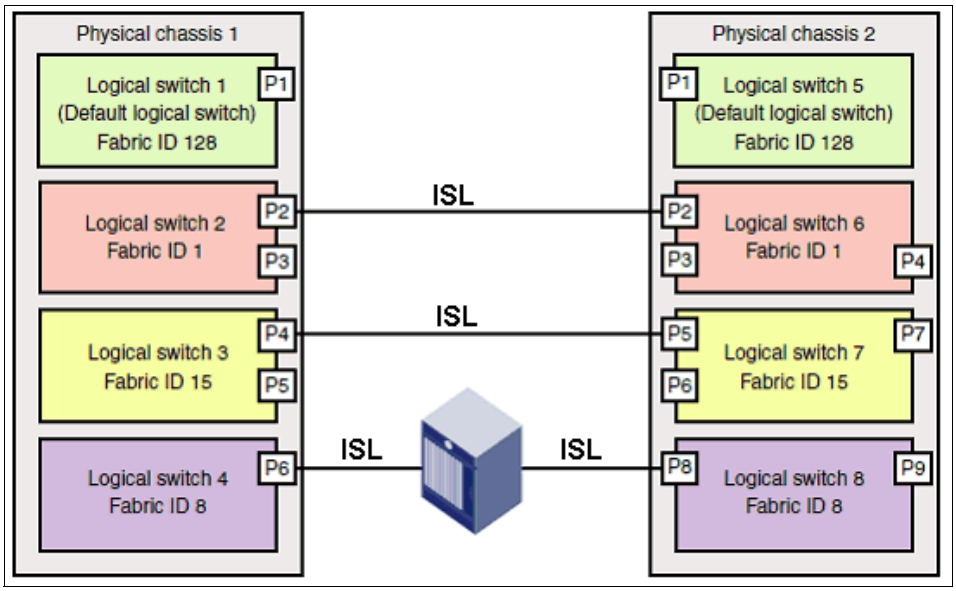


Figure 6-7 Logical fabrics with dedicated ISL

- Logical fabrics are connected by using a shared ISL, which is called an *extended ISL* (*XISL*), from a base logical switch. In this case, the separate logical switch is configured as a base switch. This separate logical switch is used only for XISL connectivity and not for device connectivity. Figure 6-8 shows a logical fabric that is formed through the XISL in the base switch.

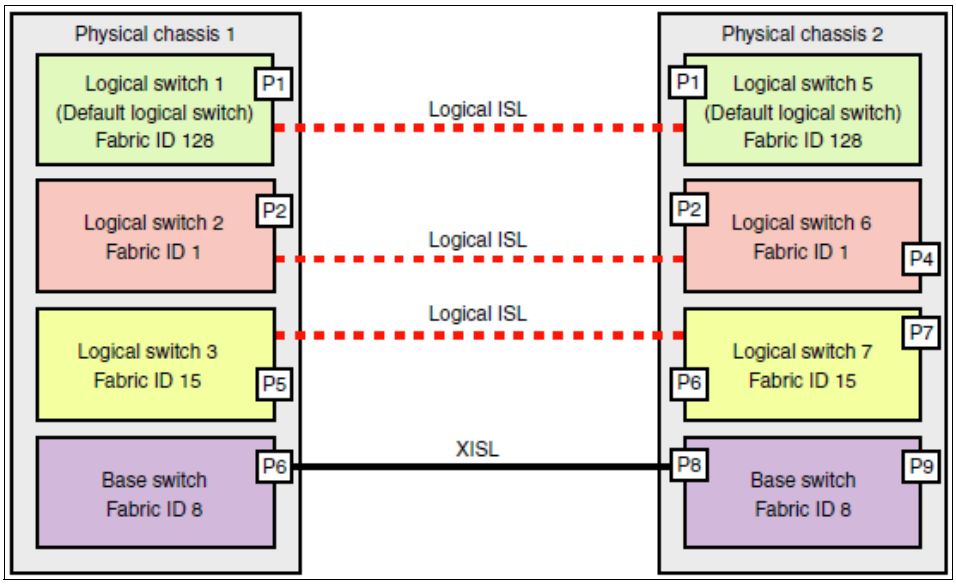


Figure 6-8 Logical ISL formed through the XISL in the base switch

6.3.2 Cisco virtual storage area network

Cisco *virtual storage area network* (*VSAN*) is a feature that enables the logical partition of SAN switches. A VSAN provides the flexibility to partition, for example, a dedicated VSAN for disk and tape. Or, a VSAN can provide the flexibility to maintain 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.

Virtual storage area network in a single storage area network switch

With VSAN, you can consolidate small fabrics into the same chassis. This consolidation can also enable more security by the logical separation of the chassis into two individual VSANs. Figure 6-9 shows a single chassis that is divided into two logical VSANs.

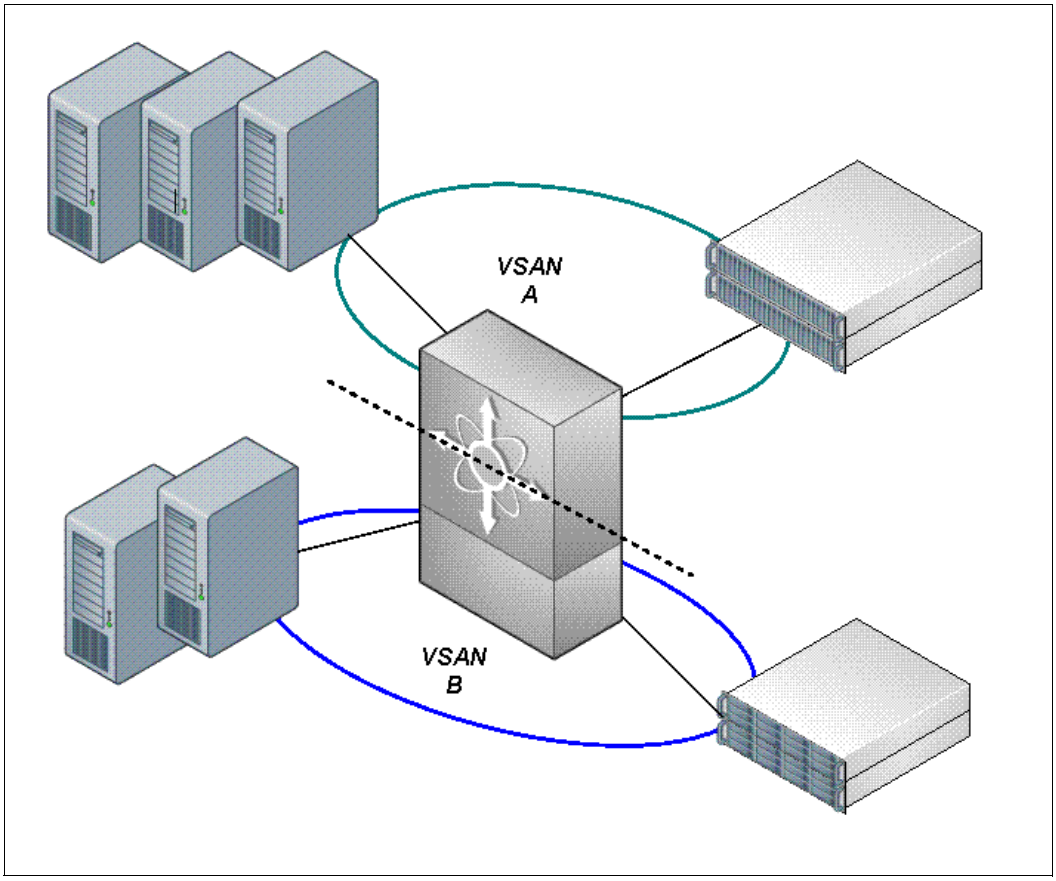


Figure 6-9 Two VSANs in a single chassis

Virtual storage area network across multiple chassis

In multiple chassis, the virtual storage area network (VSAN) can be formed with devices in one chassis to devices in another switch chassis through the *extended inter-switch link* (*XISL*).

Figure 6-10 shows the VSAN across chassis with an *enhanced inter-switch link* (*EISL*) for VSAN communication.

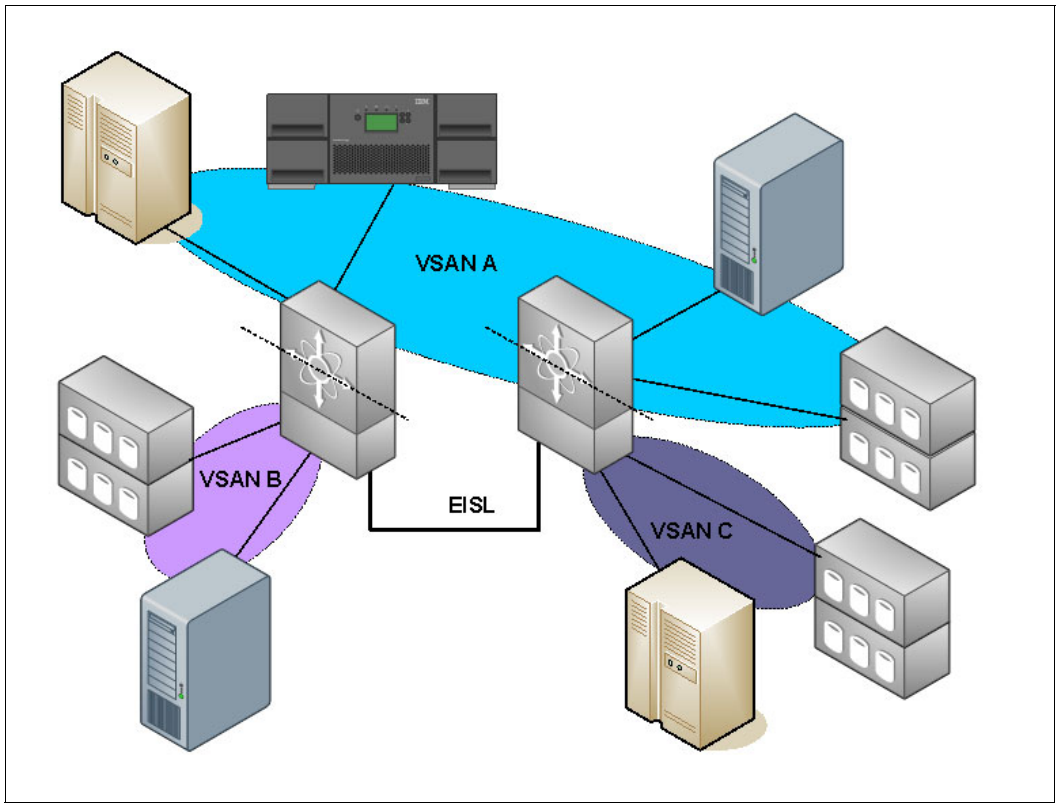


Figure 6-10 VSAN across multiple chassis

6.3.3 N_Port ID Virtualization

Server virtualization with blade servers provides enhanced scalability of servers. This scalability is supported equally in the SAN with *N_Port ID Virtualization* (*NPIV*). NPIV allows SAN switches to have one port that is shared by many virtual nodes, therefore, supporting a single HBA with many virtual nodes.

Figure 6-11 shows sharing a single HBA by multiple virtual nodes. In this case, the same HBA is defined with multiple virtual worldwide node names (WWNNs) and worldwide port names (WWPNs).

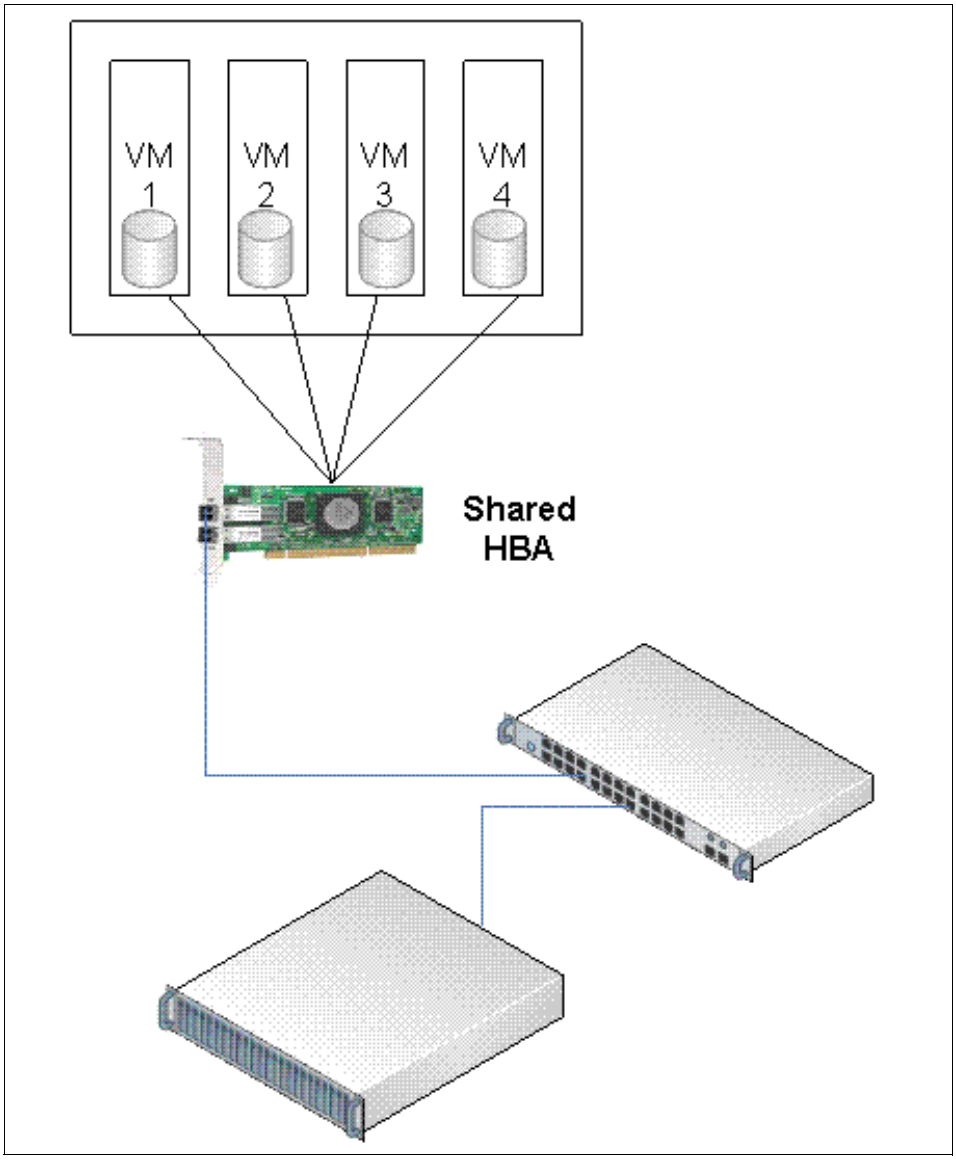


Figure 6-11 Single HBA with multiple virtual nodes

NPIV mode of blade server switch modules

On blade servers, when they are enabled with the NPIV mode, the FC switch modules that connect to an external SAN switch for access to storage act as an HBA N_port (instead of a switch E_port). The back-end ports are F_ports that connect to server blade modules.

Figure 6-12 shows the switch module in the NPIV mode.

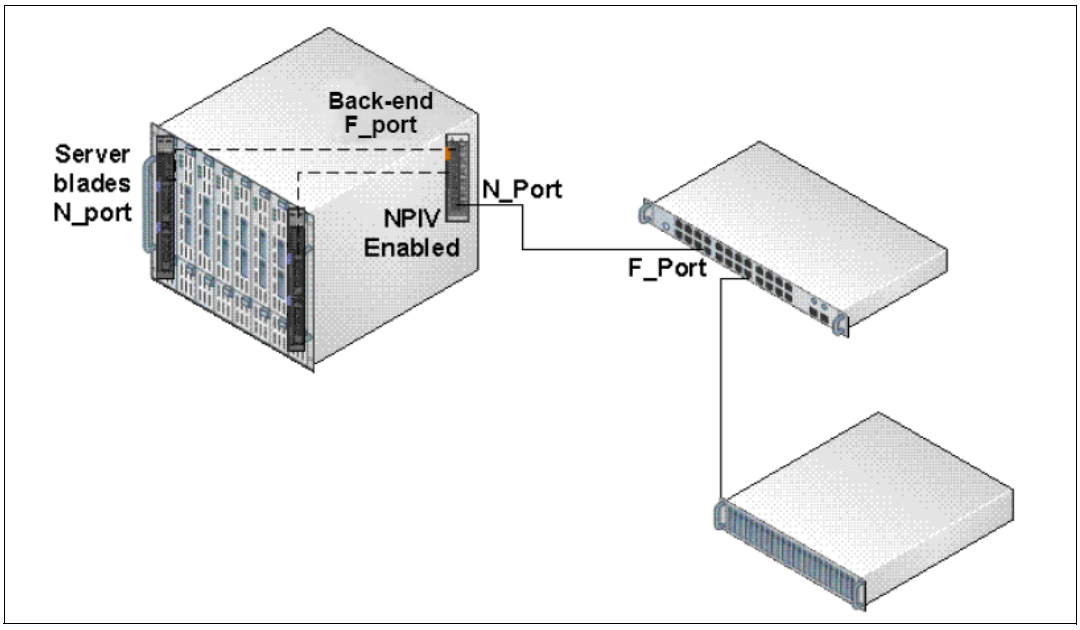


Figure 6-12 Blade server with FC switch module in the NPIV mode

With the NPIV mode, we can overcome the interoperability issues of merging external switches that might come from separate vendors to the blade server switch module. Also, management is easier because the blade switch module becomes a node in the fabric. And, we can overcome the scalability limitations of many switch domains for a switch module in blade servers.

6.4 Building a smarter cloud

Storage-as-a-Service (*SaaS*) is a business model in which a large company rents space in their storage infrastructure to a smaller company or individual. It is a good alternative when a small business lacks the capital budget or technical personnel to implement and maintain their own storage infrastructure. In certain circumstances, SaaS is 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 lead to relevant issues in terms of the planning and optimization of storage resources.

Most of these issues can be managed by ensuring that spare resources are available and by moving data, by using data mobility tools, or by using operating system features (such as host-level mirroring). However, all of 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 IBM Easy Tier® feature, which is available with the IBM 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, Easy Tier automatically and nondisruptively moves data, at the sub-LUN or subvolume 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 and with hierarchical storage management (HSM). Examples include functions that are provided by IBM Tivoli Storage Manager and 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 that is 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 that is dedicated to the volumes on the storage system. From the server’s or application’s perspective, thinly provisioned volumes are displayed and function the same as fully provisioned volumes. However, physical disk drive capacity is allocated only as needed (on demand) for write activity to the volumes. Deallocated physical capacity is available for use as needed by all volumes in a storage pool or even across an entire storage system.

Thin provisioning offers these advantages:

- It allows higher storage systems utilization, which in turn leads to a reduction in the amount of storage that you need, lowering your direct capital expenditure (CAPEX).
- It lowers your operational expenditure (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 you require more capacity, the storage will likely cost less.
- Capacity planning is simplified because you can manage a single pool of free storage. Multiple applications or users can allocate storage from the same free pool, avoiding the situation where certain volumes are capacity-constrained and other volumes have spare capacity.
- Your storage environment becomes more agile, and it is 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 needed for writing data. Large potential savings can result 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 that is used for data with the total amount of physical capacity that is allocated to a server. Historically, utilization ratios were under 50%, indicating a large amount of allocated but unused physical storage capacity. Often, the users and storage administrators are uncertain how much capacity is needed. However, they must ensure that they do not run out of space, and they also must allow for growth. As a result, users might request more capacity than they need and storage administrators might allocate more capacity than is requested, resulting in a 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 might 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. Deallocated 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 restripe data as capacity requirements grow. Without thin provisioning, if an application requires capacity beyond the capacity that is provided by its current set of volumes, two options are available:

- Existing volumes might be increased in size.
- Additional volumes might be provisioned.

In many environments, these options are challenging because of the required steps and potential disruption to make the volumes larger or additional volumes visible and optimized for the application.

With thin provisioning, large virtual or logical volumes might be created and presented to applications while the associated physical capacity grows only as needed, transparent to the application.

Without thin provisioning, physical capacity was dedicated at the time of volume creation, and storage systems typically did not display or report how much of the dedicated physical capacity was used for data. As storage systems implemented thin provisioning, physical allocation and usage became visible. Thin provisioning increases storage efficiency by making it easy to see the amount of physical capacity that is needed and used because physical space is not allocated until it is needed for data.

6.4.3 Data deduplication

Data deduplication emerged as a key technology to dramatically reduce the amount of space and the cost that are associated with storing large amounts of data. Data deduplication is the art of intelligently reducing storage needs in order of magnitude. This method is better than common data compression techniques. Data deduplication works through the elimination of redundant data so that only one instance of a data set is stored.

IBM has the broadest portfolio of data deduplication solutions in the industry, which gives IBM the freedom to solve client issues with the most effective technology. Whether it is 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 that is 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 and 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 depends on many variables, including the rate of data change, the number of backups, and the data retention period. For example, if you back up the same incompressible data one time a week for six months, you save the first copy and you do not save the next 24. This method provides a 25:1 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, this method provides a 2:1 data deduplication ratio. A more likely scenario is that a portion of your data changes from backup to backup so that your data deduplication ratio changes 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 6-13 shows the concept of data deduplication.

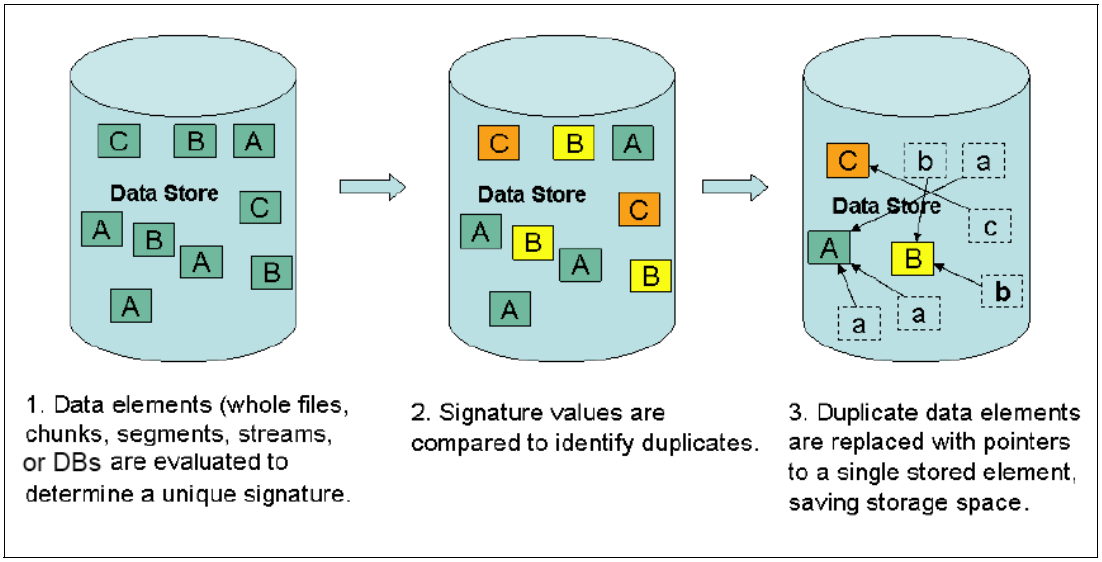


Figure 6-13 The concept of data deduplication

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 have the highest data deduplication ratios. However, product functions, such as IBM Tivoli Storage Manager Progressive Incremental or Oracle Recovery Manager (RMAN), can reduce the data deduplication ratio. Compressed, encrypted, or otherwise scrambled workloads typically do not benefit from data deduplication. Good candidates for data deduplication are text files, log files, uncompressed and non-encrypted database files, email files (PST, DBX, and IBM Domino®), and Snapshots (Filer Snaps, BCVs, and VMware images).

Types of data deduplication and IBM HyperFactor

Many vendors offer data deduplication products. Various methods are available to deduplicate data. The following three methods are used frequently for data deduplication:

- Hash-based data deduplication uses a hashing algorithm to identify chunks of data. Secure Hash Algorithm 1 (SHA-1) or Message-Digest Algorithm 5 (MDA-5) is commonly used. The details of each technique are beyond the intended scope of this publication.
- Content-aware data deduplication methods are aware of the structure of common patterns of data that is used by applications. The content-aware data deduplication method assumes that the best candidate to deduplicate against is an object with the same properties, such as a file name. When a file match is identified, a bit-by-bit comparison is performed to determine whether data changed and the changed data is saved.
- IBM HyperFactor® is a patented technology that is used in the IBM System Storage ProtecTIER Enterprise Edition and higher software. HyperFactor takes an approach that reduces the phenomenon of missed factoring opportunities, providing a more efficient process. With this approach, HyperFactor can surpass the reduction ratios that are 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 that is 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 that is used.

Figure 6-14 shows the HyperFactor technology.

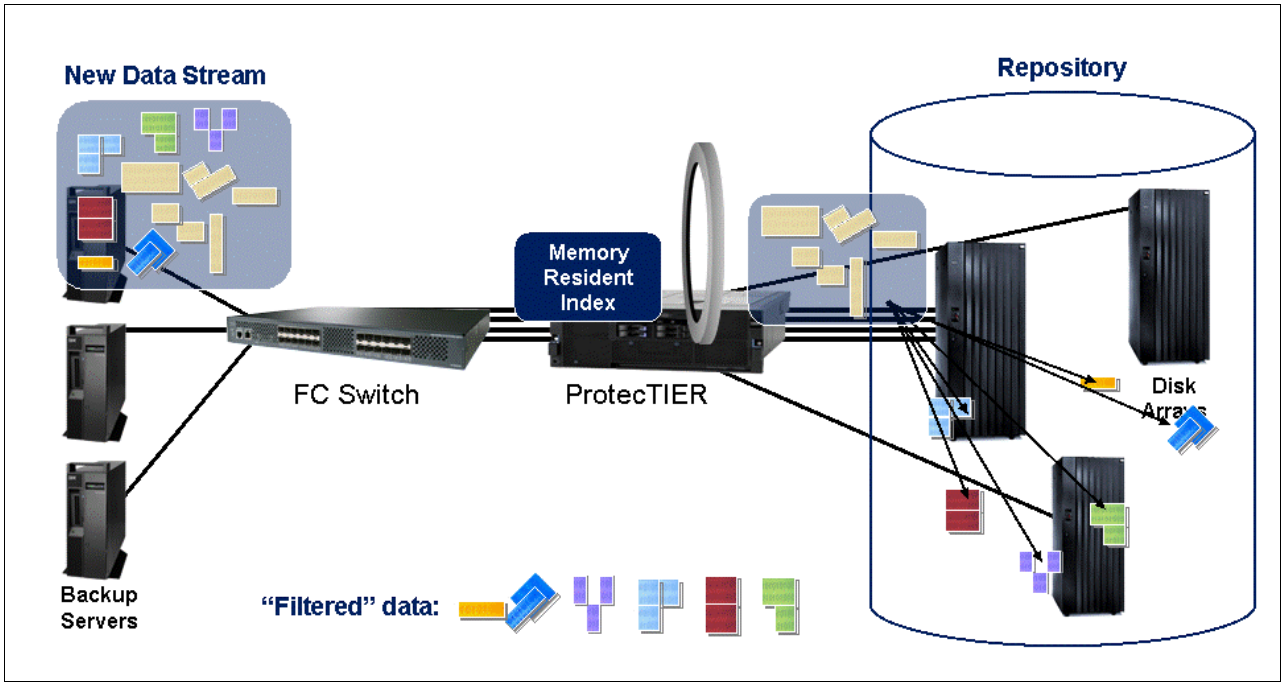


Figure 6-14 IBM HyperFactor technology

Data deduplication processing

Data deduplication can either occur while the data is backed up to the storage media (real-time or *inline*) or after the data is written to the storage media (post-processing). Each method contains positive and negative aspects. These considerations must be evaluated by the engineer or technical specialist that is responsible for the concrete solution architecture and deployment. IBM decided to use inline data deduplication processing because it offers larger target storage space without any need of a temporary disk cache pool for post-processed deduplication data.

Bit comparison techniques, such as the technique that is 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 is managed by new generation management tools because older tools generally lack the required features.

When these tools are used correctly, these tools can make the adoption of virtualization technology easier and more cost-effective. The tools offer the following additional benefits:

- Enable line-of-business insight into storage utilization and allocation, enabling easier departmental charge-back
- Offer information for more intelligent business decisions about storage efficiency so that you can respond faster to changing business demands, yet reduce costs
- Provide better understanding of application storage performance and utilization patterns that enable better information lifecycle management (ILM) of application data
- Allow an organization to manage the infrastructure proactively through the correct capacity management, rather than reactively
- Improve operational efficiency that leads 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. The most important factor to consider is how the choice of the technology affects the recovery time objective (RTO). For SANs, many possible solutions are available. The cloud design drives the selections that can meet your requirements. You must select a smart cloud that can guarantee business continuity and address any disaster recovery plans.

Disaster recovery: For more information “lessons learned” from disaster recovery and disaster recovery solutions, see *IBM Storage Infrastructure for Business Continuity*, REDP-4605:

http://www.redbooks.ibm.com/abstracts/redp4605.html?Open

6.4.6 Storage on demand

Scalable, pay-per-use cloud storage can help you to manage massive data growth and your storage budget. Massive data growth can lead to a costly procurement cycle with setup costs and implementation delays every time that more storage is needed. Cloud storage offers you the ability to expand your storage capacity immediately, and later, to shrink storage consumption, if necessary.

Cloud storage provides a ready-made data storage solution that helps in the following 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 several of the most common Fibre Channel storage area network (SAN) products and technologies. For a description of the IBM products that are in the IBM System Storage portfolio, see Chapter 12, “IBM Fibre Channel storage area network product portfolio” on page 247.

7.1 The environment

The Storage Networking Industry Association (SNIA) defines the meaning of SAN, Fibre Channel, and storage:

- Storage area network (SAN)

  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 that provides physical connections and a management layer. This layer 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 that can support multiple protocols, including access to open system storage (Fibre Channel Protocol (FCP)), access to mainframe storage (Fibre Channel connection (FICON)), and networking (TCP/IP). Fibre Channel supports point-to-point, arbitrated loop, and switched topologies with various copper and optical links that are running at speeds from 1 gigabit per second (Gbps) to 10 Gbps. The committee that is standardizing Fibre Channel is the INCITS Fibre Channel (T11) Technical Committee.

- Storage system

  A storage system that consists of storage elements, storage devices, computer systems, and appliances, plus all control software, which communicates 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. These components form a *network.* Any combination of devices that can interoperate is likely to be used.

A Fibre Channel network might consist of many types of interconnect entities, including directors, switches, hubs, routers, gateways, and bridges.

The deployment of these interconnect entities allows Fibre Channel networks of varying scale to be built. In smaller SAN environments, you can employ hubs for *Fibre Channel Arbitrated loop* (*FC-AL*) 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 component that creates a Fibre Channel SAN provides an individual management capability and participates in an often complex end-to-end management environment.

Figure 7-1 shows a generic SAN connection.

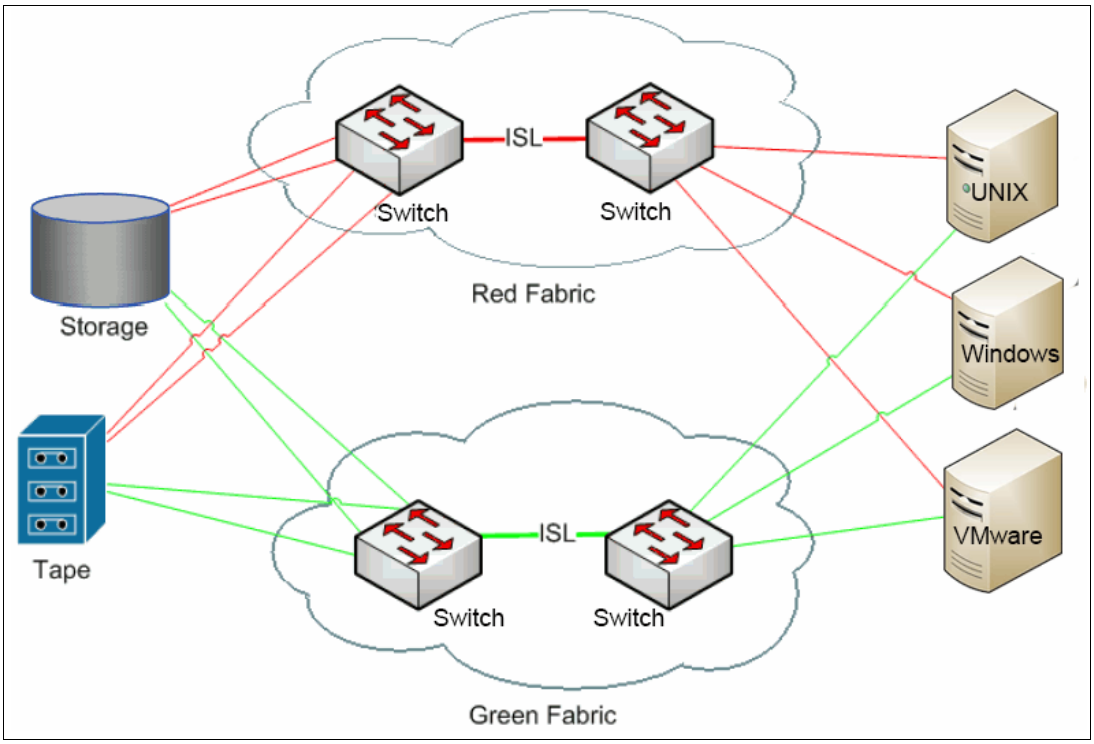


Figure 7-1 Generic SAN

7.2 Storage area network devices

A Fibre Channel SAN employs a *fabric* to connect devices, or *end points*. A fabric can be as simple as a single cable that connects two devices, similar to server-attached storage. However, the term is most often used to describe a more complex network to connect servers and storage by using 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.

Consider the following questions:

- 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 device?
- How high are my availability needs and expectations?
- Will I connect multiple platforms to the same fabric?
- What technology do I want to use, FC, Fibre Channel over Ethernet (FCoE) or Internet Small Computer System Interface (iSCSI)?

7.2.1 Fibre Channel bridges

Fibre Channel bridges allow the integration of traditional 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*.

Data center bridging: Do not confuse Fibre Channel bridges with data center bridging (DCB), although fundamentally they serve the same purpose, which is to interconnect different protocols.

A *bridge* is a device that converts signals and data from one form to another form. You can imagine these devices in a similar way as the bridges that we use to cross rivers. They act as a translator (a bridge) between two different protocols. These protocols can include the following types:

- Fibre Channel
- Internet Small Computer System Interface (iSCSI)
- Serial Storage Architecture (SSA)
- Fibre Channel over IP (FCIP)

We do not see many of these devices today, and they are considered historical devices.

7.2.2 Arbitrated loop hubs and switched hubs

Fibre Channel Arbitrated loop (*FC-AL*) is a Fibre Channel topology in which devices connect in a one-way loop fashion in a ring topology. This topology is also described in Chapter 5, “Topologies and other fabric services” on page 89.

Figure 7-2 shows an FC-AL topology.

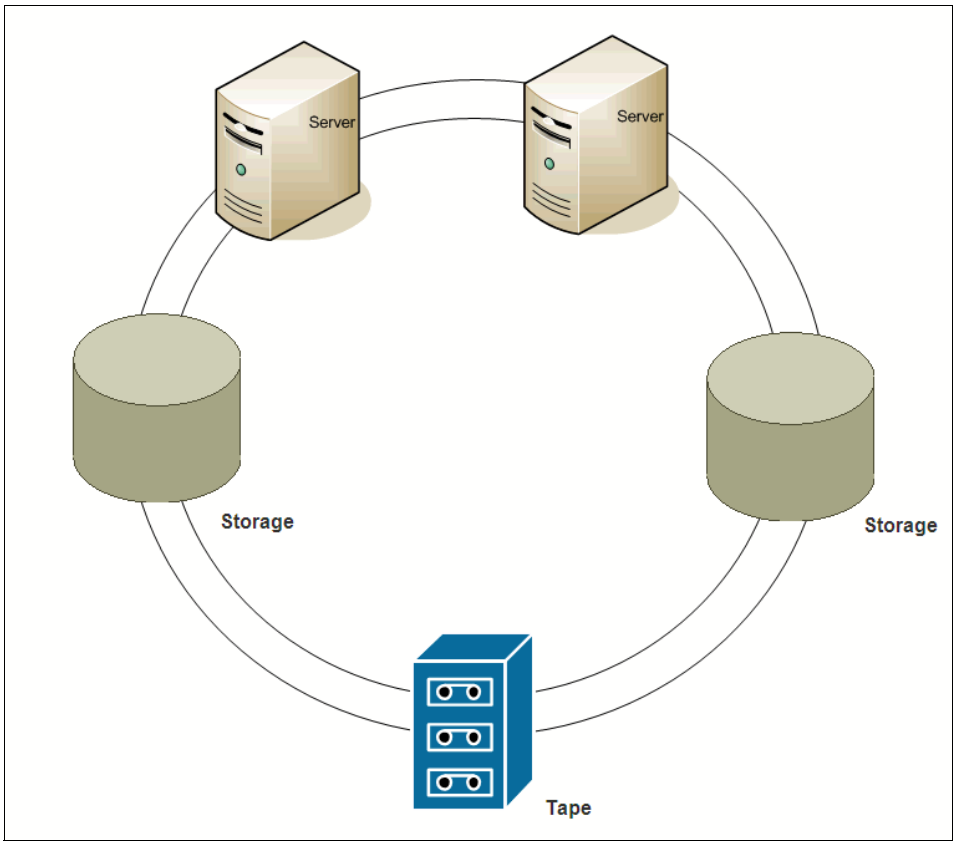


Figure 7-2 Arbitrated loop

In FC-AL, all devices on the loop share the bandwidth. The total number of devices that might 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 - 15.

Hubs are typically used in a SAN to attach devices or servers that do not support switched fabric-only FC-AL. They might be unmanaged hubs, managed hubs, or switched hubs.

Unmanaged hubs serve as cable concentrators and as a means to configure the arbitrated loop that is based on the connections that it detects. When one of the interfaces, typically a *gigabit interface converter* (*GBIC*), on the hub senses that no cable is connected, that interface shuts down. The hub port is then bypassed as part of the arbitrated loop configuration.

Managed hubs offer all of the benefits of unmanaged hubs, but in addition, they offer the ability to manage them remotely by using *Simple Network Management Protocol* (*SNMP*).

By using FC-AL, you can connect many servers and storage devices without using costly Fibre Channel switches. FC-AL is not used much today because switched fabrics now lead in the Fibre Channel market.

Switched hubs

Switched hubs allow devices to be connected in their 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 all share the bandwidth.

Switched hubs are typically managed hubs.

FC-AL: Originally, 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 an individual port basis in a switched fabric. For this reason, FC-AL implementations are, by certain observers, considered historical 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 the responder and operates at full Fibre Channel bandwidth.

Switches can be connected in cascades and meshes by using *inter-switch links (ISLs)* or *expansion ports (E_ports)*.

Note: Devices from different manufacturers might not interoperate fully.

The switch also provides various fabric services and features. The following list provides examples:

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

It is common to refer to switches as either core switches or edge switches, depending on where they are in the SAN. If the switch forms, or is part of the SAN backbone, it is the *core switch*. If the switch is mainly used to connect to hosts or storage, it is called an *edge switch*. Directors are also sometimes referred to as switches because they are essentially switches. Directors are large switches with higher redundancy than most normal switches.

7.2.4 Multiprotocol routing

Certain devices are multiprotocol routers and devices. Multiprotocol routers and devices 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, multiprotocol routers and devices support many protocols and offer their own features, such as zoning. The following list shows the supported protocols:

- Fibre Channel Protocol (FCP)
- Fibre Channel over IP (FCIP)
- Internet Fibre Channel Protocol (iFCP)
- Internet Small Computer System Interface (iSCSI)
- Internet Protocol (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 adopting a modular system approach to their components. Therefore, the 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, and storage services.

7.2.6 Multiplexers

Multiplexing is the process of simultaneously transmitting multiple signals over the same physical connection. Common types of multiplexing are used for fiber optic connections that are based on either time or wavelength:

- Time-division multiplexing (TDM)
- Wavelength division multiplexing (WDM)
- Dense wavelength division multiplexing (DWDM)

When you use multiplexers in a SAN environment, more parameters in the SAN switch configuration might be needed to ensure correct load balancing. Therefore, check with your SAN switch vendor for preferred practices.

Multiplexers: Usually multiplexers are transparent to the SAN fabric. If you are troubleshooting an ISL that covers a long distance, remember that the multiplexer, if installed, plays an important role in that path.

7.3 Components

Many components must interoperate to create a SAN. We identify several common components.

7.3.1 Application-specific integrated circuit

The fabric electronics use a personalized *application-specific integrated circuit* (*ASIC*) and its predefined set of elements. Examples of these types of elements include logic functions, I/O circuits, memory arrays, and backplanes to create specialized fabric interface components.

An ASIC provides services to Fibre Channel ports. The circuit might be used to connect to the following devices or ports:

- External node ports (N_ports), such as a fabric port (F_port) or a fabric loop port (FL_port)
- External loop devices, such as an FL_port
- Other switches, such as an expansion port (E_port)

The ASIC contains the Fibre Channel interface logic, and message and buffer queuing logic. The ASIC receives buffer memory for the on-chip ports and other support logic.

Frame filtering

Frame filtering is a feature that enables devices to provide zoning functions with finer granularity. Frame filtering can be used to set up port-level zoning, worldwide name (WWN) zoning, device-level zoning, protocol-level zoning, and logical unit number (LUN)-level zoning. Frame filtering is commonly carried out by an ASIC. After you set up the filter, the complicated function of zoning and filtering can be achieved at wire speed.

7.3.2 Fibre Channel transmission rates

Fibre Channel transmission rates are sometimes referred to as *feeds and speeds*. The number of vendor offerings for switches, HBAs, and storage devices grows constantly. Currently, the 16 Gb FC port has the fastest line rate that is supported for an IBM SAN. The 16 Gb FC port uses a 14.025 Gbps transfer rate and 64b/66b encoding, which provides approximately 1600 MBps in throughput. The 8 Gb FC port has a line rate of 8.5 Gbps that uses 8b/10b encoding, which results in approximately 800 MBps. When you compare feeds and speeds, the FC ports are sometimes referred to as *full duplex*. The transceiver and receiver parts of the FC port are then added, therefore “doubling” the MBps.

Encoding: By introducing the 64b/66b encoding to Fibre Channel, the encoding overhead is reduced from approximately 20% by using 8b/10b encoding to approximately 3% with the 64b/66b encoding.

The new 16 Gb FC port is approved by the Fibre Channel Industry Association (FCIA). This approval ensures that each port speed can communicate with at least two previously approved port speeds. For example, 16 Gb can communicate with 8 Gb and 4 Gb.

The FCIA also created a roadmap for future feeds and speeds. For more information, see this website:

`http://www.fibrechannel.org/`

7.3.3 SerDes

The communication over a fiber, whether optical or copper, is serial. Computer busses, however, use parallel busses. Therefore, Fibre Channel devices must be able to convert between the two types. For this conversion, the devices 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 that contains all of the components in a device, sometimes the design that is used is that of a backplane and blades. For example, directors and large core switches typically implement this technology.

The *backplane* is a circuit board with multiple connectors into which other cards can be plugged. These other cards are typically referred as *blades* or *modules*, but other terms can be used.

If the backplane is in the center of the unit with blades that are plugged in at the back and front, the backplane is referred to as a *midplane*.

7.4 Gigabit transport technology

In Fibre Channel technology, frames are moved from the source to the destination by using *gigabit transport*, which is required to achieve fast transfer rates. To communicate with gigabit transport, both sides must support this type of communication. You can obtain this support by installing this feature in the device or by using specially designed interfaces that can convert other communication transport into gigabit transport. The *bit error rate* (*BER*) allows for only a single bit error to occur one time in every 1,000,000,000,000 bits in the Fibre Channel standard. Gigabit transport can be used in a copper or fiber optic infrastructure.

Layer 1 of the *open systems interconnection* (*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 the common concepts of the Layer 1 level.

7.4.1 Fibre Channel cabling

Fibre Channel cabling is available in two forms: *fiber optic cabling* or *copper cabling*. Fiber optic cabling is the typical cabling type. But, *Fibre Channel over Ethernet* (*FCoE*) copper cabling is also available.

Fiber optic cabling is more expensive than copper cabling. The optical components for devices and switches and the cost of any client 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. Fiber optic cabling is resistant to signaling distortion by electromagnetic interference.

Fiber optic cabling

In copper cabling, electric signals are used to transmit data through the network. The copper cabling is 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 that is used for the transmission (Figure 7-3).

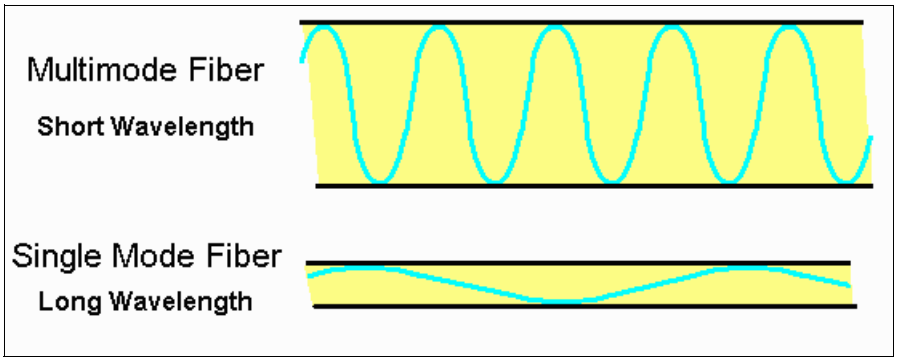


Figure 7-3 Multimode versus single-mode optic signaling

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 on page 159). The core diameter for single-mode

cabling is
nine microns in diameter (Figure 7-4).

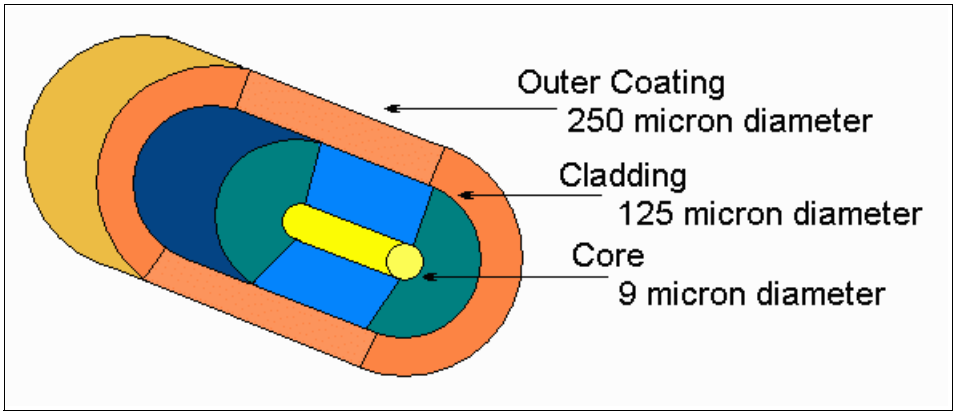


Figure 7-4 Single-mode fiber cable

Multimode fiber

Multimode fiber (*MMF*) uses short wavelength light to transmit data and requires a cable with a larger core for transmission (see Figure 7-3 on page 159). The core diameter for multimode cabling can be 50 or 62.5 microns in diameter (Figure 7-5).

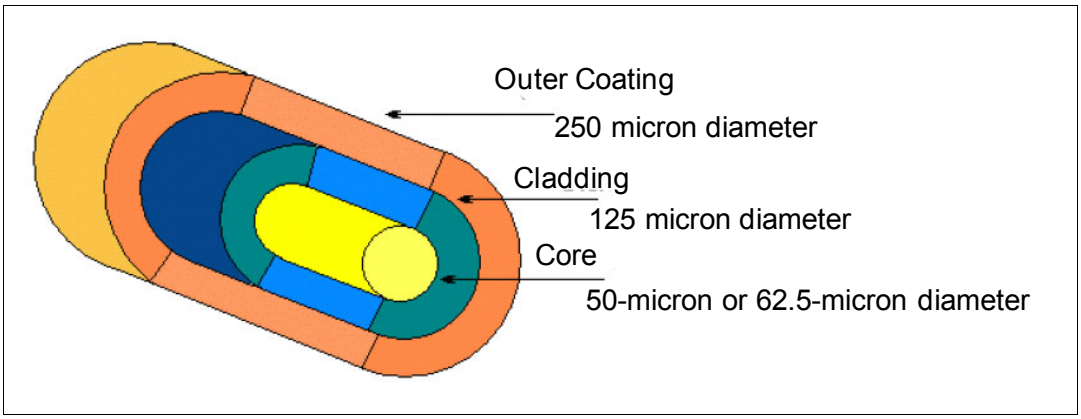


Figure 7-5 Multimode fiber cable

The color of the outer coating is sometimes used to identify whether a cable is a multimode or single-mode fiber cable, but the color is not a reliable method. The *Telecommunications Industry Association-598C* (*TIA-598C*) standard suggests a yellow outer coating for single mode fiber and an orange outer coating 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 that are printed on the outer coating of the cabling. See also Figure 7-7 on page 161 and Figure 7-8 on page 161.

Figure 7-6 Blue 62.5-micron MMF cable

Figure 7-7 shows the yellow SMF cable.

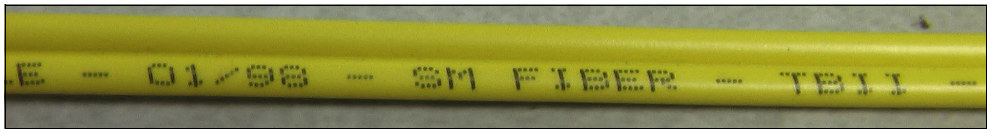


Figure 7-7 Yellow SMF cable

Figure 7-8 shows the orange 50-micron MMF cable.

Figure 7-8 Orange 50-micron MMF cable

Copper cabling

When we refer to *copper cabling,* we mean that the material that is used to transfer the signals is made of copper. The most common copper wire is the twisted-pair cable that is used for normal Ethernet. This type of cabling 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 categories of twisted-pair cabling are available (Table 7-1). These categories indicate the signaling capabilities of the cabling.

Table 7-1 TIA/Electronic Industries Alliance (EIA) cabling categories

TIA/EIA cabling category	Maximum network speeds that are 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 that is used for Ethernet twisted-pair cabling is likely the connector that most people recognize and associate with networking, which is the *RJ45 connector*. Figure 7-9 shows this connector.

Figure 7-9 RJ45 copper connector

Twisted-pair cabling contains four pairs of wire inside the cable (Figure 7-10).

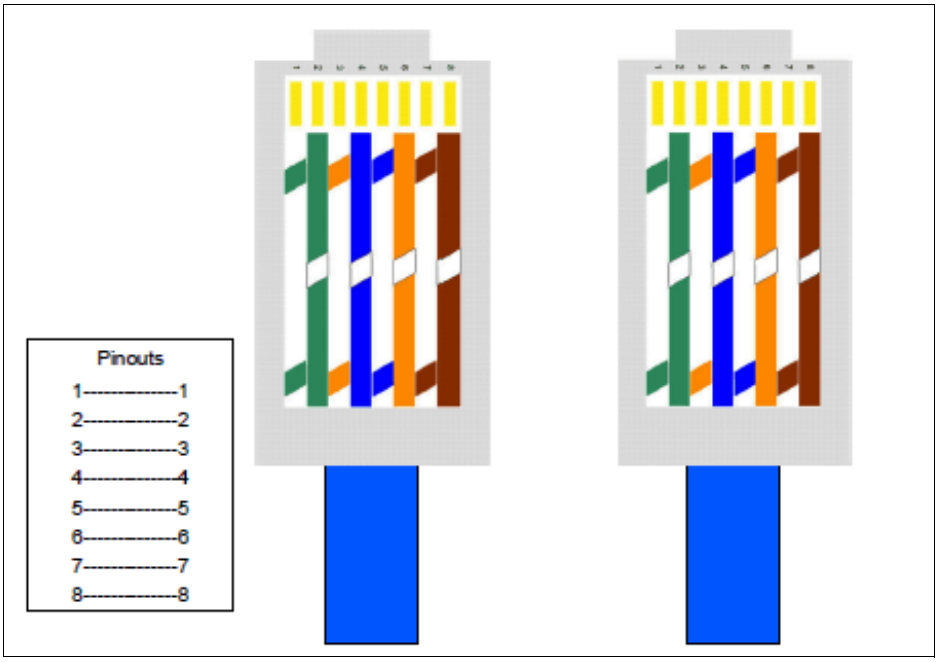


Figure 7-10 Straight-through Ethernet cable

An Ethernet that is operating in 10/100 Mb mode uses only two pairs: pairs 1-2 and 3-6. An Ethernet that is operating in 1 Gb mode uses all four pairs: pairs 1-2, 3-6, 4-5, and 7-8. Distances up to 100 meters (328.08 feet) 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 due to multiple factors, such as vendor support, cabling type, electromagnetic interference, and the number of physical connections in the segment.

Twinax cabling

Twinax cables were used by IBM for many years, but they were recently reintroduced to the market as a transport media for 10 Gb Ethernet. One of the greatest benefits of a twinax cable is its low power consumption. Also, this cable costs less 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 that is used in networking today is the *LC connector* (Figure 7-11).

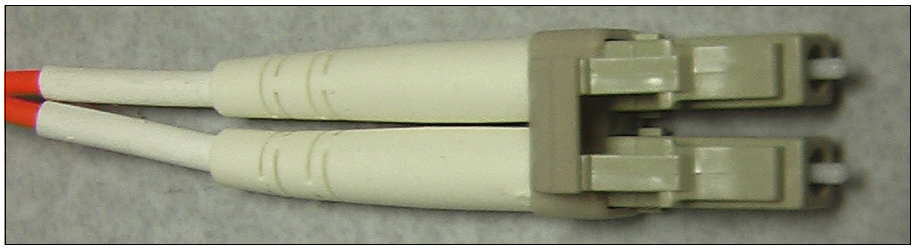


Figure 7-11 LC fiber connector

Other types of connectors are the SC connector (Figure 7-12) and the ST connector (not shown).

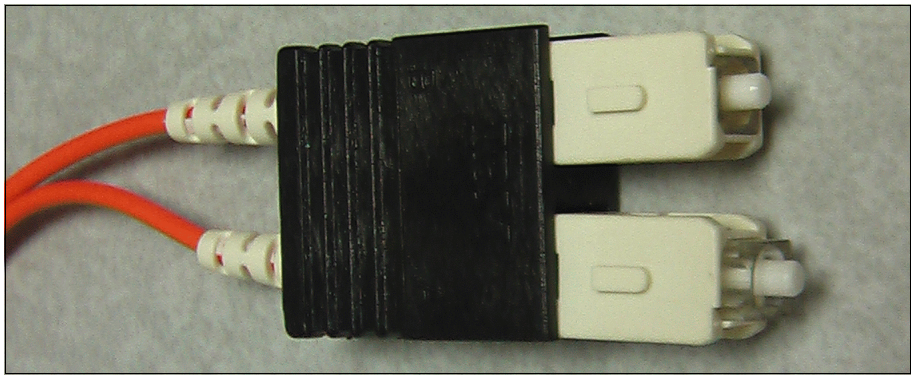


Figure 7-12 SC fiber connector

7.4.2 Transceivers

A *transceiver* or *transmitter/receiver* is the fiber optic port of a device where the fiber optic cables connect. Occasionally, a device might include an integrated transceiver, which limits the flexibility in the type of cabling that you can use. Most devices provide a slot to insert a modular transceiver, providing flexibility so that you can select either single or multimode implementations.

Certain equipment might use a larger transceiver that is known as a *Gigabit Interface Converter* (GBIC) (Figure 7-13). As technology advances, smaller transceivers are introduced. These smaller transceivers provide much higher port density, such as small form-factor pluggables (SFPs), 10 Gigabit SFP+, 10 Gigabit SFP-XFP, and Quad SFP (QSFP).

Figure 7-13 Gigabit Interface Converter (GBIC)

Figure 7-14 shows the various transceivers.

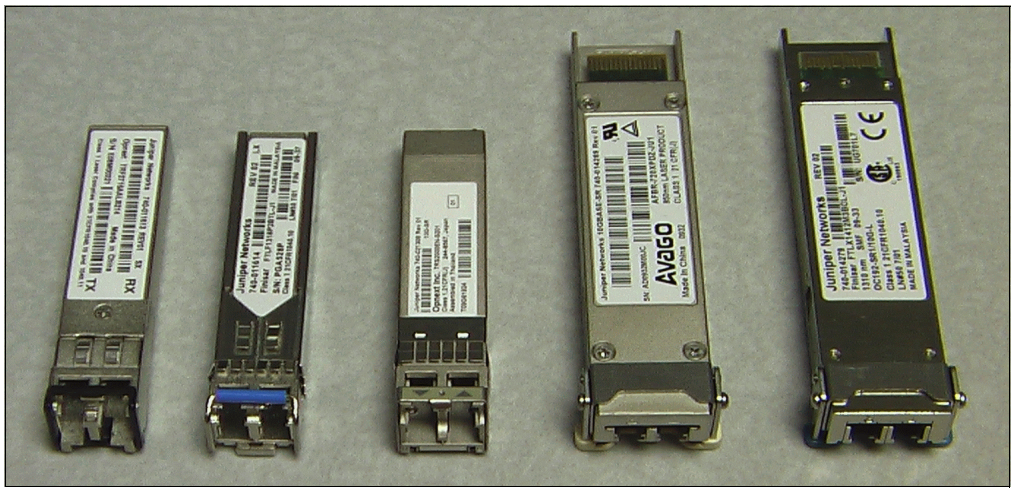


Figure 7-14 From left to right: SFP-MMF, SFP-SMF, SFP+-MMF, XFP-MMF, and XFP-SMF

Figure 7-15 shows a QSFP and cable.

Figure 7-15 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*).

Fibre Channel host bus adapter

The HBA connects to the bus of the host or storage system. Certain devices offer more than one Fibre Channel connection and include 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.

Figure 7-16 shows an HBA.

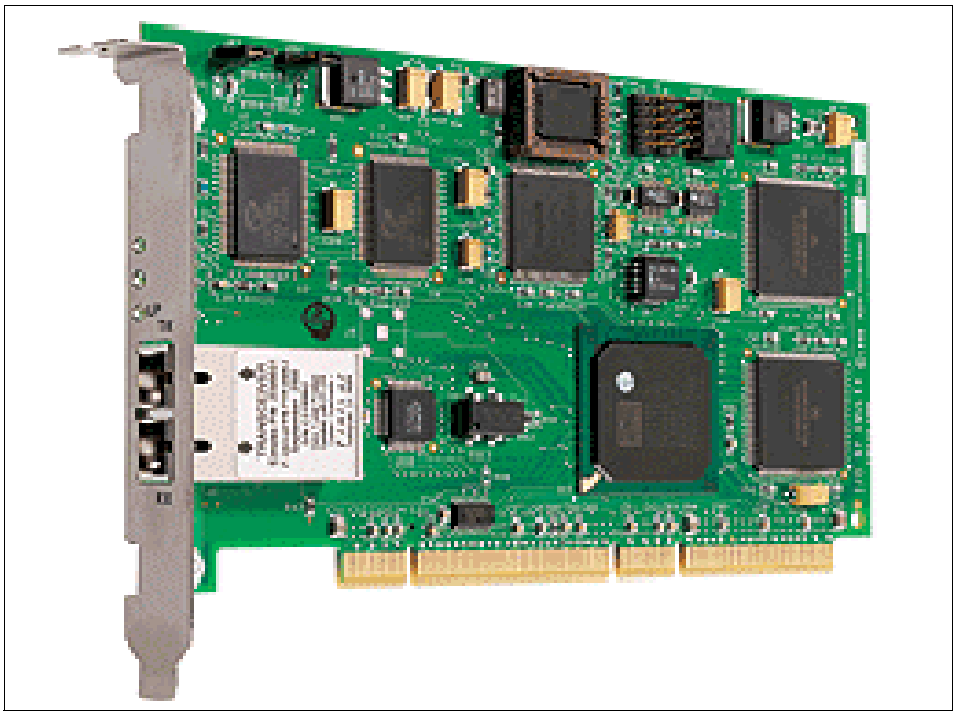


Figure 7-16 Host bus adapter (HBA)

Several manufacturers offer HBAs. The choice of the HBA is an important consideration when you plan a SAN. HBAs might include more than one port. They can be supported by certain equipment and not other equipment. HBAs might include parameters to tune the system. Many other features are available. HBAs also have a certain number of buffer-to-buffer credits.

Important: If you are considering the use of an HBA with multiple virtual machines behind it, the choice of an HBA is a critical decision.

Converged Network Adapter host bus adapter

Converged Network Adapters (*CNAs*) can run both *Converged Enhanced Ethernet* (*CEE*) and Fibre Channel traffic at the same time. These CNAs combine the functions of an HBA and a Network Interface Card (NIC) on one card. CNAs fully support FCoE protocols and allow Fibre Channel traffic to converge onto 10 Gbps CEE networks. These adapters play a critical role in the FCoE implementation.

When you implement CNAs, the CNAs can significantly reduce data center costs by converging data and storage networking. Standard TCP/IP and Fibre Channel traffic can both run on the same high-speed 10 Gbps Ethernet wire. You can save costs through reduced requirements for adapters, switches, cabling, power, cooling, and management. CNAs gained rapid market acceptance because they deliver excellent performance, help reduce data center total cost of ownership (TCO), and protect the current data center investment.

The 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, the full hardware offload for FCoE protocol processing reduces system processor utilization for I/O operations. This reduction leads to faster application performance and higher levels of consolidation in virtualized systems.

Figure 7-17 shows a dual-port CNA.

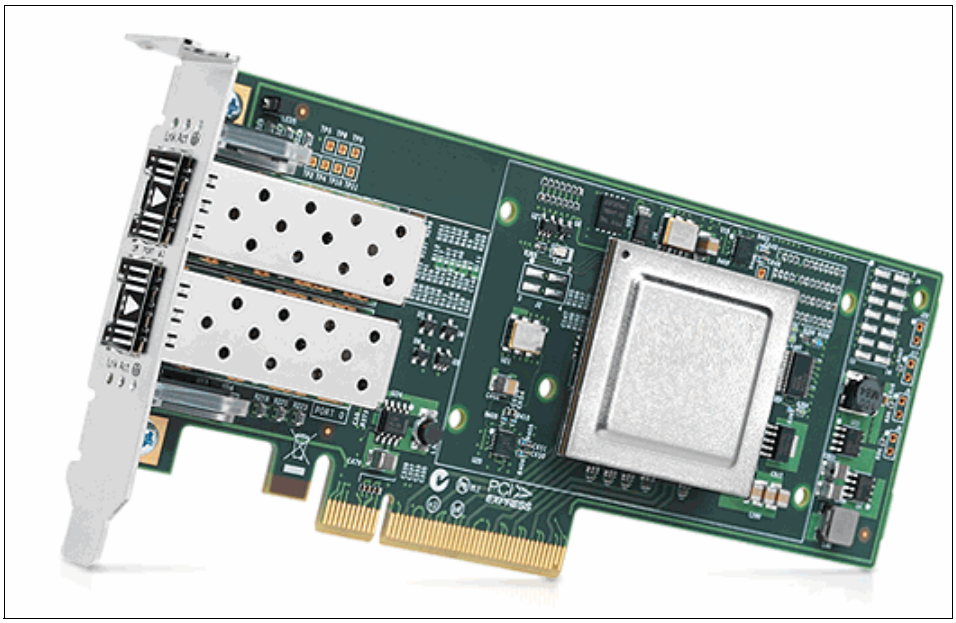


Figure 7-17 Dual-port Converged Network Adapter (CNA)

7.5 Inter-switch links

A link that joins a port on one switch to a port on another switch is called an *inter-switch link* (*ISL*). These ports are referred to as *E_ports*.

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

Before an ISL can carry frames that originate from the node ports, the joining switches undergo a synchronization process in which the operating parameters are interchanged. If the operating parameters are incompatible, the switches cannot join, and the ISL becomes *segmented*. Segmented ISLs cannot carry traffic that originates on node ports, but they can still carry management and control frames.

You can connect an E_port to a Fibre Channel router or a switch with embedded routing capabilities. Then, the port becomes an EX_port on the router side. Brocade calls these ports an *inter-fabric link* (*IFL*). However, Cisco uses *trunked E_port (TE_port)* or *extended ISL* (*EISL*), which allows traffic (from multiple virtual SANs (VSANs)) to be routed through that link.

7.5.1 Cascading

Expanding the fabric is called *switch cascading*, or *cascading*. Cascading is interconnecting Fibre Channel switches and directors by using ISLs. By cascading switches, the following benefits are possible for a SAN environment:

- The fabric can be seamlessly extended. Additional switches can be added to the fabric without powering down the existing fabric.
- You can increase the distance between various SAN participants easily.
- By adding switches to the fabric, you increase connectivity by providing more available ports.
- Cascading provides high resilience in the fabric.
- With ISLs, you can increase the bandwidth. The frames between the switches are delivered over all available data paths. Create more ISLs to increase the speed of the frame delivery. However, be careful to ensure that you do not introduce a bottleneck.
- When the fabric grows, the name server is fully distributed across all of the switches in the fabric.
- With cascading, you also provide greater fault tolerance within the fabric.

7.5.2 Hops

When Fibre Channel traffic traverses an ISL, this process is known as a *hop*. Or, to state it another way, traffic that goes from one E_port over an ISL to another E_port is one hop. ISLs are created by connecting an E_port to an E_port. Figure 7-18 on page 169 shows an illustration of the hop count from server to storage.

The hop count is limited. This limit is set by the fabric operating system. The limit is used to derive a frame hold time value for each switch. This value is the maximum amount of time that a frame can be held in a switch before the frame is dropped, or the fabric indicates that it is too busy. The hop count limits must be investigated and considered in any SAN design work because the hop count limit significantly affects the proposal.

7.5.3 Fabric shortest path first

Although this next topic is not a physical component, it is important to understand the concept of *fabric shortest path first* (*FSPF*) at this stage. According to the FC-SW-2 standard, FSPF is a link state path selection protocol. FSPF tracks the links on all switches in the fabric (in routing tables) and associates a cost with each link. The protocol computes the paths from a switch to all of the other switches in the fabric. This process is performed by adding the cost of all of the links that are traversed by the path and by choosing the path that minimizes the cost, that is, the shortest link.

For example, as shown in Figure 7-18, if a server needs to connect to its storage through multiple switches, FSPF routes all traffic from this server to its storage through switch A directly to switch C. This path is taken because it has a lower cost than traveling through more hops through switch B.

Figure 7-18 shows hops in a fabric.

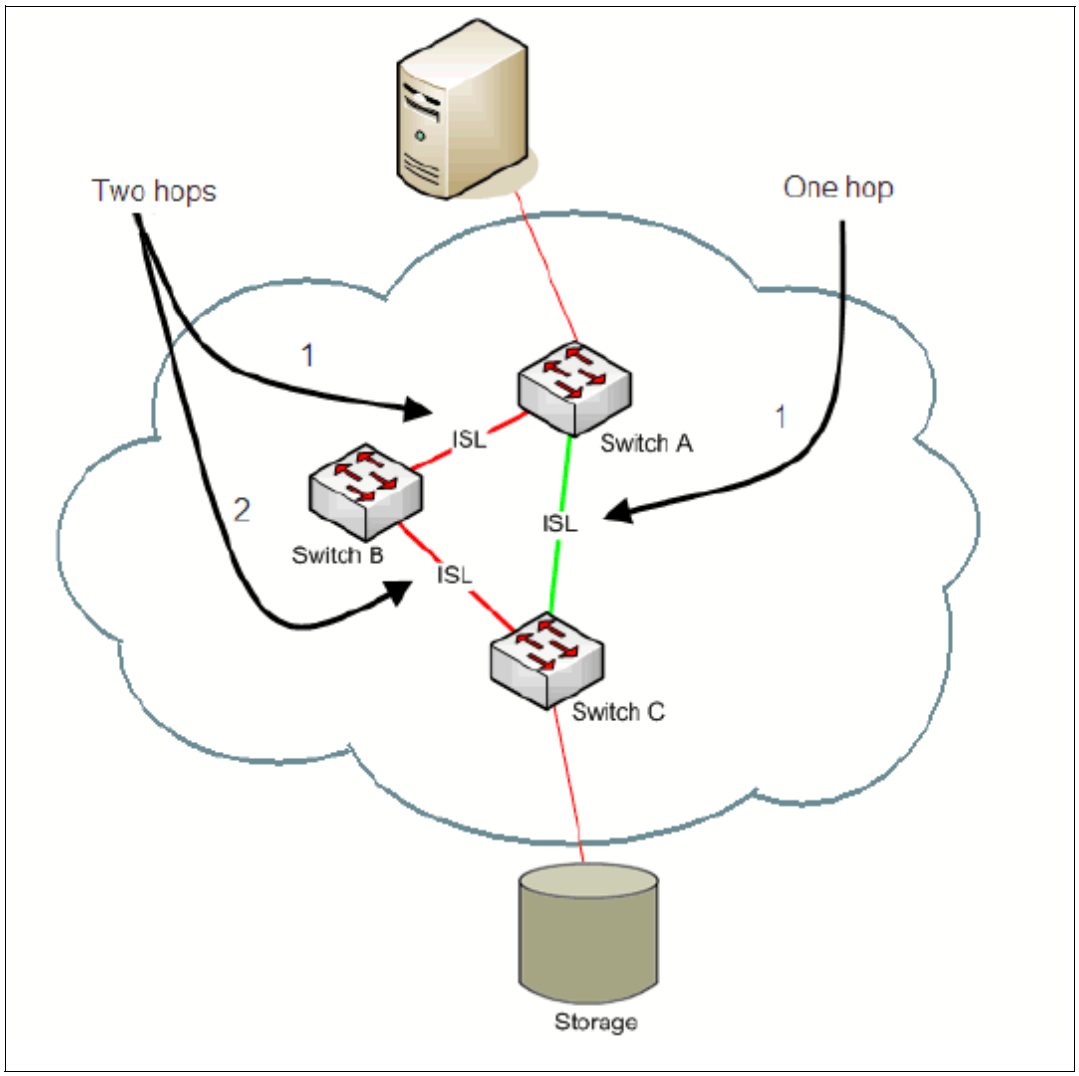


Figure 7-18 Hops example

FSPF is based on the hop count cost.

The collection of link states, including the cost, of all of the switches in a fabric constitutes the *topology database*, or *link state database*. The topology database is kept in all of the switches in the fabric. The switches are maintained and synchronized to each other. An *initial database synchronization* occurs, and an *update mechanism* is used. The initial database synchronization is used when a switch is initialized, or when an ISL comes up. The update mechanism is used when a link state changes, for example, when an ISL goes down or comes up, and on a periodic basis. This mechanism ensures consistency among all switches in the fabric.

7.5.4 Non-blocking architecture

To support high performance fabrics, the fabric components, switches, or directors must be able to move around data. These components must move the data without affecting other ports, targets, or initiators that are on the same fabric. If the internal structure of a switch or director cannot move data without an effect, the situation is called blocking.

Blocking

Blocking means that the data does not get to the destination. Blocking is not the same as *congestion* because with congestion, data is still delivered, but delayed. Currently, almost all Fibre Channel switches are created by using non-blocking architecture.

Non-blocking

A *non-blocking* architecture is used by most switch vendors. Non-blocking switches enable multiple connections that travel through the switch at the same time. Figure 7-19 shows this concept.

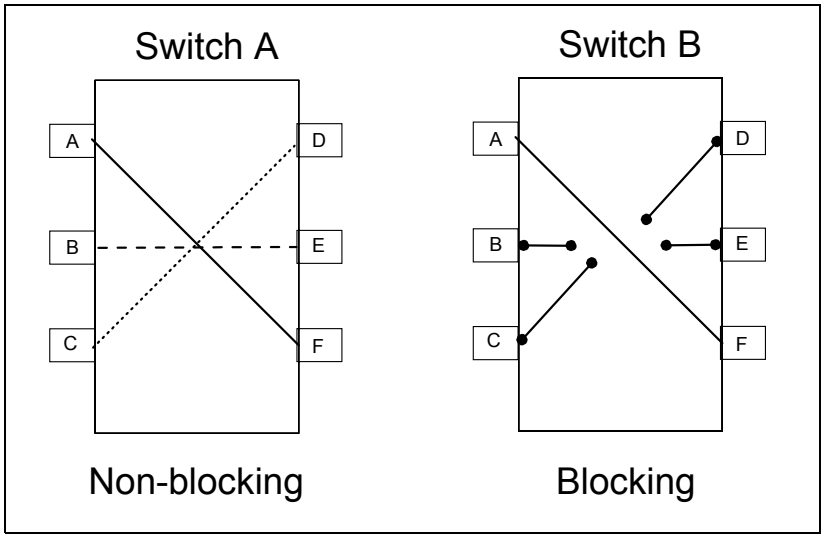


Figure 7-19 Non-blocking and blocking switching

In Figure 7-19, port A in the non-blocking Switch A communicates to port F, Port B communicates to port E, and port C communicates to port D, without any form of suspension of communication or a delay. That is, the communication is not blocked. In the blocking switch, Switch B, while port A communicates to port F, the switch is stopped or blocked from all other communication and does not continue until port A finishes communicating with port F.

7.5.5 Latency

Typically, in the SAN world, *latency* is the time that it takes for a Fibre Channel frame to traverse the fabric. When we describe the SAN, the latency in a SAN is rarely considered because it is in the low microsecond range. This concept is sometimes confused with *disk latency,* which is the measure of how quickly or slowly a storage target completes a read or write request that is sent from the server. However, when we describe long distances, all latency, both storage and SAN, plays a significant role.

Latency increases as the number of ISLs increases because the Fibre Channel frame must traverse the fabric by using ISLs. By fabric, we mean the Fibre Channel components and any latency discussion that relates to the SAN. Usually, the time that is taken is expressed in microseconds, which indicates the performance characteristics of the SAN fabric. Latency is often provided at a switch level, and sometimes at a fabric level.

7.5.6 Oversubscription

Another aspect of data flow is the *fan-in ratio,* which is also called the *oversubscription ratio* and frequently the *fan-out ratio* from the storage device's perspective, both in terms of host ports to target ports, and the device to the ISL. This ratio is the number of device ports that need to share a single port.

For example, two servers, each of which is equipped with a 4 Gb port (4+4=8 Gb) are both communicating with a storage device through a single 4 Gb port, which provides a 2:1 ratio. Therefore, the total theoretical input is higher than the input that the port can provide. Figure 7-20 on page 172 shows a typical oversubscription through an ISL.

Oversubscription can occur on storage device ports and ISLs. When you design a SAN, it is important to consider the possible traffic patterns to determine the possibility of oversubscription. An oversubscription might result in degraded performance. You can overcome the oversubscription of an ISL by adding an ISL between the switches to increase the bandwidth. Oversubscription to a storage device might be overcome by adding more ports from the storage device to the fabric.

Oversubscription: Vendors differ in how they practice utilization on their stated overall bandwidth for each chassis. Vendors use both storage port and ISL oversubscription. Verify the oversubscription preferred practices with your switch vendor.

7.5.7 Congestion

Oversubscription leads to a condition that is called *congestion*. When a node is unable to use as much bandwidth as it wants, because of contention with another node, congestion occurs. A port, link, or fabric can be congested. This condition normally affects the application directly and results in poor performance.

Congestion can be difficult to detect because it can also relate directly to buffer-to-buffer credit starvation in the switch port. Therefore, when you look at the data throughput from the switch, you see normal or less traffic flowing through the ports. However, the server I/O is unable to perform because the data cannot be transported because of a lack of buffer-to-buffer credits.

7.5.8 Trunking or port-channeling

One means of delivering high availability at the network level is the aggregation of multiple physical ISLs into a single logical interface. You can use this aggregation to provide link redundancy, greater aggregated bandwidth, and load balancing. Cisco calls this technology *port channeling*. Other vendors call this technology *trunking*.

Figure 7-20 shows the concepts of trunking.

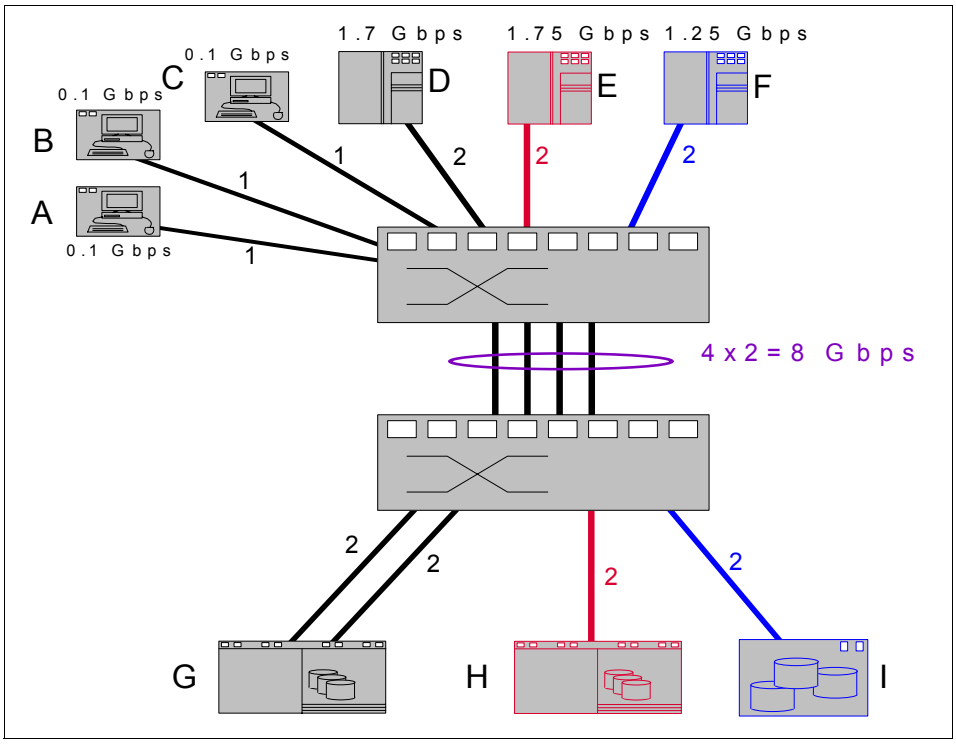


Figure 7-20 Trunking

In Figure 7-20, six computers access three storage devices. Computers A, B, C, and D communicate with storage G. Server E communicates with storage H. 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 allow FSPF alone to decide the routing, servers D and E might both use the same ISL. This situation leads to oversubscription and therefore congestion because 1.7+1.75 is greater than 2.

If all of the ISLs are gathered into a trunk, effectively they are a single, large ISL. They appear as 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. Therefore, one of the ISLs might fail and you still have sufficient bandwidth to satisfy the needs.

When the nodes come up, FSPF simply sees one route, and the nodes are all assigned a route over the same trunk. The fabric operating systems in the switches share the load over the actual ISLs, which combine to make up the trunk. This process is performed by distributing frames over the physical links and then reassembling them at the destination switch so that an in-order delivery can be assured, if necessary. And, to FSPF, a trunk is displayed as a single, low-cost ISL.

8

Management

Management is a key issue behind the concept of *infrastructure simplification*. The ability to manage heterogeneous systems at different levels as though they were a fully integrated infrastructure is a goal that many vendors and developers strive to achieve. Another goal is to offer the system administrator a unified view of the whole storage area network (SAN).

In this chapter, we look at several initiatives in the field of SAN management. These solutions incrementally smooth the way toward infrastructure simplification.

8.1 Management principles

SAN management systems typically are made up of a set of multiple-level software components that provide tools for monitoring, configuring, controlling (performing actions), diagnosing, and troubleshooting a SAN. We briefly describe the types and levels of management in a typical SAN implementation. We also describe the current efforts toward the establishment of open and general-purpose standards for building interoperable, manageable components.

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 techniques are usually independent of each other. To pretend that one SAN management solution provides a single point of control, which can perform every possible action, is premature.

We do not intend to describe each vendor’s standards fully. We present an overview of the myriad of possibilities in the IT environment. The high-level features of any SAN management solution are likely to include most of the following functions:

- Capacity management
- Device management
- Fabric management
- Proactive monitoring
- Fault isolation and troubleshooting
- Centralized management
- Remote management
- Performance management
- Security and standard compliance

8.1.1 Management types

Two philosophies are used to build management mechanisms:

- *In-band management*: The management data, such as status information, action requests, and events, flows through the same path as the storage data.
- *Out-of-band management*: The management data flows through a dedicated path, therefore not sharing the same physical path that is used by the storage data.

Figure 8-1 shows in-band and out-of-band models. These models are not mutually exclusive. In many environments, you might want a combination of the models.

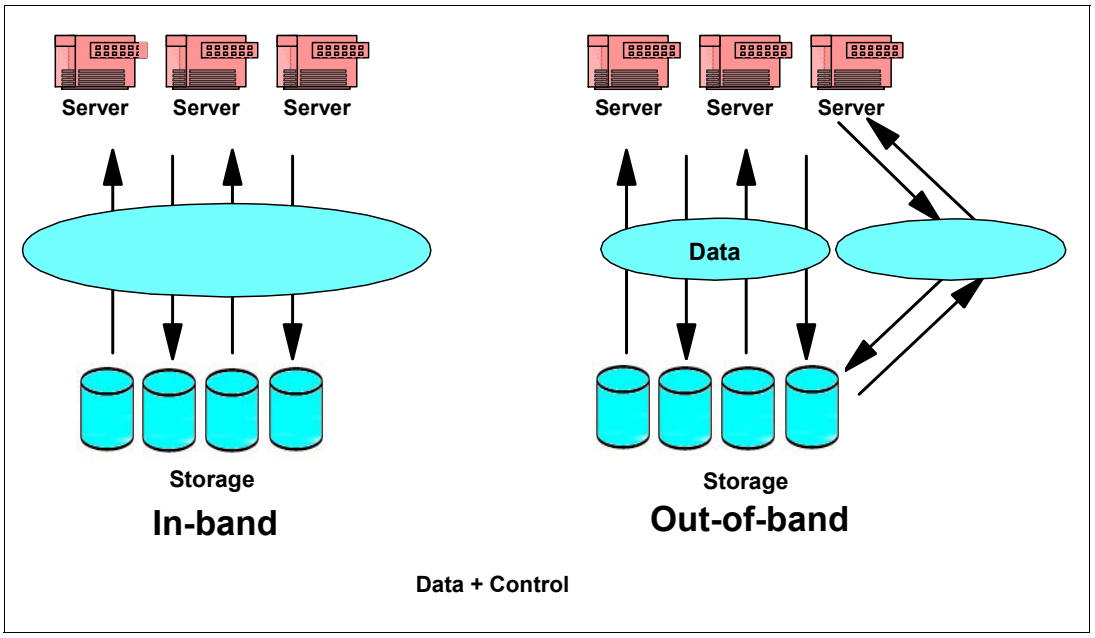


Figure 8-1 In-band and out-of-band models

The in-band approach is simple to implement. This approach requires no dedicated channels (other than LAN connections). It offers inherent advantages, such as the ability for a switch to initiate a SAN topology map with queries to other fabric components. However, if the Fibre Channel transport fails, the management information cannot be transmitted. Therefore, the access to devices and the ability to detect, isolate, and recover from network problems are lost. 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. Because in-band management is performed over the SAN, administrators are not required to manage any additional connections.

Conversely, out-of-band management does not rely on the storage network; its major 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, in-band management offers these major 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 offers these major advantages:

- Management is possible, even if a device is down.
- Management is accessible from anywhere in the routed network.
- Management traffic is kept out of the Fibre Channel, so that the management traffic does not affect the business-critical data flow on the storage network.

8.1.2 Connecting to storage area network management tools

A typical method to connect to a SAN device (Fibre Channel switches and storage devices that connect to a SAN) is by connecting through the Ethernet to a storage management device on a network segment that is intended for storage devices (Figure 8-2).

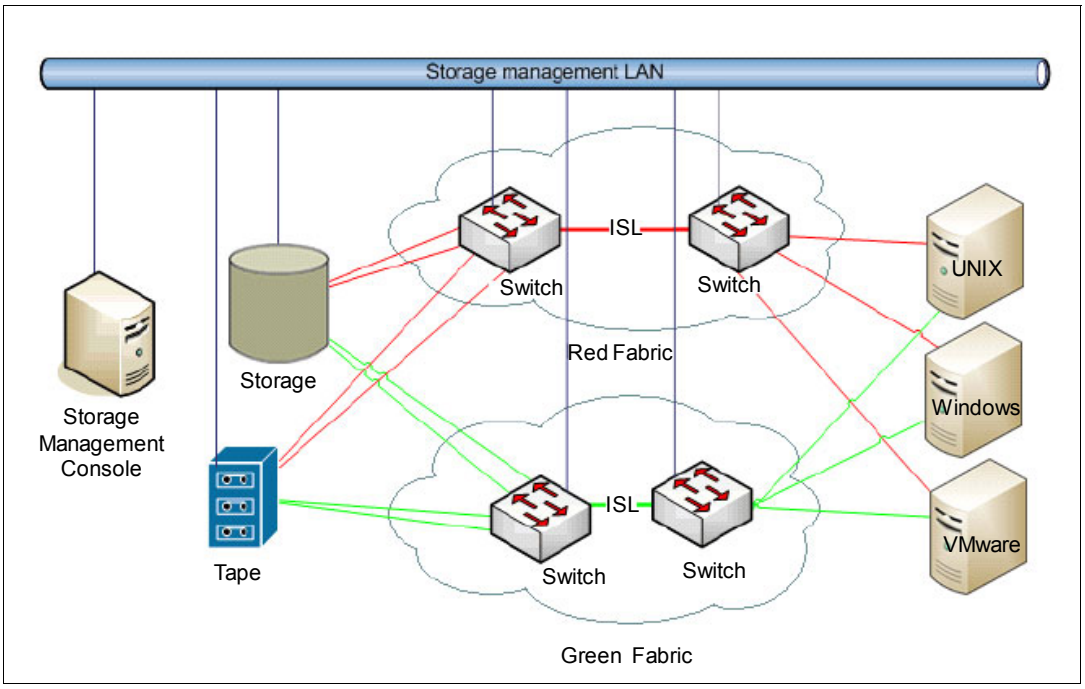


Figure 8-2 Storage management network

The SAN storage level consists of the storage devices that integrate the SAN, such as disks, disk arrays, tapes, and tape libraries. Because the configuration of a storage resource must be integrated with the configuration of the server's logical view of the storage resources, the SAN storage level management can also span both storage resources and servers.

8.1.3 Storage area network fault isolation and troubleshooting

In addition to providing tools for monitoring and configuring a SAN, one key benefit that a well-designed management mechanism offers is the ability to detect, diagnose, and solve problems efficiently in a SAN.

Many tools are available to collect the necessary data to perform problem determination and problem source identification (PD/PSI) in a SAN. Tools can offer the following capabilities:

- Monitor the SAN health
- Report failures
- Monitor and identify storage devices
- Monitor the fabric for failures or imminent bottlenecks
- Interpret message and error logs
- Send Simple Network Management Protocol (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, the configuration design information is understandable, available at any support level, and always updated to the latest configuration. Also, a database must exist where all of the following information is safely stored:

- Connections
- Naming conventions
- Device serial numbers
- Worldwide names (WWNs)
- Zoning
- System applications

A responsible person must be in charge of maintaining this infrastructure and monitoring the SAN's health status.

8.2 Management interfaces and protocols

We present the major protocols and interfaces that were developed to support management mechanisms.

8.2.1 Storage Networking Industry Association initiative

The Storage Networking Industry Association (SNIA) uses its Storage Management Initiative (SMI) to create and promote the adoption of a highly functional interoperable management interface for multivendor storage networking products. The SNIA strategic imperative is to manage all storage by the SMI interface. The adoption of this interface allows the focus to switch to the development of added value functionality. IBM is one of the industry vendors that is promoting the drive toward 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 that is called the *Common Information Model* (*CIM*) to enable interoperability. The web-based version of CIM (Web Based Enterprise Management (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 CIMOM and XML over HTTP to provide for the management of storage devices, enabling central management by using open standards.

The SNIA uses the xmlCIM protocol to describe storage management objects and their behavior. CIM allows management applications to communicate with devices by using object messaging that is 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, see this website:

http://www.snia.org

Open storage management with the Common Information Model

SAN management involves configuration, provisioning, logical volume assignment, zoning, and logical unit number (LUN) masking. Management also involves 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) and a graphical user interface (GUI) to support these kinds of administrative tasks. Many devices also provide proprietary application programming interfaces (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 various devices.

The CIM interface and the SMI-S object model, which is adopted by the SNIA, provide a standard model for accessing devices. This ability allows management applications and devices from various vendors to work with each other's products. This flexibility means that clients have more choice as to which devices work with their chosen management application, and which management applications they can use with their devices.

IBM embraces the concept of building open standards-based storage management solutions. IBM management applications are designed to work across multiple vendors' devices. Devices are being CIM-enabled to allow them to be controlled by other vendors' management applications.

Common Information Model object manager

The SMI-S standard designates that either a proxy or an embedded agent can be used to implement CIM. In each case, the CIM objects are supported by a *CIM object manager* (*CIMOM*). External applications communicate with CIM through 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 that are defined by the SMI-S standard. This standard allows storage management applications, regardless of vendor, to query status and perform command and control by using XML-based CIM interactions.

IBM developed its storage management solutions based on the CIMOM architecture (Figure 8-3).

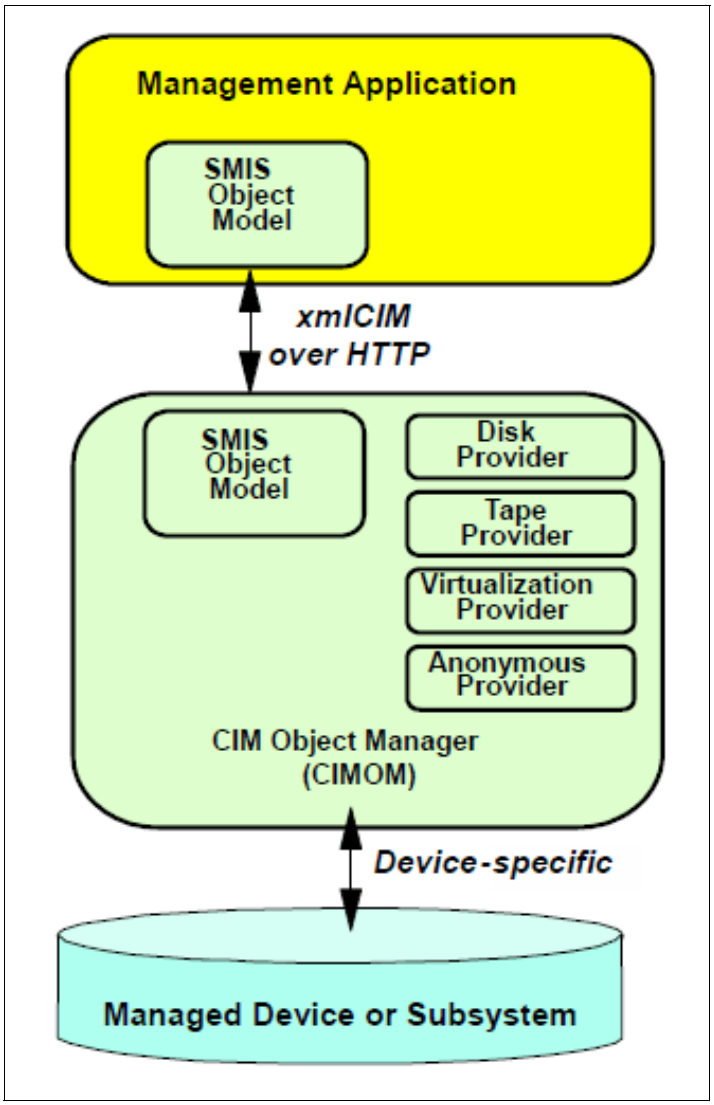


Figure 8-3 CIMOM component structure

8.2.2 Simple Network Management Protocol

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 SNMP agent is loaded on the managed devices. Management data is organized in a hierarchical data structure that is called the *Management Information Base* (*MIB*). The MIBs are defined and sanctioned by various industry associations.

The objective is for all vendors to create products in compliance with the MIBs so that inter-vendor interoperability at all levels can be achieved. If a vendor wants to include more device information that is not specified in a standard MIB, that additional information is specified through MIB extensions.

This protocol is widely supported by LAN/wide area network (WAN) routers, gateways, hubs, and switches. SNMP is the predominant protocol that is 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 an error condition. The device symbol, or icon, which is displayed on the SNMP manager console, can be changed to red or yellow, or any warning color, and messages can be sent to the network operator.

Out-of-band developments

SNMP MIBs are implemented for SAN fabric elements that allow out-of-band monitoring. The ANSI Fibre Channel Fabric Element MIB provides significant operational and configuration information about individual devices. The emerging Fibre Channel Management MIB provides more 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

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 located 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 the network host that supports a service. Rather, the user supplies the wanted type of service and a set of attributes that describe the service. Based on that description, the SLP resolves the network address of the service for the user.

SLP provides a dynamic configuration mechanism for applications in LANs. Applications are modeled as clients that need to locate servers that are attached to any of the available networks within an enterprise. For cases where many clients and services are available, the protocol is adapted to use the nearby Directory Agents that offer a centralized repository for advertised services.

8.2.4 Vendor-specific mechanisms

Many vendor-specific mechanisms are deployed by major SAN device providers.

Application programming interface

Many SAN devices are available from many vendors, and everyone has their own management and configuration software. In addition, most SAN devices can also be managed through a CLI over a standard Telnet connection, where an IP address is associated with the SAN device. Or, they can be managed by an RS-232 serial connection.

With multiple vendors and many management and configuration software tools, many products are available to evaluate, implement, and learn. In an ideal world, one product can manage and configure all of the functions and features on the SAN platform.

Application programming interfaces (*APIs*) are one way to help this simplification become a reality. Many vendors make the API of their product available for other vendors to make it possible for common management in the SAN. This openness allows the development of upper-level management applications that interact with multiple vendor devices and offer the system administrator a single view of the SAN infrastructure.

Common Agent Services

Common Agent Services is a component to provide a way to deploy agent code across multiple user machines or application servers throughout an enterprise. The agents collect data from and perform operations on managed resources for Fabric Manager.

The 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 that are 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 that are running the common agent software, which is referred to in this document as the *agent*.

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.

Common Agent Services consists of the following subcomponents:

- Agent manager

  The *agent manager* is the server component of the 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 that is 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 of the subagents to run within. It enables multiple management applications to share resources when managing a system.

- Resource manager

  Each product that uses Common Agent Services has its own resource manager and subagents. For example, IBM Tivoli Provisioning Manager has a resource manager and subagents for software distribution and software inventory scanning.

Figure 8-4 shows the Common Agent topology.

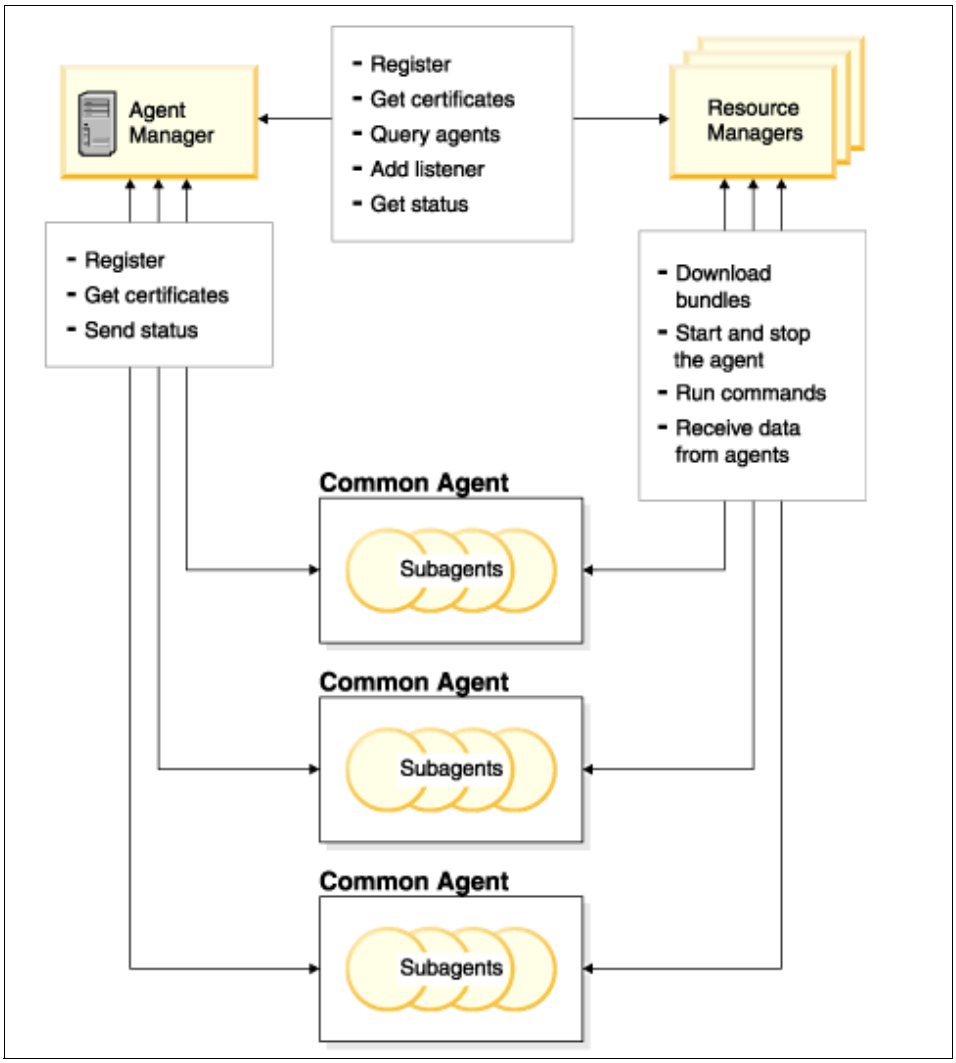


Figure 8-4 Common Agent Services

8.3 Management features

SAN management requirements are typified by having a common purpose, but they are implemented in different ways by the vendors. Certain vendors prefer to use web browser interfaces, other vendors prefer to use embedded agents, and still other vendors prefer to use the CLI. Many vendors use a combination of all of these interfaces. Typically, the selection of SAN components is based on a combination of the hardware and software functions, not on the ease of use of the management solution. The high-level features of any SAN management solution are likely to include most of the following benefits:

- Cost-effectiveness
- An open approach
- Device management
- Fabric management
- Proactive monitoring
- Fault isolation and troubleshooting
- Centralized management

- Remote management
- Adherence to standards
- Resource management
- Secure access
- Standards compliance

8.3.1 Operations

When we describe management, it automatically includes the operational aspects of the environment. The SAN administrators are responsible for all configuration of the SAN switches.

Typically, the initial design and creation of a SAN environment includes only a handful of servers and few storage systems. However, the environment grows and new technology needs to be added. At this stage, it tends to get more complicated. Therefore, it is necessary to ensure that comprehensive documentation exists that describes all aspects of the environment. And, the documentation needs to be reviewed regularly to ensure that it is current.

The following standards and guidelines need to be documented:

- Zoning standards:
  - How to create zones by using preferred practices
  - Naming standards that are 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 within the team.
- SAN and storage installation preferred practices: Agreed-to process to install and configure the equipment.
- SAN and storage software and firmware upgrade roadmaps:
  - High-level overview of how to ensure that the environment is kept current
  - Change schedules
- Monitoring and performance guidelines, such as defining the components, software, and processes that are monitored and explaining how to handle exceptions.

8.4 Vendor management applications

Each vendor in the IBM SAN portfolio provides their own bespoke applications to manage and monitor the SAN. In the following topics, we provide a high-level overview of each application.

8.4.1 Storage Networking SAN b-type

The IBM Storage Networking SAN *b-type* family switch management framework supports the widest range of solutions, from the small workgroup SANs up to large enterprise SANs. You can use the tools in the following sections with b-type SANs to centralize control and enable the automation of repetitive administrative tasks.

IBM Network Advisor

Under pressure to reduce costs and increase agility, organizations can become stuck in a reactive mode, overspending on inefficient solutions aimed at keeping their networks up and running. By standardizing and automating processes, however, network teams can proactively address availability and performance issues, and dramatically reduce costs.

To address this issue, the IBM Network Advisor software management tool provides comprehensive management for data, storage, and converged networks. This single application can deliver end-to-end visibility and insight across different network types; it supports Fibre Channel SANs—including Gen 5 and Gen 6 Fibre Channel platforms—IBM FICON, and IBM SAN b-type extension solutions. In addition, this tool supports comprehensive lifecycle management capabilities across different networks through a simple, seamless user experience. It is available both in SAN Professional Plus and SAN Enterprise options.

The IBM Network Advisor includes the following features:

- Deploys 20 years of storage networking best practices in one click to simplify the deployment of monitoring with predefined, threshold-based rules, actions, and policies
- Monitors Fibre Channel and Fibre Channel over IP (FCIP) health and performance indicators using customizable, browser-accessible dashboards
- Enables fast troubleshooting and problem isolation through comprehensive Fabric Vision diagnostics and dashboard playback
- Helps reduce costs by automating tasks across the network management lifecycle
- Provides unprecedented visibility and insight into the SAN fabric through integration with Fabric Vision technology
- Simplifies operations by providing centralized, end-to-end network management of SANs, including Gen 5 and Gen 6 Fibre Channel platforms, IBM FICON, and IBM SAN b-type extension environments
- Integrates seamlessly with hypervisors and management solutions from other vendors

For more information about IBM Network Advisor, see this website:

`http://www.ibm.com/systems/storage/san/b-type/na/`

Fabric Vision technology

Fabric Vision technology with IO Insight is an extension of Gen 6 Fibre Channel. It provides unprecedented insight and visibility across the storage network with powerful built-in monitoring, management, and diagnostic tools that enable organizations to simplify monitoring, increase operational stability, and dramatically reduce costs.

The benefits of IT virtualization, flash storage, and automation have allowed applications and services to be deployed faster while shattering performance barriers. However, due to the unprecedented growth of application and service interactions, complexity has increased in IT ecosystems. This has resulted in greater risk and instability for mission-critical operations and in access to critical data on storage.

To embrace high-density virtualization, flash storage, and cloud infrastructures, IT organizations need flexible storage networks that are both dynamic and high performing. Increased complexity and higher service level agreement (SLA) objectives mean that storage networks must respond with new tools. This will help to ensure nonstop operations, access to critical data, quick identification of potential congestion points, and maximized application performance—while simplifying administration.

Fabric Vision includes the following features:

- Provides built-in monitoring, management, and diagnostics to simplify storage administration, increase operational stability, and reduce costs
- Deploys 20 years of storage networking best practices in one click with predefined, threshold-based rules, actions, and policies
- Automatically detects degraded application or device performance with built-in device latency and input/output (I/O) performance monitors
- Reduces maintenance costs and network problems with proactive monitoring and advanced diagnostic tools
- Reduces capital expenses by eliminating the need for expensive third-party tools through built-in monitoring and diagnostics capabilities
- Utilizes dashboards that enable at-a-glance views of switch status and conditions contributing to performance issues
- Monitors individual storage devices to gain deeper insight into the performance of the network
- Enables tuning of device configurations with integrated I/O metrics to optimize storage performance
- Leverages predefined MAPS policies to automatically detect and alert to latency severity levels, and to identify slow-drain devices that might impact network performance
- Identifies, monitors, and analyzes specific application flows to simplify troubleshooting

For more information about Fabric Vision, see this website:

https://ibm.biz/BdH2nB

8.4.2 Cisco

Cisco Data Center Network Manager (DCNM) is a management system for the Cisco Unified Fabric. With Cisco DCNM, you can provision, monitor, and troubleshoot the data center network infrastructure. Cisco DCNM provides visibility and control of the unified data center so that you can optimize for the quality of service (QoS) that is required to meet service level agreements (SLAs).

Cisco DCNM increases overall data center infrastructure uptime and reliability, improving business continuity. It provides a robust framework and comprehensive feature set that meets the routing, switching, and storage administration needs of data centers. Cisco DCNM streamlines the provisioning for the unified fabric and monitors the SAN and LAN components. Cisco DCNM provides a high level of visibility and control through a single web-based management console for Cisco Nexus, Cisco MDS, and Cisco Unified Computing System products.

Cisco DCNM offers these capabilities:

- Configures and manages the fabric on multiple efficient levels
- Groups multiple SAN objects and SAN management functions intelligently 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 so that you can see the real-time state of all fabrics
- Provides the multi-fabric administration of secure Fabric OS SANs through a single encrypted console
- Monitors ISLs
- Manages switch licenses
- Performs fabric stamping
- Works across all Cisco Nexus and MDS switching families
- Supports automatic configuration for multi-tenant automation
- Offers integrated storage visualization, provisioning, and troubleshooting
- Integrates with Cisco UCS Director, vSphere, and OpenStack
- Data Center Network Manager offers intuitive, multi-fabric topology views for LAN fabric and storage. Supported overlays include:
  - VXLAN
  - FabricPath
  - Layer 2
  - Virtual port channel
  - Virtual device context
  - Virtual SAN
- Power On Auto Provisioning (POAP) with validated VXLAN templates for Nexus 5000 to 9000 switches
- Automatic provisioning for VXLAN, Fasbric Path, or VLAN-based multi-tenant networks using auto-configuration or top-down methods
- Multiple site visibility and search with data synchronization
- Image management, graceful insertion and removal, and software module updates
- Performance and storage zoning management

For more information about Cisco DCNM, see this website:

https://ibm.biz/BdH2G5

8.5 SAN multipathing software

In a well-designed SAN, your device is accessed by the host application over more than one path to potentially obtain better performance. This solution can also facilitate recovery if a controller, an adapter, a small form-factor pluggable (SFP), a cable, or a switch fails.

Multipathing software provides the SAN with an improved level of fault-tolerance and performance because it provides more than one physical path between the server and storage.

Traditionally, multipathing software was supplied by each vendor to support that vendor's storage arrays. The multipathing software often is embedded in the operating system. This approach offers a server-centric approach to multipathing that is independent of the storage array. This approach is often easier to implement from a testing and migration viewpoint.

Difference between storage and a SAN: It is important to understand the key difference between a SAN and storage, although sometimes they are referred to as one.

Storage is where you keep your data.

SAN is the network that the data travels through between your server and storage.

Figure 8-5 shows an example of a dual-fabric environment where the hosts use multipathing software and can access the storage if a path fails, or if a fabric fails.

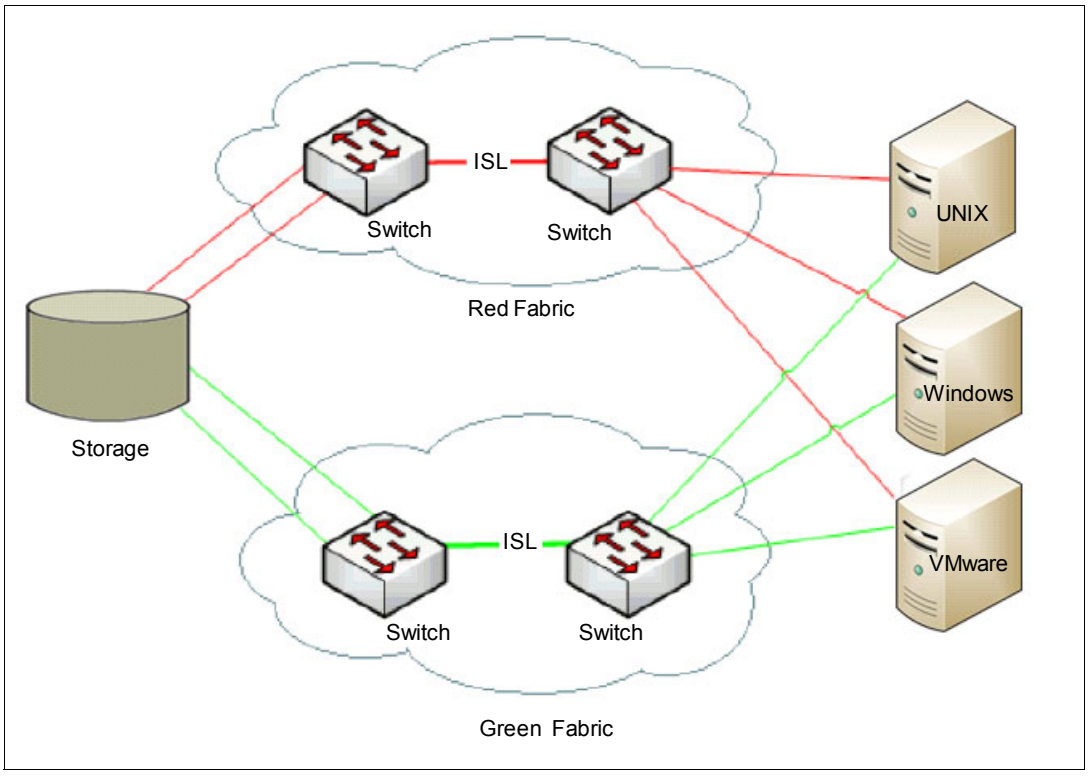


Figure 8-5 SAN overview

The IBM multipathing software is *IBM Subsystem Device Driver* (*SDD*), which offers the following benefits:

- Enhanced data availability
- Dynamic I/O load-balancing across multiple paths
- Automatic path failover protection
- Concurrent download of licensed machine code

When you determine the number of paths to configure to each volume, never exceed the level that is supported by the storage device. When you implement zoning to a storage device, you must decide the number of paths.

For detailed information about multipath drivers for IBM storage, see the following website:

`http://www.ibm.com/support/docview.wss?rs=540&context=ST52G7&q=ssg1*&uid=ssg1S7000303&loc=en_US&cs`

We provide a multipathing example in separate scenarios. In Figure 8-6, the servers are connected to the SAN with two HBAs. The HBAs access the volumes through two storage ports on the storage device. This access is controlled by zoning, which provides four working paths for each server to their volumes: two from the Red Fabric and two from the Green Fabric.

Figure 8-6 shows a single path failure, which is indicated by the STOP sign.

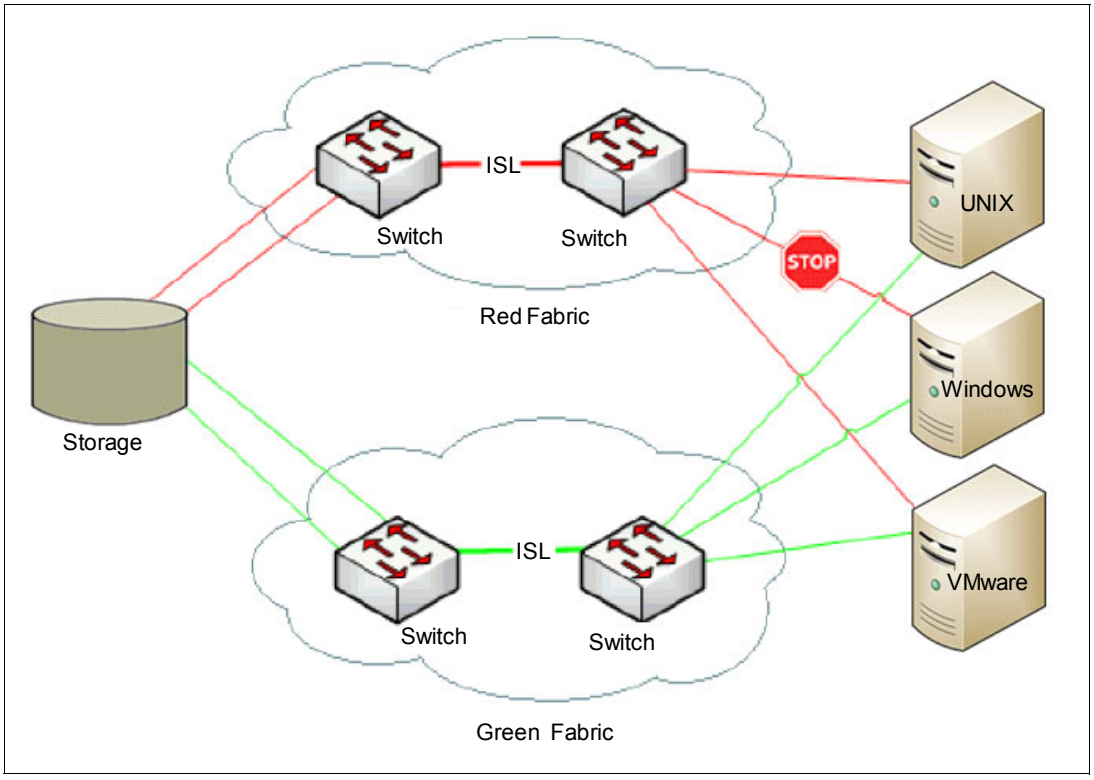


Figure 8-6 HBA failure in a single server

In Figure 8-6, the Windows server lost connectivity to the SAN and cannot access the Red Fabric. However, the Windows server has working paths through the Green Fabric. All other servers are running without any issues.

Figure 8-7 shows that a switch in the Red Fabric failed.

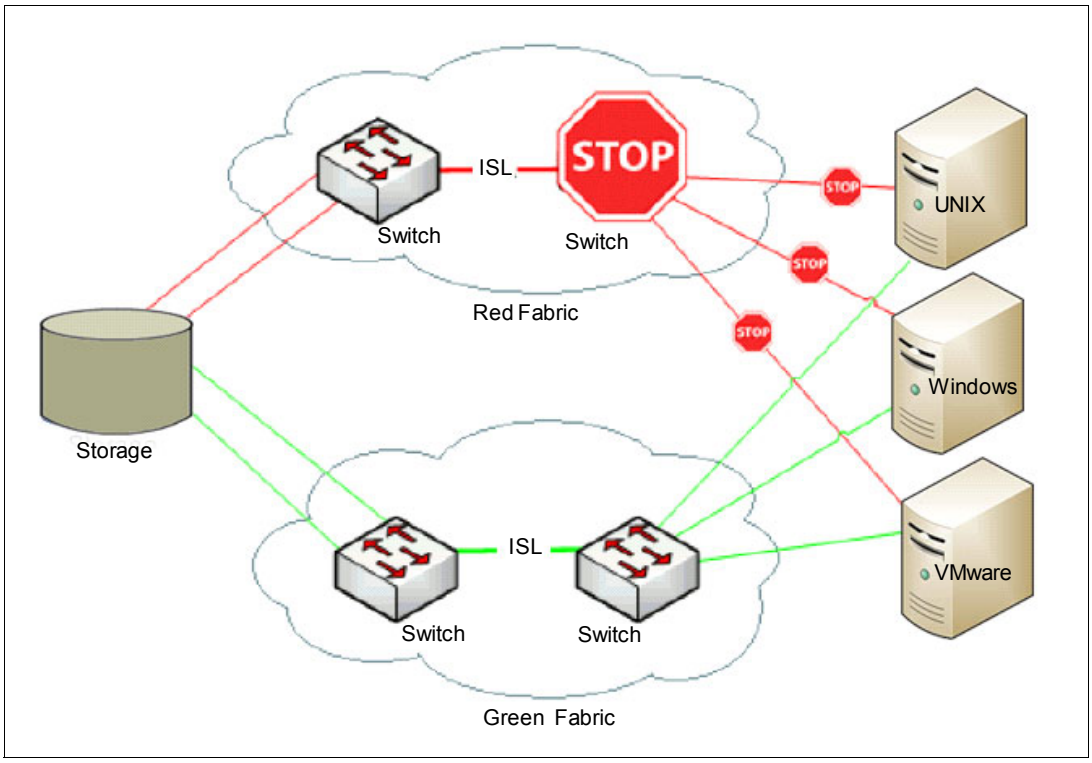


Figure 8-7 A non-functional switch affects all attached devices

In Figure 8-7, our servers cannot access a switch. Working paths still exist from the servers through the Green Fabric.

Figure 8-8 shows that a link from the storage device to a switch failed in the Red Fabric.

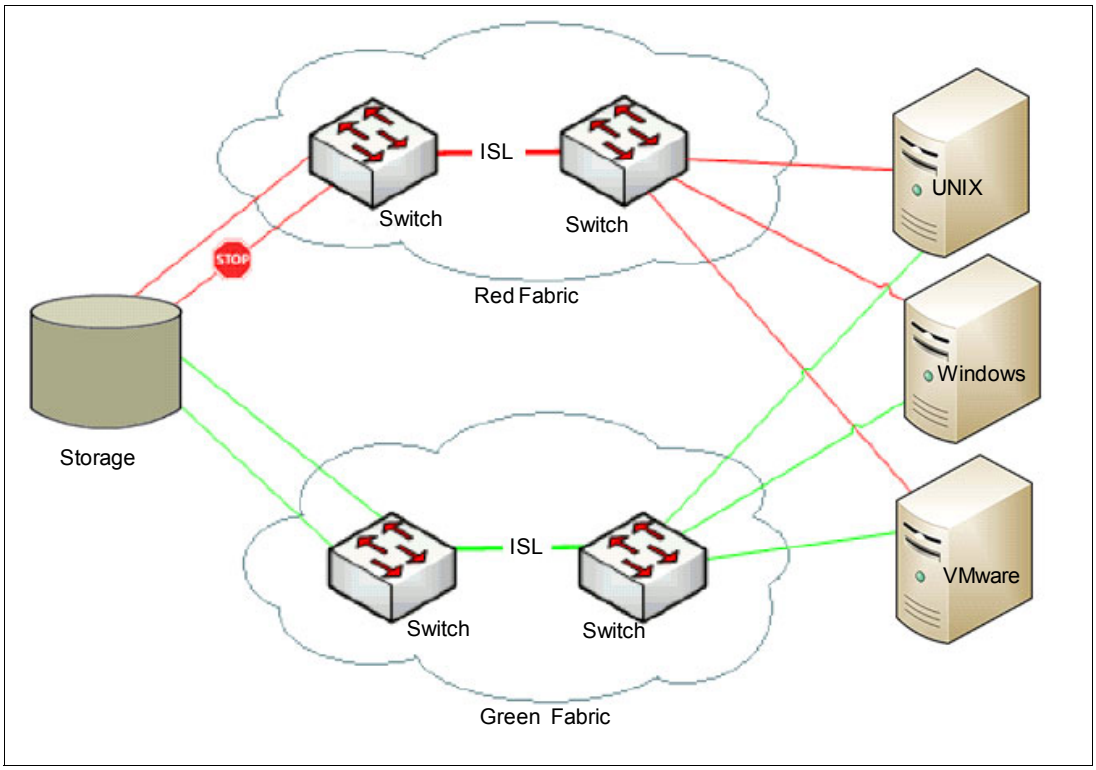


Figure 8-8 Storage device with a single failed connection to a switch

In Figure 8-8, the storage device lost one of four connections. One connection to the Red Fabric does not function. Therefore, all servers that use the same storage port now see three working paths out of four possible paths. All servers that are zoned to the failed storage port are affected.

Figure 8-9 shows the storage device lose access to the Red Fabric.

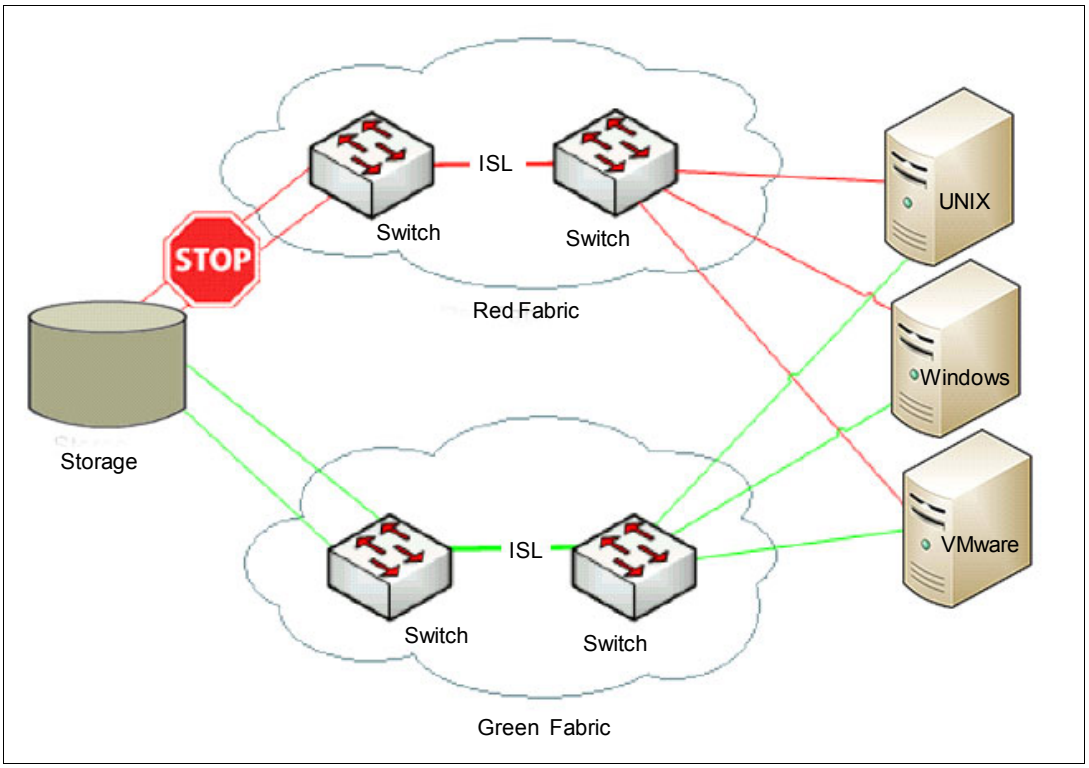


Figure 8-9 Storage device that lost two out of the four connections to the SAN

In Figure 8-9, our storage device lost access to the Red Fabric. All devices in the Red Fabric are running normally. Only two specific storage ports failed. Our servers have only two working paths through the Green Fabric. This configuration affects all servers that are zoned to these storage ports.

Figure 8-10 shows the storage device offline.

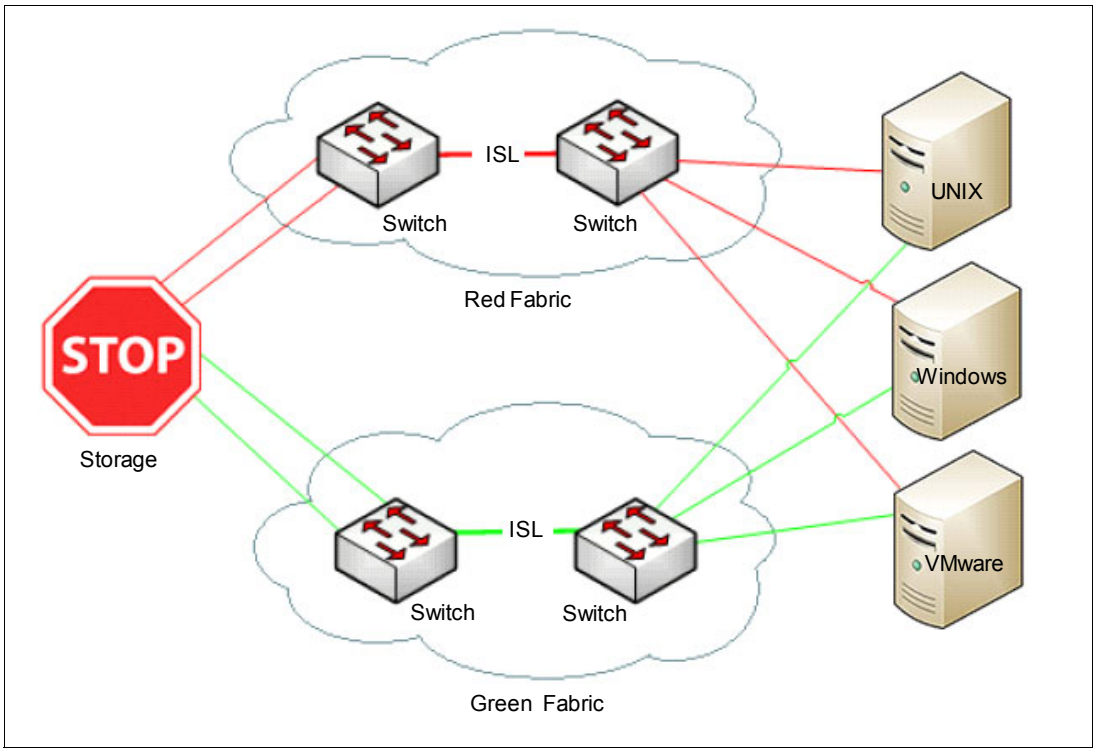


Figure 8-10 The storage device is offline and no connections work

In Figure 8-10, we lost our storage device. No paths to any volumes on this device are available. No data is accessible. All servers that are zoned to this storage device are severely affected and cannot access this storage device.

If we install a supported version of a multipath driver correctly on all of the servers, we survive all scenarios except the last scenario, in which only a minimum impact occurs.

9

Security

In this chapter, we provide an overview of the need for security. We describe the available techniques and several key points.

9.1 Security in the storage area network

Security is always a major concern for networked systems administrators and users. Even for specialized networked infrastructures, such as a storage area network (SAN), special care must be taken so that information does not get corrupted, either accidentally or deliberately, or fall into the wrong hands. And, we also must ensure that the correct security is in place at a fabric level, for example, to ensure that a user does not inadvertently change the configuration incorrectly.

Now that SANs no longer fit the traditional direct-attached storage concept where storage is cabled directly to the server, the inherent security of that design is lost. The SAN and its resources might be shared by many users and many departments. The SAN might be shared by different operating systems with 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 consideration, which is called *zoning*.

The fabric enforces the separation of data so that only the intended users can access their data.

Zoning, however, does not provide security in that sense; it implements the means of segregation (isolation) only. The real security issue is the vulnerability when the data must travel outside of the data center and over long distances. This type of travel often involves transmission over networks that are owned by different carriers.

We must look at security from two angles: for data-in-flight, as explained in 9.4.2, “Data-in-flight” on page 200, and for data-at-rest, as explained in 9.4.3, “Data-at-rest” on page 201.

More often than not, data is not encrypted when it is sent from the source to a target. Therefore, any information is readable with the correct tools, even though it is slightly more complicated than simply eavesdropping on a telephone line. Because all of the data is sent at a block level with the Fibre Channel Protocol (which means that all data that is sent is squeezed into the Fibre Channel frame before the data is sent), “sniffing” a frame or two might give you 2112 bytes of data. For an example of the difficulty, this amount is similar to 1/333.000 of a normal CD or 13 milliseconds of a CD that spans 74 minutes. Obviously, this comparison does not give you much information without putting it in the correct context.

Security is more of a concern if the whole Fibre Channel port or disk volumes and arrays are mirrored, or tapes that contain information end up in the wrong hands. However, tampering with information from a SAN is trivial. It takes a concerted effort.

The storage architect and administrators must understand that in a SAN environment, often with a combination of diverse operating systems and vendor storage devices, a combination of technologies is required. This mixture ensures that the SAN is secure from access from unauthorized systems and users, whether accidental or deliberate.

We briefly explore the technologies and their associated methodologies that can be used to ensure data integrity, and to protect and manage the fabric. Each technology offers advantages and disadvantages. You must consider each technology based on a carefully thought-out SAN security strategy that is 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”. The same concept applies to computer security. There is no point in locking all of 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 (SPOF).

The levels of defense must be complementary, and work with each other. If you have a SAN, or any other network, that crumbles after a single penetration, this level of defense is insufficient.

Many unique entities must be considered in any environment. We describe several of the most important entities.

9.2.1 Access control

Access control can be performed with authentication and authorization techniques:

Authentication The secure system must challenge the user (typically with a password) so that this user is identified.

Authorization After the system identifies a user, the system knows what this user is allowed to access and what they are not allowed to access.

As in any IT environment, including SAN, access to information and to the configuration or management tools must be restricted. Access must be granted to only those individuals that need access and that are authorized to make changes. Any configuration or management software is typically protected with several levels of security. Levels usually start with a user ID and password that must be assigned to personnel based on their skill level and responsibility.

9.2.2 Auditing and accounting

An audit trail must be maintained for auditing and troubleshooting, especially when you create a *root cause analysis* (*RCA*) after an incident occurs. Inspect and archive logs regularly.

9.2.3 Data security

Whether we describe data-at-rest or data-in-flight, data security consists of data confidentiality and integrity:

Data confidentiality The system must guarantee that the information cannot be accessed by unauthorized people, that it remains confidential, and that it is only available for authorized personnel. Confidentiality is typically accomplished by using data *encryption*.

Data integrity The system must guarantee that the data is stored or processed within its boundaries and that it is not altered or tampered with in any way.

The data security and integrity requirement aims to guarantee that data from one application or system does not become overlaid, corrupted, or otherwise destroyed. This requirement applies whether data is intentionally destroyed or destroyed by accident, either by other applications or systems. This requirement might involve a form of authorization, and the ability to fence off the data in one system from another system.

This data security necessity must be balanced with the requirement for the expansion of SANs to enterprise-wide environments, with an emphasis on multiple platform connectivity. True cross-platform data sharing solutions, as opposed to data partitioning solutions, are also a requirement. Security and access control also must be improved to guarantee data integrity.

We overview several common data security approaches for the SAN environment. This list is not meant to be an in-depth description. It is merely an attempt to acquaint you with the technology and terminology that you are likely to encounter in a discussion about SAN security.

9.2.4 Securing a fabric

Several 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 host bus adapters (HBAs). The joint initiative between Brocade and Emulex establishes Fibre Channel Authentication Protocol (FCAP) as the next-generation implementation of SLAP. Clients gain the assurance that a region of trust extends over the entire domain.

FCAP was incorporated into its fabric switch architecture and 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 that consists of switches and HBAs.

The fabric authorization database is a list of the worldwide names (WWNs) and associated information, such as 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 allow 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, SANs did not have any zoning. It was an any-to-any communication. No real access control mechanism protected storage that was used by one host from being accessed by another host. When SANs grew, this drawback became a security risk as SANs became more complex and ran more vital parts 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 that is stored on the server. Persistent binding is implemented through the HBA driver of the server. This 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 Small Computer System Interface (SCSI) target IDs and LUNs.

Logical unit number masking

One approach to securing storage devices from hosts that want to take over already assigned resources is logical unit number (LUN) masking. Every storage device offers its resources to the hosts with LUNs. For example, each partition in the storage server has its own LUN. If the host (server) wants to access the storage, it must 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 the hosts that can access a specific LUN with the storage device control program. Whenever the host accesses a particular LUN, the storage device checks its access list for that LUN. The device allows or disallows the host to gain 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 specific port that does not allow any other device to plug into the port.

Role-based access control

A *role-based access control* (*RBAC*) feature is available in most SAN devices. By using RBAC, you can control user access and user authority simply. With RBAC, you can provide users with access or permission to run tasks that are within their skill set or job role only.

Typically, RBAC has three definitions:

- Role assignment
- Role authorization
- Permission authorization

Each role can contain multiple users, and each user can be part of multiple roles. For example, if role1 users are allowed access to configuration commands only, and role2 users are allowed access to debug commands only, if John belongs to both role1 and role2, he can access configuration and debug commands.

These predefined roles in a SAN environment are important to ensure that the correct login and access is defined for each user.

9.2.5 Zoning, masking, and binding

Although zoning, masking, or binding are not classified as security products or mechanisms, combining all of their functionality can increase the security of the SAN.

9.3 Data security

These data security standards are intended to secure Fibre Channel (FC) traffic between all FC ports and the domain controller.

The following methods are used for data security standards:

- Fibre Channel Password Authentication Protocol (FCPAP) refers to Secure Remote Password Protocol (SRP), Request for Comments (RFC) 2945.
- Diffie Hellman - Challenge Handshake Authentication Protocol (DH-CHAP) refers to Challenge Handshake Authentication Protocol (CHAP), RFC 1994.
- Fibre Channel Security (FCSec) refers to Internet Protocol (IP) Security (IPSec), RFC 2406.

  The focus of the FCSec is to provide authentication of the following entities:

  – Node-to-node
  – Node-to-switch
  – Switch-to-switch

An additional function that might 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 authenticated correctly are allowed to join the fabric.

Authentication in the secure fabric is twofold. The fabric wants to verify the identity of each new switch before it joins the fabric, and the switch that wants to join the fabric wants to verify that it is connected to the correct fabric. Each switch needs a list of the WWNs of the switches that are authorized to join the fabric. The switch also needs a set of parameters that are used to verify the identities of the other switches that belong to the fabric.

Manual configuration of this information within all of the switches of the fabric is possible, but not advisable in larger fabrics. And, you need a mechanism to manage and distribute information about authorization and authentication across the fabric.

9.4 Storage area network encryption

What is data encryption, and symmetric and asymmetric encryption? What is in-flight data or data-at-rest? This terminology is explained to help you to understand the fundamentals in encryption and key management.

9.4.1 Basic encryption definition

You must first determine whether you need encryption. We describe basic encryption, cryptographic terms, and ideas about 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 was 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.

Two major types of encryption are available:

Symmetric The same secret password, or *key*, is used to encrypt a message and decrypt the corresponding cipher text.

Asymmetric One key is used to encrypt a message, and another key is used to decrypt the corresponding cipher text. Asymmetric encryption is also called *public-key encryption*.

A *symmetric cryptosystem* follows a fairly straightforward philosophy: Two parties can securely communicate if both parties use the same *cryptographic algorithm* and possess the same secret key to encrypt and decrypt messages. This algorithm is the simplest and most efficient way of implementing secure communication, if the participating parties are able to securely exchange secret keys (or passwords).

Figure 9-1 shows symmetric encryption.

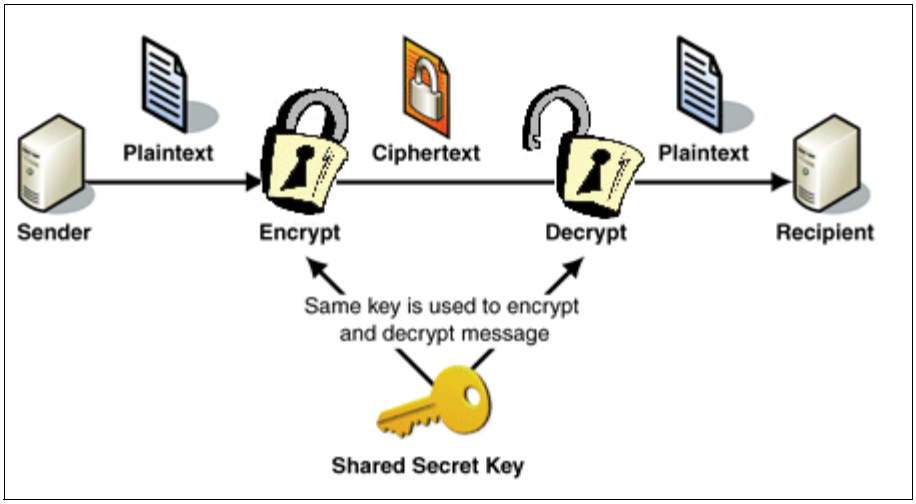


Figure 9-1 Symmetric cryptography

An *asymmetric* (or *public-key*) *cryptosystem* is a cryptographic system that uses a pair of unique keys that are typically referred to as *public keys* and *private keys*. Each individual is assigned a pair of these keys to encrypt and decrypt information. A message that is encrypted by one of these keys can be decrypted only 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 they encrypt information that is sent to an individual. For example, people can use a person's public key to encrypt information that they want to send to that person. Similarly, people can use the user's public key to decrypt information that is sent by that person.
- The other key is called a *private key* because it is accessible only to its owner. The individual can use the private key to decrypt any messages that are encrypted with the public key. Similarly, the individual can use the private key to encrypt messages so that the messages can be decrypted only with the corresponding public key.

Therefore, 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 who knows the street address can go to the mailbox 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 shows the asymmetric cryptography process.

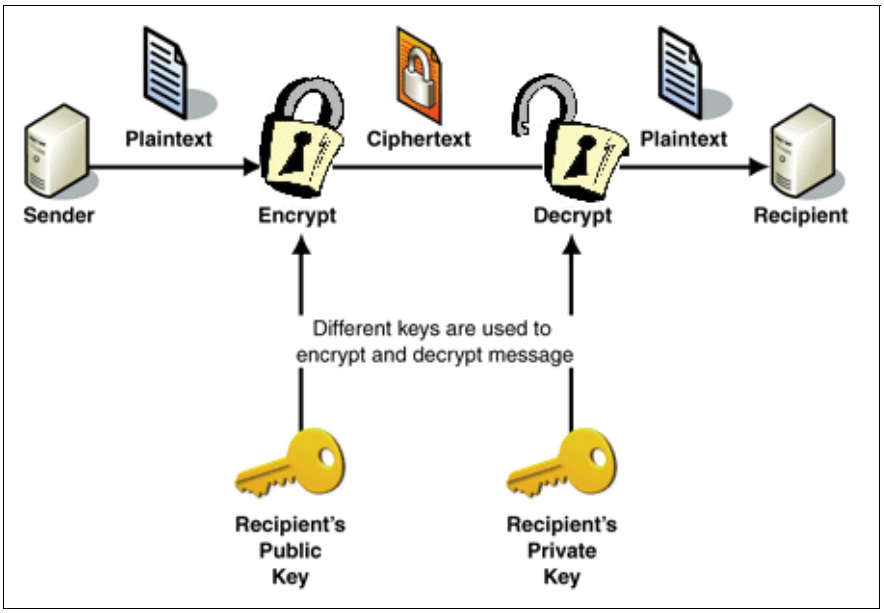


Figure 9-2 Asymmetric cryptography

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 a *session key*.

However, an issue still exists about public-key cryptosystems. When you receive a public key from someone for the first time, how do you know that this individual is really who they claim to be? If "spoofing" the identity of someone 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 that is issued by a trusted institution that vouches for the identity and key ownership of an individual. The certificate guarantees authenticity and integrity.

In the next sections, we present several common encryption algorithms and tools and explain the terminology.

9.4.2 Data-in-flight

Also known as *data-in-motion*, this term generically refers to protecting information any time that 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 use technologies, such as Secure Sockets Layer (SSL), Virtual Private Network (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, an 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 protecting *data-at-rest*.

If encryption is used as part of the strategy for the protection of data-at-rest, this protection also indirectly addresses the issue of displayed tape media. This issue is addressed because, even if tapes fall into the wrong hands, the data that is stored on them is unreadable without the correct key. These security measures assume that you enacted the appropriate key management techniques.

To gain the needed security level, you build layers of security on your SAN. You first increase the level of difficulty for an unauthorized user to even gain access to the data. You 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 receive a public key from someone for the first time, how do you know that this individual is the correct person? 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 that is issued by a trusted institution that vouches for the identity and key ownership of an individual. A digital certificate guarantees authenticity and integrity.

Trusted institutions all over the world generate trusted certificates. We use this type of mechanism also for the first time by using a certificate that is generated by our switch. For more information, see 9.4.6, “Key management considerations and security standards” on page 202.

9.4.5 Encryption algorithm

After you decide that encryption is required, you must also be aware that several encryption schemes are available to choose from. The most popular encryption schemes in use today include the following algorithms:

- Triple Data Encryption Standard (3DES)
- Data Encryption Standard (DES)
- Advanced Encryption Standard (AES)
- Rivest-Shamir-Adleman algorithm (RSA)
- Elliptic curve cryptography (ECC)
- Diffie-Hellman
- Digital signature algorithm (DSA)
- Secure Hash Algorithm (SHA)

For more information about IBM System Storage Data Encryption, see *IBM System Storage Data Encryption*, SG24-7797. For an example of how IBM implements encryption on the IBM System Storage SAN Volume Controller, see *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, were focused on establishing a perimeter of defense around system assets. Although 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 clients, partners, employees, suppliers, and their data. Although open networks offer significant advantages, they raise concerns about how a business protects 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, many of these concerns can be mitigated.

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 IBM Security Key Lifecycle Manager.

IBM Security Key Lifecycle Manager

Because of the nature, security, and accessibility of encryption, encrypted data depends 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 so that the unauthorized agent also can 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 that is contained in the ciphertext is said to be cryptographically erased. If the only copies of certain data are cryptographically erased, access to that data is permanently lost for all practical purposes.

This problem 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 and data is kept from others that do not have sufficient authority. At the same time, data must be accessible to any agent that has both the authority and the requirement to gain access.

Two security 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 of the information that is 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 redundancy ensures that the backup of each key server's data is maintained. Failure of any one key server or any one network does not prevent devices from obtaining access to the data keys that are 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 result 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 that are associated with key management for encrypted devices. However, you must still be sufficiently aware of how these products interact to provide the correct management of the computer environment.

Master key: Even with a key server, generally at least one encryption key, which is normally called the *master key* (*MK*), must be maintained manually. For example, this master key manages access to all other encryption keys. This master key encrypts the data that is used by the key server to exchange keys.

Fundamentally, IBM Security Key Lifecycle Manager works by allowing administrators to connect to storage devices and then create and manage *keystores*. These stores are secure repositories of keys and certificate information that are used to encrypt and decrypt data, or to use existing keystores already in place.

Over the course of the key lifecycle, all management functions, including creation, importation, distribution, backup, and archiving, are easily accomplished. These functions can be performed by using the lifecycle manager's graphic interface, which can be accessed by using any standard browser in the network.

IBM Security Key Lifecycle Manager therefore serves as a central point of control, unifying key management even when different classes of storage devices are involved. For more information about IBM Security Key Lifecycle Manager, see this website:

http://www.ibm.com/software/products/en/key-lifecycle-manager

Two security standards are 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 (EALs), 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

In-flight encryption of expansion port (E_port) links was introduced with FOS 7.0 and Gen5 FC 16 Gbps technology.

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 inter-switch link (ISL) between two b-type switches, and then to be decrypted or extracted at the ingress point of the ISL. This feature uses port-based encryption and compression. It is supported on 16 Gbps E_ports only.

Note: The ports can run at 2 Gbps, 4 Gbps, 8 Gbps, 10 Gbps, or 16 Gbps.

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 distances. An average compression ratio of 2:1 is provided. Frames are never left in an encrypted or compressed state when they are delivered to an end device, and both ends of the ISL must terminate at 16 Gbps ports.

For more information, see the *Metro Cloud Connectivity: Integrated Metro SAN Connectivity in Gen 5 Fibre Channel Switches* white paper:

http://www.brocade.com/content/dam/common/documents/content-types/whitepaper/brocade-metro-cloud-connect-wp.pdf

Figure 9-3 shows the b-type in-flight encryption architecture.

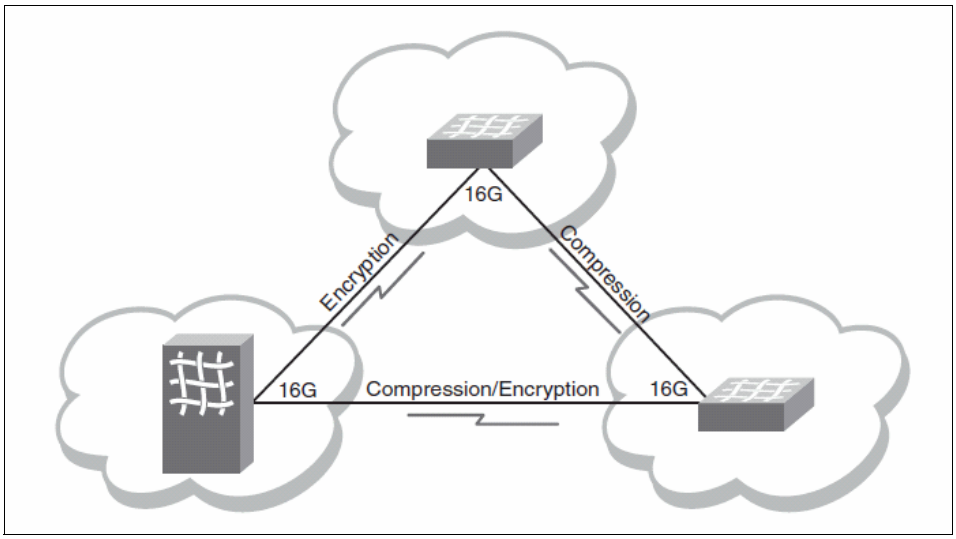


Figure 9-3 In-flight architecture

Encryption at rest

The b-type fabric-based encryption solutions work transparently with heterogeneous servers, tape libraries, and storage subsystems. Although host-based encryption works only for a specified operating system and storage-based encryption works only for a specific vendor, b-type products are deployed in the core of the fabric to encrypt Fibre Channel-based traffic. Users deploy b-type encryption solutions through either the FS8-18 Encryption Blade or the 2U, rack-mounted IBM SAN32B-E4 Encryption Switch.

The *Device Encryption Key* (*DEK*) is important. Because the DEK is needed to encrypt and decrypt the data, it must be random and 256 bits. B-type encryption devices use a True Random Number Generator (TRNG) to generate each DEK. For encrypting data that is destined for a disk drive, one DEK is associated with one logical unit number (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 unencrypted data. Therefore, the b-type encryption device can encrypt the data, block by block, without expanding the size of the data. The key management is performed by using external software, such as IBM Security Key Lifecycle Manager.

Figure 9-4 shows a simple b-type encryption setup.

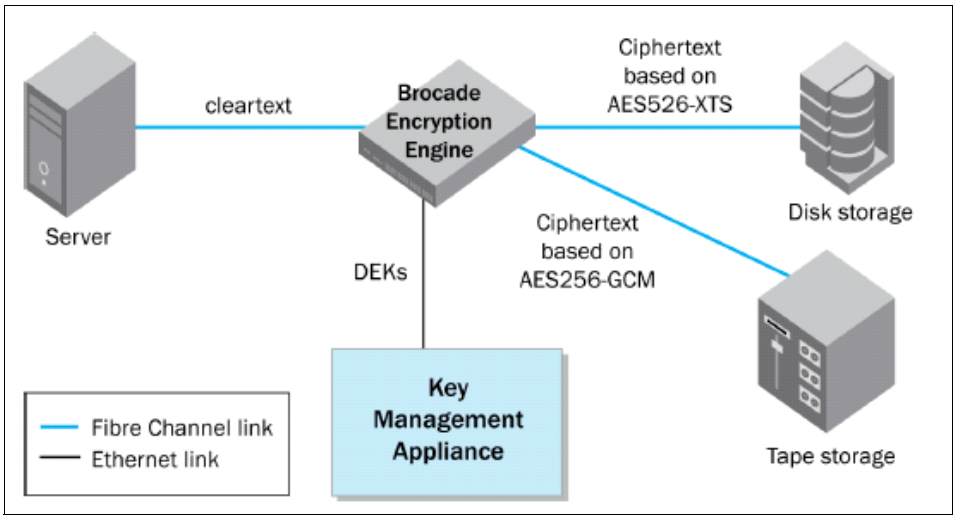


Figure 9-4 b-type encryption and key management

9.4.8 Cisco encryption methods

Cisco has two methods of encrypting SAN information: in-flight encryption and storage media encryption. For more information about both of these methods, see the following websites:

- http://www.cisco.com/en/US/prod/collateral/ps4159/ps6409/ps5990/white_paper_c11-545124.html
- 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. It uses the existing FC-SP architecture. Fibre Channel data that travels 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 passed between the two peers.

Encryption is performed at line rate by encapsulating frames at egress with encryption by using the GCM authentication mode with 128-bit AES encryption. At ingress, frames are decrypted and authenticated with integrity checks.

Two primary use cases for Cisco TrustSec Fibre Channel Link Encryption exist. In the first use case, clients 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 clients, such as defense and intelligence services.

Figure 9-5 shows Cisco TrustSec Fibre Channel Link Encryption.

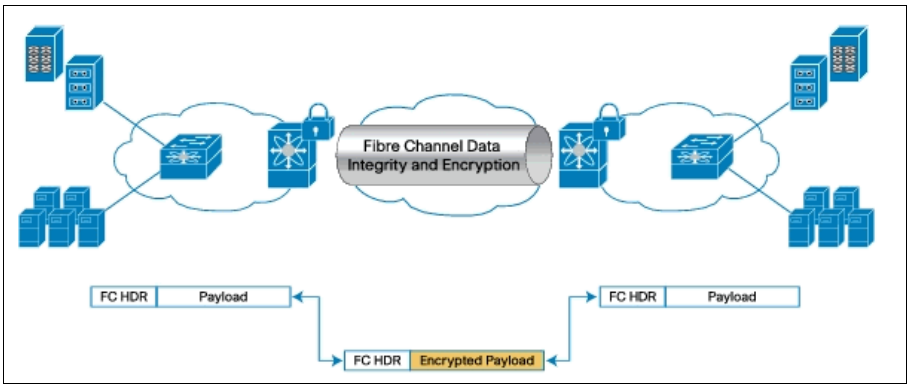


Figure 9-5 Cisco TrustSec encryption

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 by 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-site and multiple-site environments. Provisioning and key management for Cisco SME are both integrated into Cisco Fabric Manager and Data Center Network Manager (DCNM). No additional software is required for key management.

Figure 9-6 shows the SME architecture.

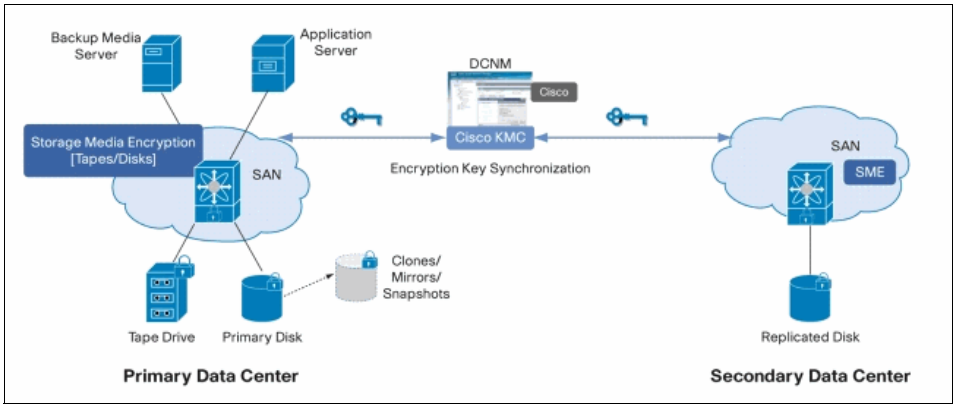


Figure 9-6 SME architecture

9.5 Encryption standards and algorithms

The following encryption algorithms are in use today:

- Advanced Encryption Standard (AES)

  AES is a symmetric 128-bit block data encryption technique that was developed by Belgian cryptographers: Joan Daemen and Vincent Rijmen. The US government adopted the algorithm as its encryption technique in October 2000, replacing the DES encryption that it used. AES works at multiple network layers simultaneously. The National Institute of Standards and Technology (NIST) of the US 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 that is approved by the National Security Agency (NSA) for top secret information.

- Rivest-Shamir-Adleman algorithm (RSA)

  The RSA algorithm involves three steps: key generation, encryption, and decryption. This algorithm was created in 1977 by Ron Rivest, Adi Shamir, and Len Adleman at MIT. The letters RSA are the initials of their surnames. It was the first algorithm that is known to be suitable for digital signing and data encryption, and one of the first great advances in public key cryptography. RSA is still widely used in electronic commerce protocols, and it is believed to be secure with sufficiently long keys and the use of up-to-date implementations.

- Elliptic curve cryptography (ECC)

  ECC is an approach to public-key cryptography that is 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. An example of this algorithm is Lenstra elliptic curve factorization, but this use of elliptic curves is *not* usually referred to as elliptic curve cryptography.

- Diffie-Hellman (D-H)

  The D-H key exchange is a cryptographic protocol that 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 by using a symmetric key cipher.

- Digital Signature Algorithm (DSA)

  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, and 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 US Patent 5,231,668, which was filed 26 July 1991, and attributed to David W. Kravitz, a former NSA employee.

- Secure Hash Algorithm (SHA)

  The 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 Transport Layer Security (TLS), SSL, Pretty Good Privacy (PGP), Secure Shell (SSH), secure/MIME (S/MIME), and IPSec. The algorithm was also used on the Nintendo Wii gaming console for signature verification when booting occurs.

9.6 Security common practices

At a high level, consider implementing the following security preferred practices at a minimum:

- Change default configurations and passwords often.
- Check and double-check configuration changes to ensure that only the data that is supposed to be accessed can be accessed.
- Ensure that the management of devices usually takes a *Telnet* form, with the use of encrypted management protocols.
- Ensure that the network is secure because remote access often relies on unsecured networks. Ensure that a form of protection is in place to guarantee that only those people with the correct authority are allowed to connect.
- Ensure that the operating systems that are connected are as secure as they can be. If the operating systems are connected to an internal and external LAN, ensure that this connection cannot be used. Access might be gained by using loose configurations.
- Assign the correct roles to administrators.
- Ensure that the devices are in physically secure locations.
- Ensure that the passwords are changed if the administrator leaves. Also, ensure that passwords are changed regularly.

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 safeguards do not guarantee that your information is 100% secure, but they can go far in ensuring that all but the most ardent “thieves” are kept out.

10

Solutions

The added value of a storage area network (SAN) lies in the use of its technology to provide tangible and desirable benefits to the business. These benefits are provided by the use of fast, secure, reliable, and highly available networking solutions. Benefits range from increased availability and flexibility to more functionality that can reduce application downtime.

In this chapter, we provide a description of general SAN applications, and the types of components that are required to implement them. Far more complexity exists than is presented here. For instance, this text does not cover how to choose one switch over another, or how many inter-switch links (ISLs) are necessary for a specific SAN design. These strategic decisions must be always considered by experienced IT architects, and these decisions are beyond the intended scope of this book. We introduce the basic principles and key considerations to choose an optimal solution for SAN deployments.

10.1 Introduction

With the continued development of communication and computing technologies and products, SANs are becoming much more complex. We are not referring to merely a simple fiber-optic connection between SAN devices. Examples of these devices include SAN switches, routers, tape drives, disk device subsystems, and target host systems that use standard Fibre Channel host bus adapters (HBAs). Technology moved beyond those solutions and continues to do so.

Today, businesses are looking for solutions that enable them to increase the data transfer rate within the most complex data centers. Businesses also want solutions that provide high availability of managed applications and systems, implement data security, and provide storage efficiency. At the same time, businesses want to reduce the associated costs and power consumption.

Organizations must find a smooth, effective, and cost-efficient way to upgrade their current or traditional SAN infrastructure. The upgraded infrastructure provides a less complex and more powerful and flexible data center of the next generation.

SAN solutions can be classified into many categories. We chose to classify our SAN solutions as infrastructure simplification, business continuity, and information lifecycle management (ILM). In the following topics, we describe the use of basic SAN design patterns to build solutions for different requirements. These requirements range 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 SAN solutions and requirements are described, we present basic principles to consider when you plan a SAN implementation or upgrade.

10.2 Basic solution principles

Many important decisions must be made by the system architect, either when a new SAN is designed, or when an existing SAN is expanded. These decisions typically refer to the choice of the connectivity technology, the preferred practices for adding capacity to a SAN, or the most suitable technology for achieving data integration.

10.2.1 Connectivity

Connecting servers to storage devices through a SAN fabric is often the first step in a phased SAN implementation. Fibre Channel attachments offer the following benefits:

- Extended connection distances (sometimes called *remote storage*)
- Enhanced addressability
- Improved performance by running Small Computer System Interface (SCSI) over Fibre Channel

Many implementations of Fibre Channel technology are simple configurations that remove many of the restrictions of the existing storage environments. With these implementations of Fibre Channel technology, you can build one common physical infrastructure. The SAN uses common cabling to the storage and other peripheral devices.

The handling of separate sets of cables, such as Original Equipment Manufacturer's Information (OEMI), Enterprise Systems Connection (ESCON), SCSI single-ended, SCSI differential, and SCSI Low Voltage Differential (LVD), caused IT organization management much difficulty as it attempted to treat each of these types differently. One of the biggest issues 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, only a few 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 when compared to a 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 reap the benefits of Fibre Channel technology, while also protecting their current storage investments.

The flexibility and simplification of the SAN infrastructure can be dramatically improved by using Fibre Channel over Ethernet (FCoE), which evolved over the last few years. FCoE can easily replace dedicated switching solutions for LAN and SAN with a single device that can transfer both types of data: Internet Protocol (IP) packets and storage data. We call these deployments *converged networks*. In the following topics, we briefly present the basic migration steps to convergency.

10.2.2 Adding capacity

The addition of storage capacity to one or more servers might be facilitated while the device is connected to a SAN. Depending on the SAN configuration and the server operating system, it might 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 primitive* (*LIP*) might affect the operation of other devices on the loop. This setback might be overcome by slowing down the operating system activity to all of the devices on that particular loop before you attach the new device. This setback is far less of a problem with the latest generation of loop-capable switches. If the storage devices attach to a SAN by a switch, the use of the switch and management software makes it is possible to make the devices available to any system that connects to the SAN.

10.2.3 Data movement and copy

Data movement solutions require that data is 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. 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. The data travels one time from the 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. This practice frees up server processor cycles and LAN or WAN bandwidth. Today, this data replication can be accomplished in a SAN by using intelligent tools and utilities and between data centers that use, for example, FCoE protocol on a WAN.

The following sections list several of the available copy services functions.

Data migration

One of the critical tasks for a SAN administrator is to move data between two independent SAN infrastructures. The administrator might move data from an old storage system that is being discontinued to the new enterprise and high-performance disk system. Two basic scenarios exist. SANs are independent and cannot be interconnected even if they reside in the same data center, and the disk systems can be cross-connected through SAN switches.

Data replication over storage area networks

In this scenario, we can interconnect both storage devices (both SANs) together and migrate data directly from an old device to the new storage box. This step is completed without interruption of the service or performance degradation of the application or host server. This type of migration is referred to as a *block-level data copy*. In this type of migration, storage systems do not analyze the data on disks, they merely split the data into blocks and copy the data that changed or was modified. Many storage vendors, including IBM, offer replication services for their disk storage systems as an optional feature of service delivery, typically as part of a backup and recovery solution. Copy services can be even further extended to long distances through a WAN to fulfill disaster recovery requirements or to increase the availability of application services across geographies.

Figure 10-1 shows how this data (logical unit number (LUN)) migration works, in principle.

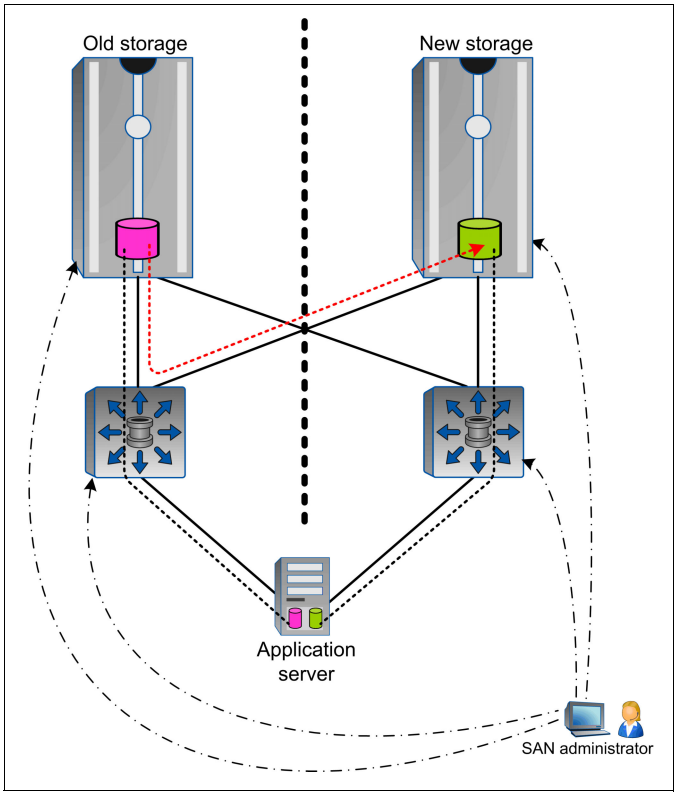


Figure 10-1 SAN-based replication of LUNs

In Figure 10-1, the storage administrator is challenged to migrate data to the newly deployed, high performance disk storage system without interrupting the client's critical SAP applications. Luckily, we can manage both source and target storage systems. These systems are configured to communicate through SAN switches. Disk copy services can replicate specific LUNs from the old storage device to the new storage device without affecting the performance of the SAP application.

In addition, you can use this procedure to prepare a standby application server that connects to the replicated disk LUNs. Or, you can use this procedure to replace the old server hardware where the SAP application is running, with the minimum outage that is necessary to switch the application over to the prepared server.

Host-based data migration

Host-based migration of storage data is the option that is used when the storage administrator is unable to establish a connection between the source and target disk storage system. This type of migration typically happens in data centers with two independent SANs. In most cases, each of these SANs is managed by a different team of administrators or even by different vendors.

Figure 10-2 shows the principle of host-based data migration.

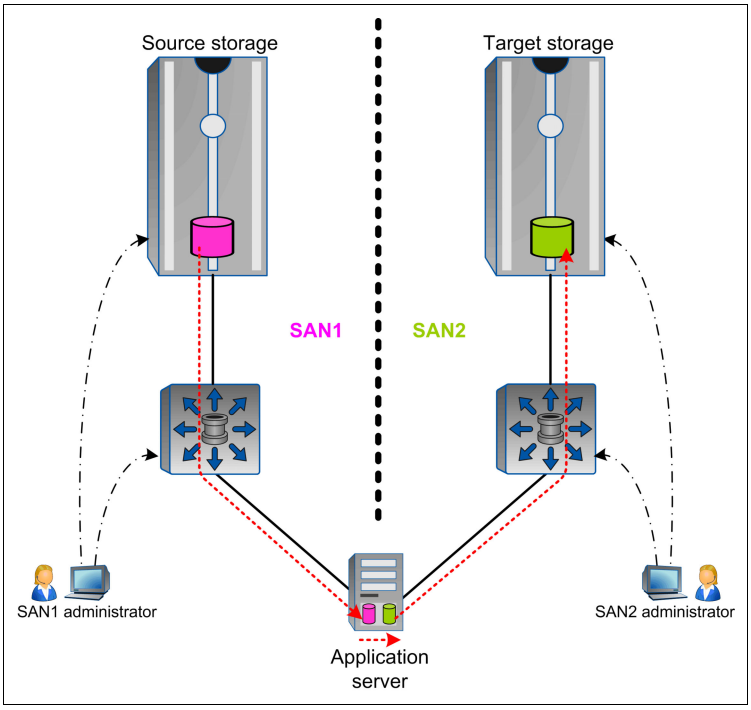


Figure 10-2 Host-based migration of data

The application server is connected to both SANs by using independent HBAs. Application owners and SAN2 administrators analyze the current disk structure that is assigned from the source storage system. The SAN2 administrator assigns the same disk capacity to the application server. The application or system owner then migrates the data from the source disk to the target disk. This migration is performed manually by using the operating system functions. (The application is offline.) Or, disk mirroring must be enabled.

When the data is synchronized between the source and target disks, the mirror can be broken, source disks can be unassigned, and the source storage system can be disconnected from the application server. The disadvantage of this solution is a significant I/O operation on the source and target LUNs that can potentially affect the performance of critical applications.

Remote data copy and migration

Remote copy or *data migration* is a business requirement that is 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 data centers, typically over long distances, sometimes even continents.

Remote copy solutions are either *synchronous* or *asynchronous*, and they require different levels of automation 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. Complete solutions are available from various vendors to support data migration projects to optimally schedule and use client network resources and to eliminate the effect on critical production environments.

Products, such as the vendor solutions, help clients efficiently and effectively migrate all of the SAN data volumes from small remote data centers to a central data center across a WAN without interruption to the service.

With advanced storage management techniques, such as outboard hierarchical storage management (HSM) and file pooling, remote copy solutions need to be implemented at the file level. These techniques imply that more data needs to be copied, and they require more advanced technologies to guarantee data consistency across files, disks, and tape in multiple server heterogeneous environments. The data center networking infrastructure is required to support various data transfer protocols to support these requirements. Examples of these interfaces include FCoE, Converged Enhanced Ethernet (CEE), or simple Internet SCSI (iSCSI).

Real-time snapshot copy

Another outboard copy service that is enabled by Fibre Channel technology is the *real-time snapshot*, which is also known as *time=zero* (*T0*) copy. This service is the process of taking an online snapshot, or freezing the data (databases, files, or volumes) at a certain time. This process allows the applications to update the original data while the frozen copy is duplicated. With the flexibility and extensibility of Fibre Channel, 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 the 24x7 availability of key database systems. This solution is optimal in homogeneous infrastructures that consist of devices from a single vendor.

10.2.4 Upgrading to faster speeds

One of the other considerations of any SAN environment is how to introduce newer, faster technology. Both 8 gigabit per second (Gbps) Fibre Channel and 10 gigabit over Ethernet (GbE) products are prevalent in the market and participate in data center networking. Now, vendors are moving forward with even faster technologies and products, such as 16 Gbps Fibre Channel ports and HBAs.

For most applications, this faster technology does not mean that the application can immediately benefit. Applications with random or "*bursty*" I/O might 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 logical place to use 16 Gbps is in the inter-switch link (ISL) scenario. This scenario has two advantages. The increased speed between switches is an obvious advantage. And, the increased bandwidth might mean that you need fewer ISLs. If fewer ISLs are required, it might be possible to reassign ISLs to attach hosts or storage.

Another consideration is cost. IT architects and investors must evaluate their current SAN solutions in their data centers and make strategic decisions to determine whether it is beneficial to continue with the upgrade to a dedicated Fibre Channel solution that is running 16 Gbps devices. Or, the architects and investors must determine whether now is the right time to consider an upgrade to converged networks to use, for example, FCoE. Many products are available on the market that support these transformations and transitions and protect client investments for the future.

10.3 Infrastructure simplification

An important critical business requirement is the need for IT infrastructures to better support business integration and transformation efforts. Often, the simplification and streamlining of core storage provisioning services and storage networking are at the center of these efforts.

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 verge of becoming true on-demand businesses. These businesses are highly competitive in the market.

Is your IT infrastructure a complex set of disparate, server-specific, and siloed applications that operate across an endless area of servers, for example:

- Transaction processing servers
- Database servers
- Tiered application servers
- Data gateways
- Human resource servers
- Accounting servers
- Manufacturing servers
- Engineering servers
- Email servers
- Web servers

If so, you must be able to answer these questions:

- Where can we deploy the next application?
- Where can we physically put the next server?
- How can we extend our storage resources?
- How can we connect more virtual or physical servers?
- Or, does a simpler way exist to manage all of these servers?

We try to answer all of these questions in the following topics.

10.3.1 The origin of the complexity

A SAN, in theory, is a simple thing. A SAN is a path from a server to a common storage resource. Therefore, where did all of the complexity come from?

Limited budgets and short-sighted strategic thinking push IT organizations into looking for short-term solutions to pain points. When a new application or project becomes available, the easy, inexpensive option is to add another low-cost server. Because 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, storage systems are also added to the sprawl. Every server has two or four HBAs and a share of the consolidated storage. As more servers are added, 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 are probably purchased from multiple vendors, with decisions made on cost, suitability to a specific application, or merely someone’s personal preference. The servers of different vendors are tested on specific SAN configurations. Every server vendor has its own interoperability matrix or a list of SAN configurations that the vendor tested and that the particular vendor supports. It might be difficult for a SAN administrator to identify the correct 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 not possible. When it was possible, expensive and unique extension technology was required. 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 major types of storage device pooling: *disk pooling* and *tape pooling*.

Disk pooling

Disk pooling allows multiple servers to use 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 that are supported by the operating systems on the servers. The servers are potentially a heterogeneous mix of UNIX, Microsoft Windows, and even IBM z Systems servers.

Storage can be dynamically added to the disk pool and assigned to any SAN-attached server when and where necessary. This function provides efficient access to shared disk resources without a level of indirection that is associated with a separate file server. This scenario is possible because storage is effectively *directly attached* to all of the servers, and efficiencies of scalability result from the consolidation of storage capacity.

When storage is added, you can use *zoning* to restrict access to the added capacity. Because many devices (or LUNs) can be attached to a single port, access can be further restricted by using LUN-masking. You can use LUN masking to specify who can access a specific device or LUN.

You can attach and detach storage devices under the control of a common administrative interface. Storage capacity can be added without stopping the server, and the storage capacity can be available to the applications immediately.

Figure 10-3 shows an example of disk storage pooling across two servers.

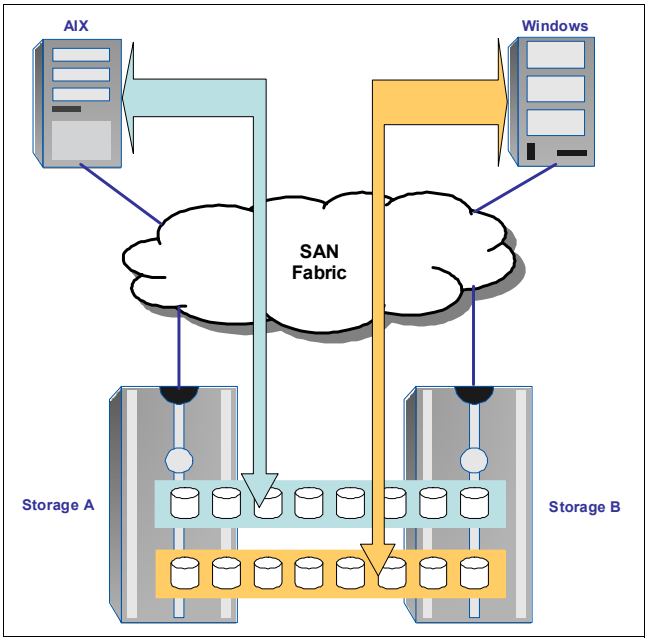


Figure 10-3 Disk pooling concept

In Figure 10-3, one server is assigned a pool of disks that are 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 might be the space that is not yet allocated, or the third pool can be a preformatted disk for future use. Again, all of the changes in the disk structure can be performed dynamically without interruption to the service.

Tape pooling

Tape pooling addresses a problem in an open systems environment where 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 the tape device is 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. These devices 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 to serialize tape access. Tape pooling is an 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 that use, for example, IBM Tivoli Storage Manager with SAN-attached IBM tape libraries.

Logical volume partitioning

At first sight, an individual might ask, “How will logical volume partitioning make my infrastructure simpler? It looks as though we are creating more and more pieces to manage in my storage”. Conceptually, this thought is correct, but the benefit of *logical volume partitioning* is to address the need for maximum volume capacity and to effectively use it within target systems. Logical volume partitioning 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 might be a subset or superset of disks that are 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 the logical disk is defined.

For example, two logical disks, with different capacities (for example, 50 GB and 150 GB) might be created from a single 300 GB hardware-addressable disk. Each of the two disks is assigned to a different server, which leaves 100 GB of unassigned capacity. A single 2,000 GB logical disk might 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 a storage controller can be further used by certain storage virtualization engines, such as the IBM SAN Volume Controller. This engine, when compared to environments that do not use this controller, offers better and more scalability and virtualization of storage resources. The SAN Volume Controller provides these benefits with less management effort and clearer visibility to the target host systems.

Figure 10-4 shows multiple servers that are accessing logical volumes that were created by using the different alternatives that we mentioned. (The logical volume, which is called the *unallocated volume,* is not assigned to any server.)

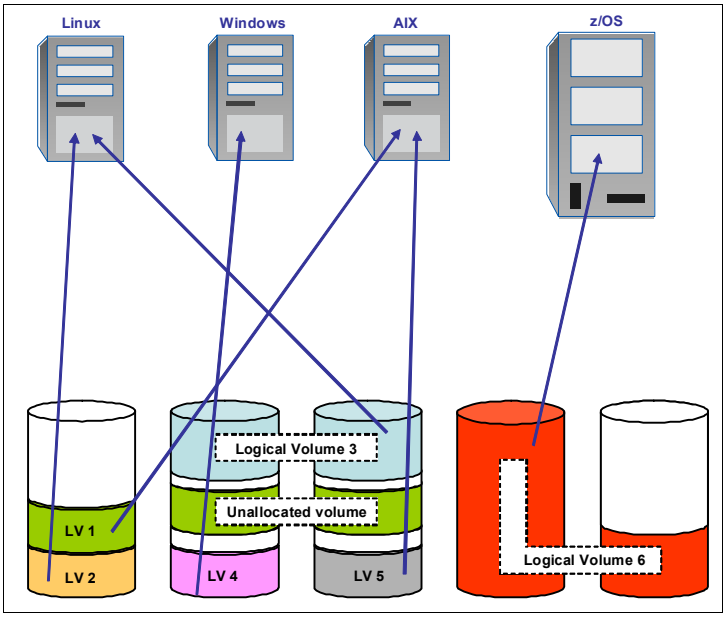


Figure 10-4 Conceptual model of logical volume partitioning

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. This consolidation occurs as clients' experience and requirements grow and change. This type of solution is also known as *data center bridging*.

Clients deploy multiple SAN islands for different applications with different fabric switch solutions. The growing availability of iSCSI server capabilities creates the opportunity for low-cost iSCSI server integration and storage consolidation. Additionally, depending on the choice of router, iSCSI servers can provide *Fibre Channel over IP* (*FCIP*) or Internet Fibre Channel Protocol (iFCP) capability.

The available multiprotocol SAN routers provide an iSCSI Gateway Service to integrate low-cost Ethernet-connected servers to existing SAN infrastructures. The iSCSI Gateway Service also provides a Fibre Channel-to-Fibre Channel (FC-FC) Routing Service to interconnect multiple SAN islands without requiring the fabrics to merge into a single large SAN.

Figure 10-5 shows an example of using a multiprotocol router and converged core switch to extend SAN capabilities across long distances or merely over metropolitan areas.

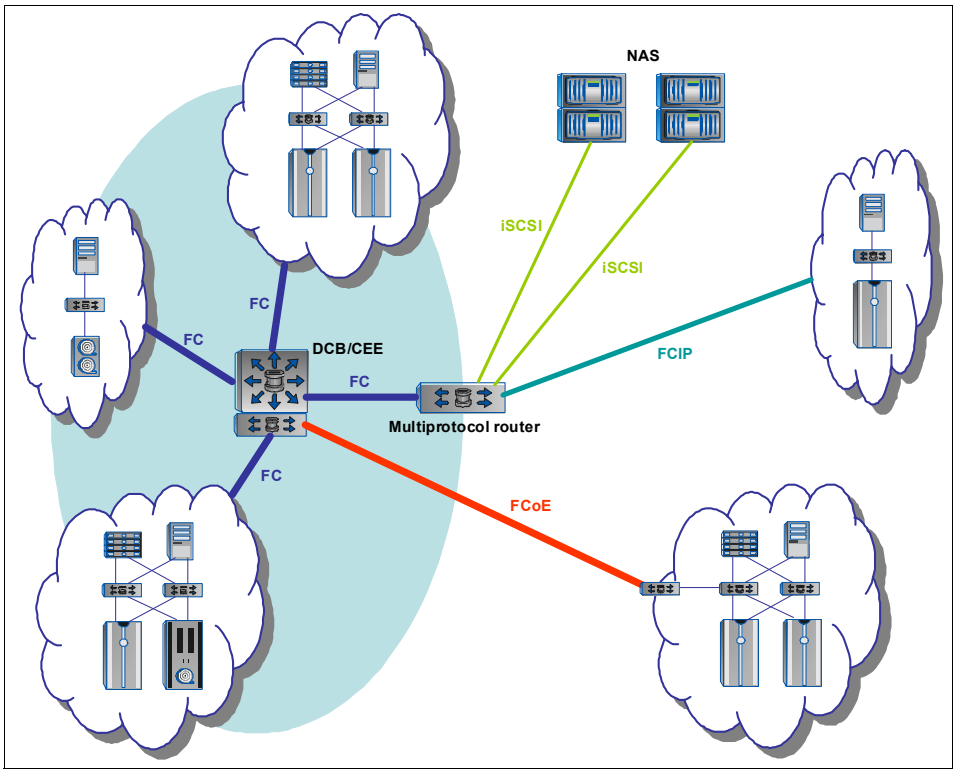


Figure 10-5 The concept of SAN consolidation

A multiprotocol-capable router solution offers many benefits to the marketplace. In our example, discrete SAN islands and several protocols are involved. Many disruptive and potentially expensive actions are involved to merge these SAN fabrics:

- Downtime
- Purchase of more switches and ports
- Purchase of HBAs
- Migration costs
- Configuration costs
- Purchase of more licenses
- Ongoing maintenance

However, many advantages are available by installing a multiprotocol router or core FCoE-enabled switch or director:

- 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 the router
- Support of other protocols
- Increased return on investment (ROI) by consolidating resources
- Capability to use router to isolate the SAN environment for greater security

The router and core switch can provide more benefits. In this example, which is an FC-FC routing service that negates the need for a costly SAN fabric merger, the advantages are apparent and real. A router can also provide the following benefits:

- Device connectivity across multiple SANs for infrastructure simplification
- Tape-backup consolidation for information lifecycle management (ILM)
- Long-distance SAN extension for business continuity
- Low-cost server connectivity to SAN resources

10.3.4 Migration to a converged network

Medium-sized and enterprise data centers typically run multiple separate networks. These networks include 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, data centers 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. The 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 the use of FCoE and helps address these challenges efficiently and effectively. In the following topics, we briefly describe how to upgrade your current infrastructure to a converged network in three principal steps. The prerequisite of the converged network is lossless 10 Gbps over Ethernet (10 GbE or higher), inline with the data center bridging (DCB) standards.

Figure 10-6 presents the concept of the migration to convergency.

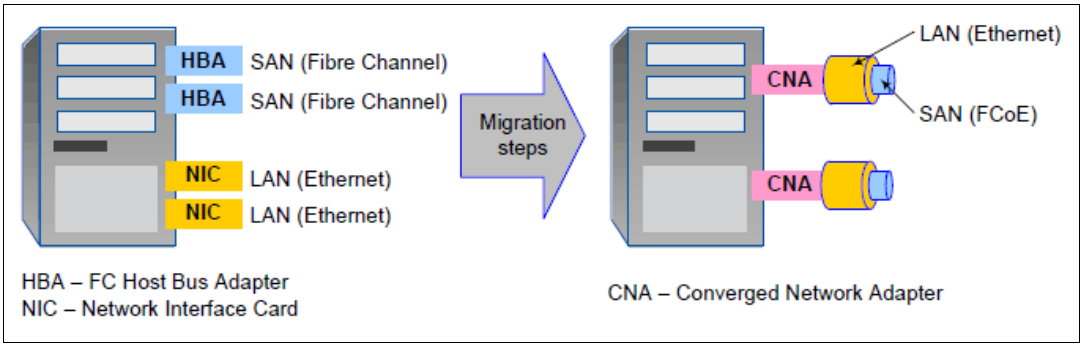


Figure 10-6 Conceptual model of migration to a converged network

The following list provides a summary of the key benefits of upgrading to a converged network:

- Reduced capital expenditures by 15% - 30%, depending on current infrastructure
- Reduced operational and management costs by 30% - 50%
- Improved network and storage efficiencies
- Increased asset and storage utilization
- Improved flexibility and business agility
- Reduced power consumption by 15% - 20%

The following steps describe the process of migrating to a converged network:

1. Access layer convergence.

   Assume that we have separate adapters for a 1 Gbps or 10 Gbps Ethernet communication (2 - 8 adapters for each server) and FC HBAs for storage networking (2 - 6 adapters, typically dual-port). In this step, we replace these combinations by using Converged Network Adapters (CNAs). See Figure 10-7.

   Fabric deployment: For illustration, only a single fabric data center solution is presented in all of the figures. In real data centers, dual-fabric deployment is essential.

   Additionally, we must install a switch that supports both protocols, IP and FCoE, typically as a *top-of-rack* (*TOR*) device. Therefore, the TOR device that supports DCB standards and multiple protocols can continue to work with the existing environment. The device can also separate the network traffic from the data storage traffic and direct each of them to the correct part of the data center networking infrastructure. All of the traffic is segmented at the access layer, which is the first step of the overall process.

   Figure 10-7 shows the first step in the migration process, which is access layer convergence.

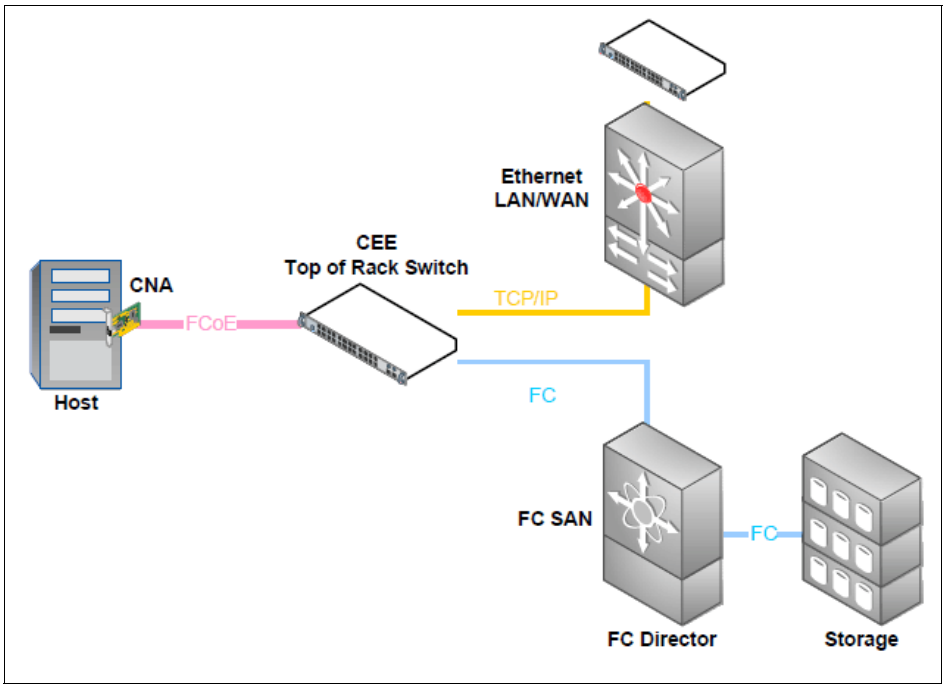


Figure 10-7 Access layer convergence

2. Fabric convergence.

   The second step is to implement more core types of switches that support data center bridging and converged network protocols. Therefore, rather than implementing a converged network on TOR switches or blades, we move this function to the core directors. A second stage of the development of DCB standards introduces *multi-hop* bridging because different solutions are available from each of the vendors of the SAN networking products (Figure 10-8).

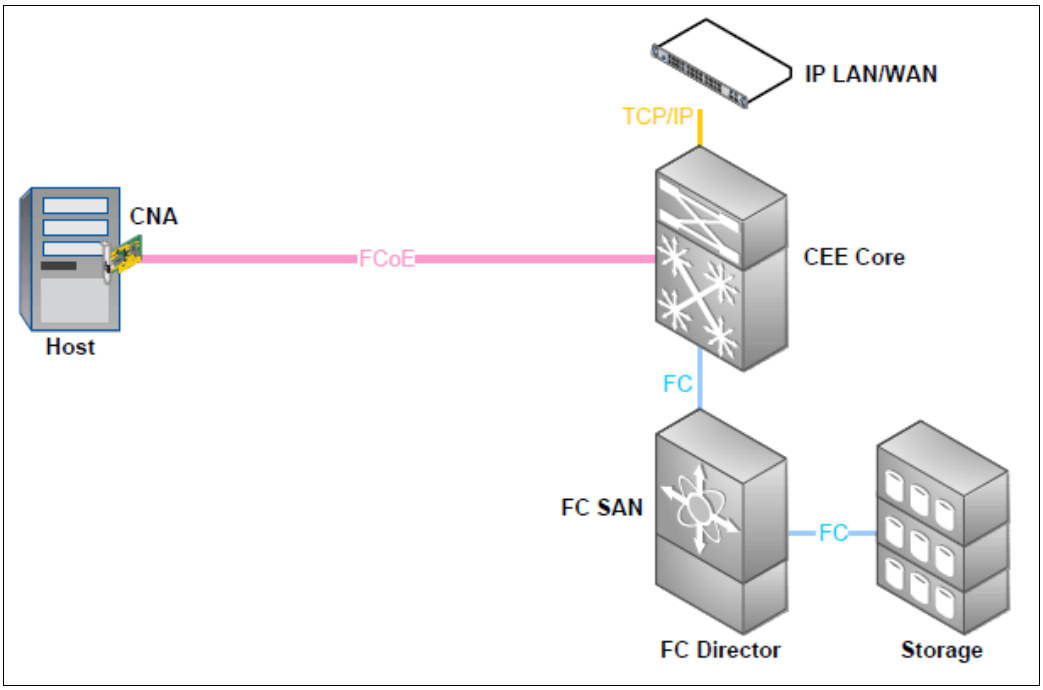


Figure 10-8 Fabric convergence

3. Storage convergence.

   For the final step of the migration, we implement native FCoE-enabled storage devices. Now, various vendors with mid-range to enterprise disk storage systems already offer FCoE. This step enables 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 step dramatically reduces the requirements for operation and management of the infrastructure, reduces the power consumption, and simplifies the complexity of the network (rack space and cabling). Figure 10-9 shows the final status of the converged network infrastructure.

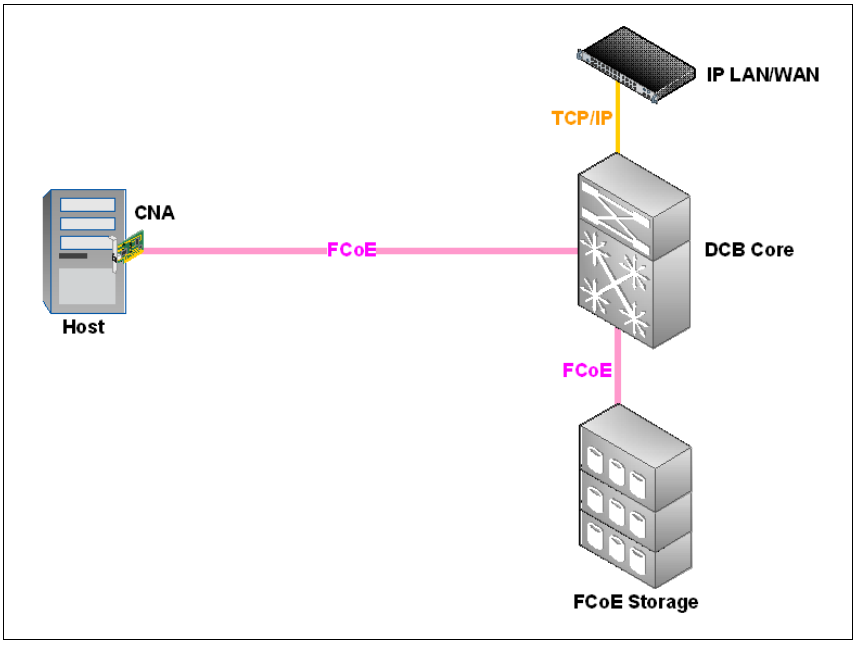


Figure 10-9 Storage convergence

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 Fibre Channel concepts. Examples of these concepts include virtual SANs (VSANs), worldwide names (WWNs), FC IDs (FCIDs), multipathing, and zoning to servers and storage arrays.
- A high level of performance, which is comparable to the performance of current Ethernet and Fibre Channel networks. These networks are 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 GbE, 10 GbE, and 40 GbE, and eventually 100 GbE).
- Simplified operations and management (no change to the management infrastructure that is deployed in the SANs).

10.4 Business continuity and disaster recovery

On-demand businesses rely on their IT systems to conduct business. Everything must work all of the time. Failure truly is not an option. A sound and comprehensive business continuity strategy that encompasses high availability, near-continuous operations, fault-tolerant systems, and disaster recovery is essential.

Data protection of multiple network or SAN-attached servers is performed according to one of two backup and recovery approaches: local backup and recovery, or network backup and recovery.

The *local backup and recovery* solution offers the advantage of speed because the data does not travel over the network. However, with a local backup and recovery approach, costs occur for overhead because local devices must be acquired for each server and are therefore difficult to use efficiently. Also, costs occur for management overhead because of the need to support multiple tape drives, libraries, and mount operations.

The *network backup and recovery* approach that uses shared tape libraries and tape drives a SAN is cost-effective. This approach is efficient because it centralizes storage devices that use one or more network-attached devices. This centralization shortens the ROI because the installed devices are used 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 a network backup and recovery environment eliminates the potential need to perform manual tape mount operations on multiple servers.

SANs combine the best of both approaches because of the central management of backup and recovery. You can assign one or more tape devices to each server and use FCP to transfer data directly from the disk device to the tape device, or vice versa, over the SAN.

Another important topic in this category is instant business continuity if a device fails. Critical applications cannot afford to wait until their data is restored to a 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 refer to these environments as *highly available* or *fault-tolerant systems*.

10.4.1 Clustering and high availability

SAN architecture naturally allows multiple systems (target hosts) to access the same disk resources in medium-sized or enterprise disk storage systems, even concurrently. This feature enables specific applications, such as AIX IBM HACMP™, IBM PowerHA®, and Microsoft Windows Cluster Services, that run on the hosts to introduce highly available or fault-tolerant application systems.

These systems ensure that in a single host failure, the application is automatically (without any manual administrator intervention) moved over to the backup cluster host. The high availability means a short outage. Or, the failed host is isolated from application processing, and the workload is balanced among other working cluster nodes. This method is called *fault-tolerant*, which means no disruption to service.

Figure 10-10 shows the concept of a highly available cluster solution.

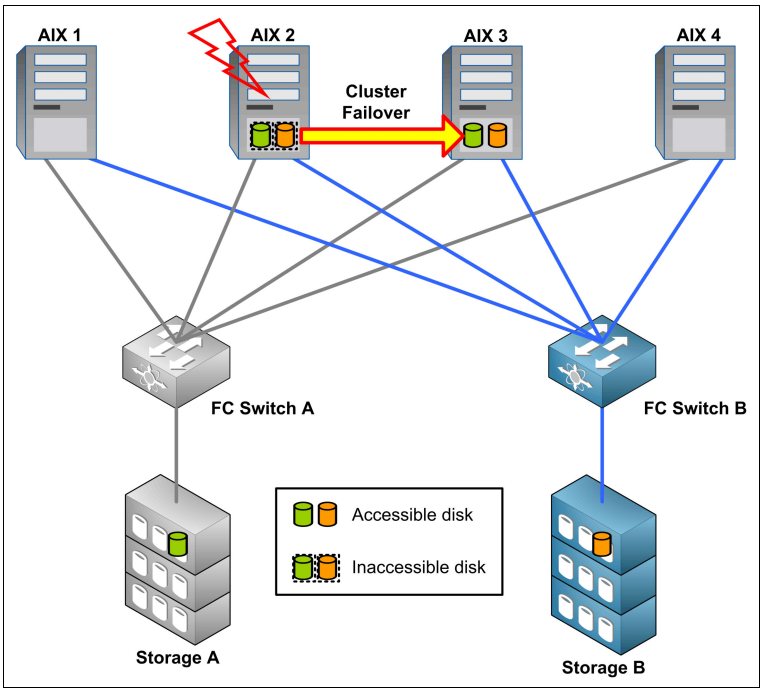


Figure 10-10 Highly available cluster system

In Figure 10-10, an application runs on system AIX2. The system manages mirrored disks from both of the storage systems (green and orange). SAN zoning allows both cluster nodes (AIX2 and AIX3) to operate the same set of disks. The cluster has one primary cluster node that is active (an active-passive cluster). When AIX2 fails, cluster services that run on AIX3 recognize the failure and automatically move all of the application resources to AIX3. The disk sets are activated on AIX3, and the application is started in the correct sequence.

Figure 10-11 shows the configuration of a fault-tolerant clustered environment.

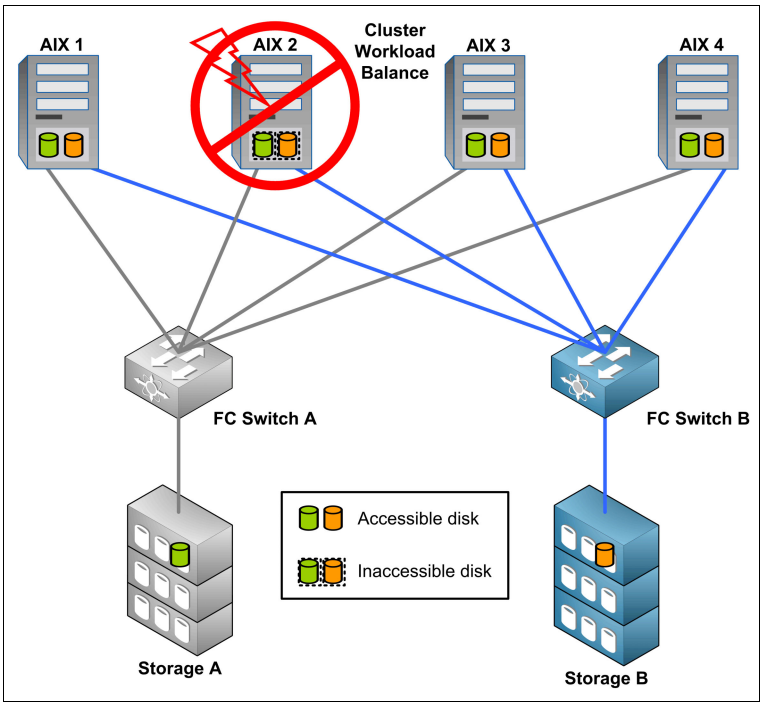


Figure 10-11 Fault-tolerant cluster system

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. If the AIX2 system fails, the cluster application automatically deactivates the assigned disks and redistributes the workload among the remaining active cluster nodes. No interruption occurs to the business. This configuration is costly and typically only employed for business critical applications, such as banking systems and air traffic control systems.

10.4.2 LAN-free data movement

The network backup and recovery approach implies that data flows from the backup and recovery client (typically a file or database server) to the centralized backup and recovery server. Or, the data flows between the backup and recovery servers over the Ethernet network. The same approach applies for the archive or storage management applications. Often, the network connection is the bottleneck for data throughput, especially in large database systems, because of the network connection bandwidth limitations. The SAN offers an advantage to offload the backup data out of the 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 the central backup tape repository and backup and recovery clients with large database files. Systems with a high number of small files still use the network for data transportation because they cannot benefit from a LAN-free environment.

In the tape drive and tape library sharing approach, the backup and recovery server or client that requests that a backup copy is copied to or from tape reads or writes the data directly to the tape device by 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. The central backup and recovery server controls only the tape mount operations and stores the references (metadata) in its embedded database system.

Figure 10-12 shows an example of tape library sharing and LAN-free backups.

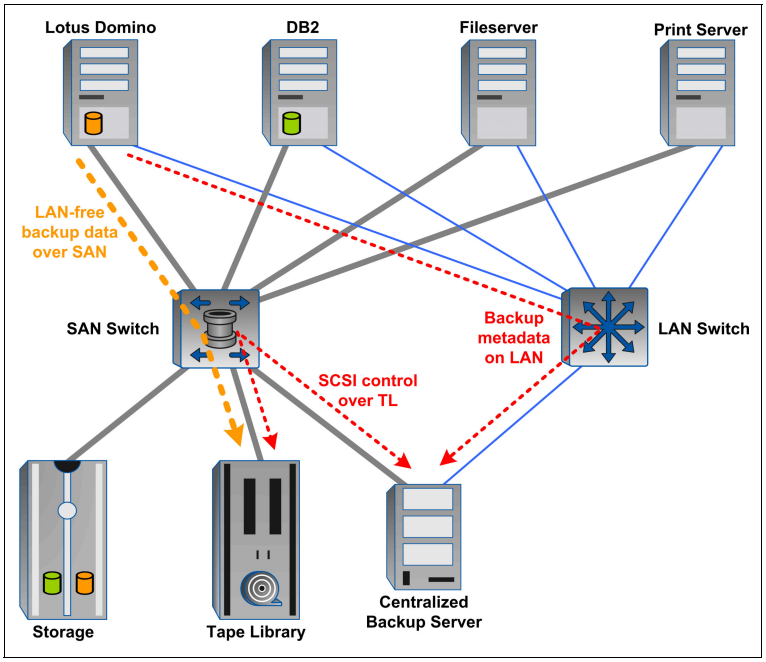


Figure 10-12 LAN-free and LAN-based backups

IBM Lotus® Domino and IBM DB2 database systems benefit from the improved performance of backups directly to tapes over Fibre Channel. However, small servers with a high number of files still continue to back up to a LAN or WAN.

IBM offers enterprises a centralized backup and recovery solution that supports various platforms and database systems. IBM Tivoli Storage Manager and its component IBM Tivoli Storage Manager for SANs enable clients to perform online backups and archives of large application systems directly to tape over a SAN without a significant effect on performance.

10.4.3 Disaster backup and recovery

A SAN can facilitate disaster backup solutions because of the greater flexibility that is allowed in connecting storage devices to servers. Backup solutions are also simplified because of the greater distances that are supported by a SAN when compared to SCSI restrictions.

You can now perform extended-distance backups for disaster recovery within a campus or city (Figure 10-13).

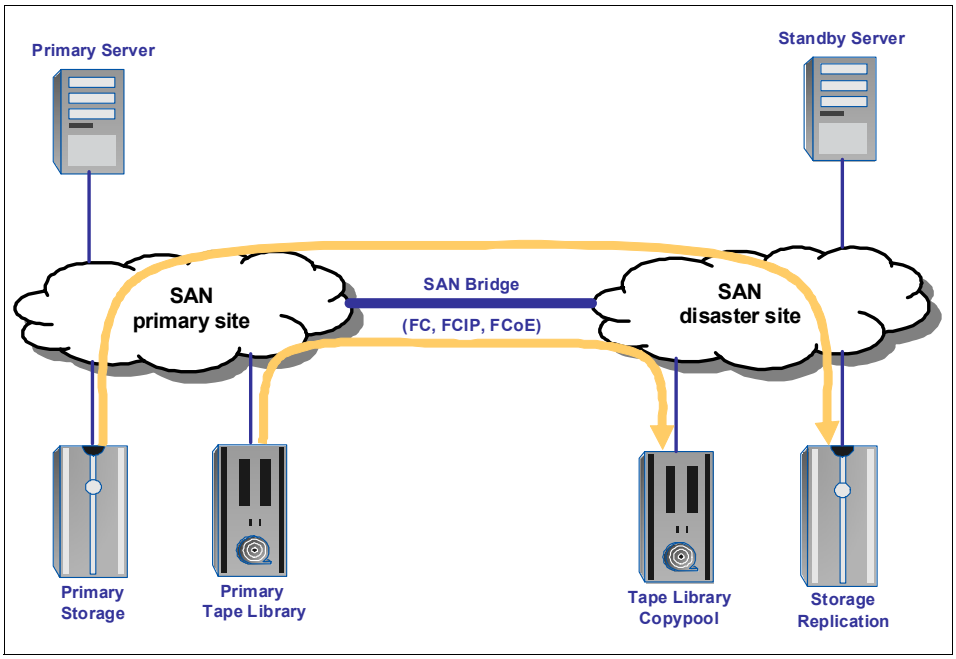


Figure 10-13 Disaster backup at a remote site by using SAN bridging

When longer distances are required, SANs must be connected by using gateways and WANs. One of the solutions is FCoE.

Depending on your business requirements, disaster protection deployments might use copy services that are implemented in disk subsystems and tape libraries (that might be achieved by using SAN services), SAN copy services, or 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 merely hardware or software; it includes processes and policies to manage the information. ILM is designed on the recognition that different types of information can have different values at different points in their lifecycles. Predicting storage needs and controlling costs can be especially challenging as your 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. To effectively implement ILM, owners of the data need to determine how information is created, how it ages, how it is modified, and when it can safely be deleted. ILM segments the 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 Information lifecycle management

To manage the data lifecycle and to prepare your business for on-demand services, four main elements can address your business in an ILM-structured environment:

- Tiered storage management
- Long-term data retention and archiving
- Data lifecycle management
- Policy-based archive management

10.5.2 Tiered storage management

Most organizations 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 to maintain business continuity. Tiered storage can help reduce overall disk-storage costs by providing the following benefits:

- Reducing overall disk-storage costs by allocating the most recent and most critical business data to higher-performance disk storage. Costs can also be reduced by moving older and less critical business data to lower-cost disk storage.
- Accelerating business processes by providing high-performance access to the 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 (Figure 10-14), which limits the ability to optimize cost and performance.

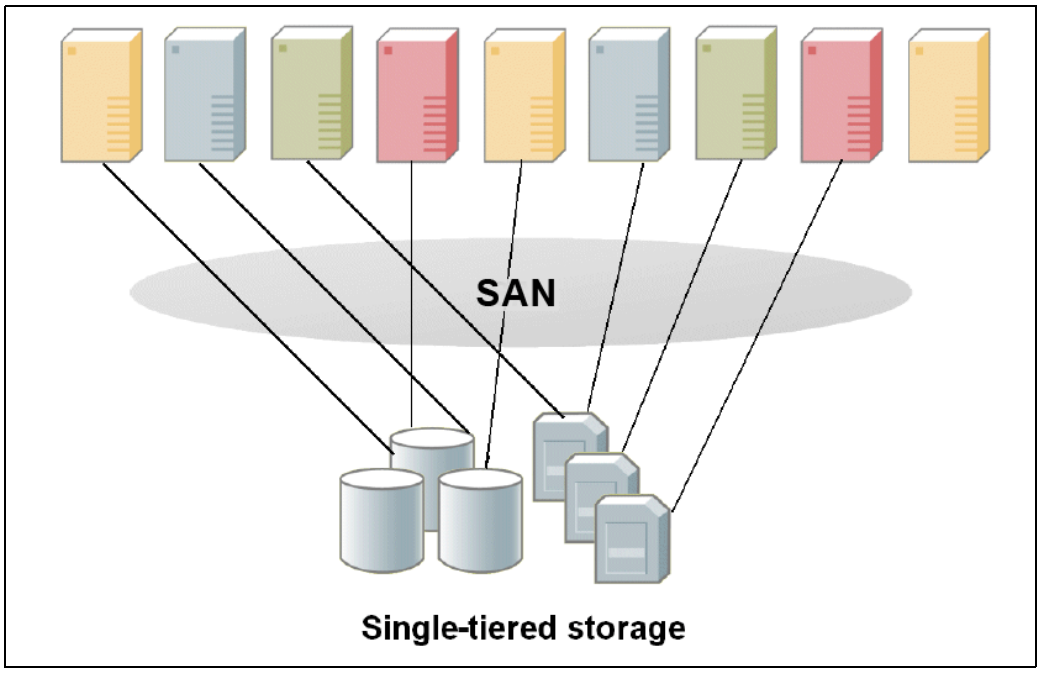


Figure 10-14 Traditional single-tiered storage environment

Multi-tiered storage environment

A tiered storage environment that uses the SAN infrastructure offers you the flexibility to align the storage cost with the changing value of information. The tiers relate to data value. The most critical data is allocated to higher-performance disk storage. The less critical business data is allocated to lower-cost disk storage.

Each storage tier provides different performance metrics and disaster recovery capabilities. The creation of the 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.

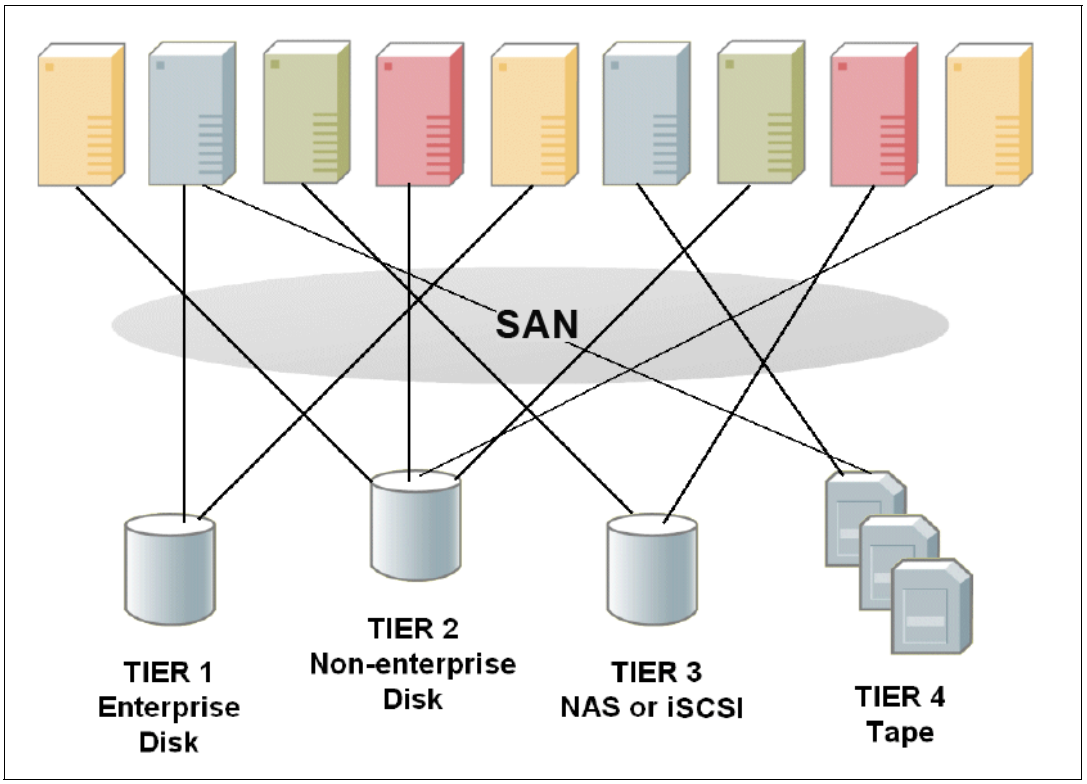


Figure 10-15 ILM tiered storage environment

An IBM ILM solution in a tiered storage environment is designed with the following factors:

- Reduces the TCO of managing information. It can help optimize data costs and management by freeing expensive disk storage for the most valuable information.
- Segments data according to value. Segmenting data can help to create an economical balance and sustainable strategy to align storage costs with business objectives and information value.
- Helps you decide about moving, retaining, and deleting data because ILM solutions are closely tied to applications.
- Manages information and determines how the information is managed based on content, rather than migrating data that is based on technical specifications. This approach can help result in more responsive management. This solution offers you the ability to retain or delete information according to your business rules.
- Provides the framework for a comprehensive enterprise content management strategy.

10.5.3 Long-term data retention

A rapidly growing class of data is best described by how it is managed rather than the arrangement of its bits. The most important attribute of this data is its retention period; therefore, this data is called *retention-managed data*. This data is typically kept in an archive or a repository. In the past, this data was known as archive data, fixed content data, reference data, unstructured data, and other terms that imply its read-only nature. This data is often measured in terabytes and kept for long periods of time, sometimes forever.

In addition to the sheer growth of data, certain laws and regulations that govern the storage and secure the retention of business and client information are increasingly part of the business landscape. These regulations make data retention a major challenge to any institution. An example of this challenge is the Sarbanes-Oxley Act, which was enacted in the US in 2002.

Businesses must comply with these laws and regulations. Regulated information includes the following types of data:

- Email
- Instant messages
- Business transactions
- Accounting records
- Contracts
- Insurance claims processing

All of these types of information can have different retention periods. These periods can be two years, seven years, or forever (permanent retention). Data is an asset when it must be kept; however, data that is kept past its mandated retention period might also become a liability. Furthermore, the retention period can change because of factors, such as litigation. All of these factors mandate tight coordination and the need for ILM.

In addition to the numerous state and governmental regulations that must be met for data storage, industry-specific and company-specific regulations also must be met. And these regulations are constantly updated and amended. Organizations must develop a strategy to ensure that the correct information is kept for the correct period of time, and that the information is readily accessible when it must be retrieved at the request of regulators or auditors.

It is easy to envision the exponential growth in data storage that results 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 you consider 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. These resources include high-performance, high-capacity disk arrays; lower-cost disk arrays, such as Serial Advanced Technology Attachment (SATA); tape libraries; and permanent archival media, where appropriate. Yet ILM involves more than just data movement; it also encompasses scheduled deletion and regulatory compliance. Because decisions about moving, retaining, and deleting data are closely tied to the application use of data, ILM solutions are typically closely tied to applications.

ILM can potentially provide the framework for a comprehensive information management strategy and help ensure that information is stored on the most cost-effective media. This framework helps enable administrators to use tiered and virtual storage, and to process automation. By migrating unused data off more costly, high-performance disks, ILM can help by performing the following functions:

- Reduce the cost to manage and retain data
- Improve application performance
- Reduce backup windows and ease system upgrades
- Streamline data management
- Allow the enterprise to respond to demands in real time
- Support a sustainable storage management strategy
- Scale as the business grows

ILM recognizes that different types of information have different values at different points in their lifecycles.

Data can be allocated to a specific storage level that is aligned to its cost, with policies that define when and where data is moved (Figure 10-16).

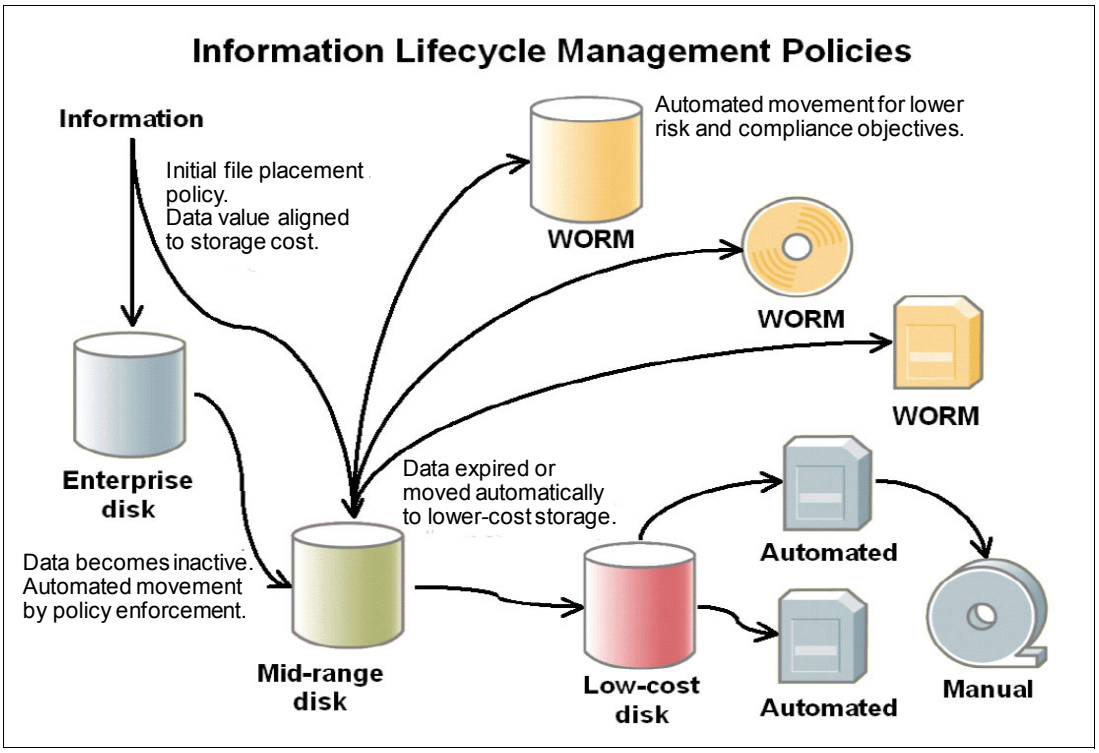


Figure 10-16 ILM policies

10.5.5 Policy-based archive management

Businesses have mountains of data that was captured, stored, and distributed across the enterprise. This wealth of information provides a unique opportunity. By incorporating these assets into solutions and, at the same time, by delivering newly generated information to their employees and customers, a business can reduce costs and information redundancy. Organizations can also take advantage of the potential profit-making aspects of their information assets.

The growth of information in corporate databases, such as enterprise resource planning (ERP) and email systems, makes organizations consider moving unused data off high-cost disks. To effectively manage information in corporate databases, businesses must take the following steps:

- Identify database data that is no longer 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 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.

Email archiving solutions are designed to reduce the size of corporate email systems by moving email attachments and messages to an archive from which they can easily be recovered, if needed. This action helps reduce the need for user management of email, improves the performance of email systems, and supports the retention and deletion of email.

The way to archive is to migrate and store all information assets in an enterprise content management system. ERP databases and email solutions generate large volumes of information and data objects that can be stored in content management archives. By using an archive solution, you can free up system resources, while you maintain access to the stored objects for later reference.

Allowing the archive to manage and migrate data objects gives a solution the ability to have ready access to newly created information that carries a higher value. And at the same time, you are still able to retrieve data that is archived on less expensive media.

For more information, see the *ILM Library: Information Lifecycle Management Best Practices Guide*, SG24-7251.

11

Storage area networks and green data centers

System storage networking products and their deployment in large enterprise data centers significantly participate in the rapid growth of floorspace, power, and cooling resources.

In this chapter, we briefly introduce the concepts of a green data center strategy and how a storage area network (SAN) and IBM Storage align with the green goal. In addition, we also describe the IBM smarter data center that facilitates the evolution of energy-efficient operations.

11.1 Data center constraints

Many data centers are running out of power and space. They cannot add more servers because they reached either their power or space limits, or perhaps they reached the limit of their cooling capacity.

In addition, environmental concerns are becoming priorities because they can also impede the ability of a company to grow. Clients all over the world prefer to purchase products and services from companies that have a sustainable approach to the environment. Clients also want products and services that are able to meet any targets that might be imposed on them, whether from inside their company or from outside in the form of government legislation.

Because environmental sustainability is a business imperative, many data center clients are looking at ways to save energy and cut costs so that their company can continue to grow. However, it is also a time to consider transformation in spending, not just cutting costs. Only smarter investments in technology and perhaps a different way of thinking are needed to achieve green efficiency in data centers.

Data centers must 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 expenditures (CAPEX) and operational expenditures (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.

IT architects need to consider four key factors that drive the efficiency of data center operations:

- Energy cost

  The cost of a kilowatt of electricity rose only slightly in recent years, but the cost of operating servers 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 accelerated the adoption of virtualization technologies and increased virtual image densities. Therefore, this increased demand drives total server energy consumption higher, while the amortized cost of operation for each workload is decreasing.

- Power capacity

  Certain companies cannot deploy more servers because more electrical power is not available. Many suppliers, especially those utilities in crowded urban areas, are telling clients that power feeds are at capacity limits and that they have no more power to sell. New server, storage, and networking products give better performance at lower prices, but they can also be power hungry. The effort to overcome a power supply threshold is a huge investment.

- Cooling capacity

  Many data centers are now 10 - 20 years old, and the cooling facilities are not adapted to their present needs. Traditional cooling methods allowed for 2 - 3 kW of cooling for each rack. Today's requirements are 20 - 30 kW for each 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, growth becomes expensive.

11.1.1 Energy flow in the data center

To understand how to reduce energy consumption, you need to understand where and how the energy is used. You can study energy use in a data center by taking three different views:

- How energy is distributed among IT equipment (servers, storage, and 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 on).
- How the energy that is allocated to IT resources is used to produce business results. (Are idle resources that are powered on spending energy without any effect?)

Figure 11-1 shows how energy is used by several components of a typical non-optimized data center. Each component is divided into two parts: IT equipment (servers, storage, and network) and the infrastructure around it that supports the IT equipment (chillers, humidifiers, air conditioners, power distribution units, uninterruptible power supplies (UPS), lights, and so on).

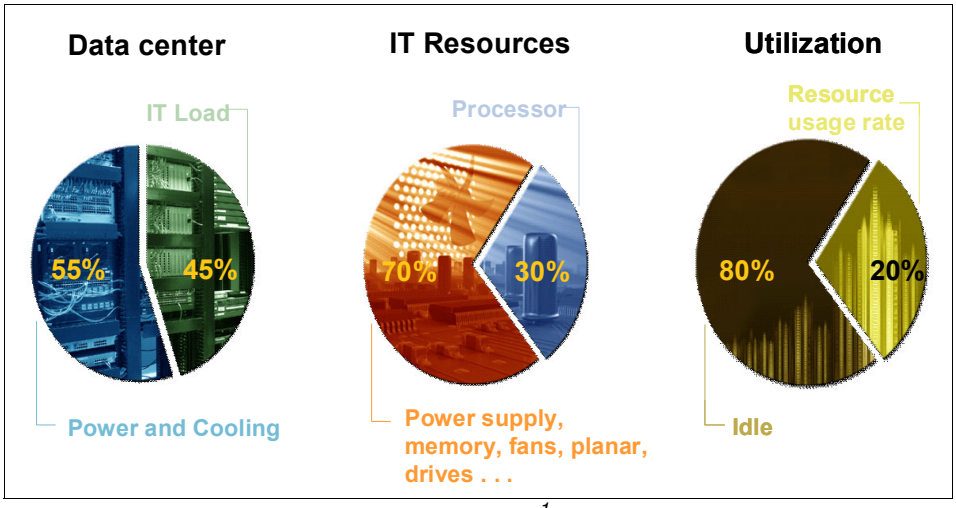


Figure 11-1 Energy usage in a typical data center[1]

In typical data centers, the IT equipment does not use 55% of the overall energy that is brought into the data center. Therefore, this portion of the energy is not producing calculations, data storage, and so on. The concept of a green data center is to eliminate this waste and reduce such inefficiency.

Energy conversion: Basic laws of thermodynamics state that energy cannot be created or destroyed; it changes only in form. The efficiency of this conversion is less than 100% (in a real world, much 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, such as the virtualization of physical servers, can help to change this ratio to a more reasonable value.

[1] Data source: Creating Energy-Efficient Data Centers, US Department of Energy

Companies need to consider the use of IT resources in the data center. A typical server utilization rate is around 20%. Underutilized systems can be a significant issue because significant energy is expended on non-business activities, therefore wasting a major investment. Again, server virtualization, consolidation, and the addressed provisioning of IT resources help to use the entire capacity of your IT equipment.

Data centers must become immensely more efficient to meet their needs while they keep costs in check as the demand for and price of resources continue to rise. But the realization of this efficiency requires a deep and pervasive transformation in how data centers are designed, managed, operated, populated, and billed. These aspects mandate a unified and coordinated effort across organizational and functional boundaries toward a common set of goals.

We introduce the concept of green data centers and how IBM supports the migration to next-generation data centers that are effective, cost-efficient, and environment friendly.

11.2 Data center optimization

To enable your data center to become more effective, consume less power, and become more cost-efficient in terms of infrastructure management and operation, IT architects must consider two components of the migration strategy:

- Optimization of the site and facilities

  This component includes data center cooling, heating, ventilation, air conditioning (HVAC), UPS, and power distribution to the site and within the data center. A standby power supply or alternative power sources must also be considered.

- Optimization of the IT equipment

  This component relates to IT equipment in the data center that generates business value to the clients, such as servers (physical and virtual), disk and tape storage devices, and networking products.

Applying innovative technologies within the data center can yield more computing power per kilowatt. The IT equipment continues to become more energy efficient. Technology evolution and innovation outpace the life expectancy of data center equipment. Therefore, many companies are discovering 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 demonstrate that blade servers reduce power and cooling requirements 25% - 40% over 1U technologies. Replacing equipment before it is fully depreciated might seem unwise. However, 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 typically enough to offset any lost asset value.

11.2.1 Strategic considerations

The strategy of moving toward a green data center and the overall cost-effective IT infrastructure consist of four major suggested areas:

- Centralization:
  - Consolidate many small remote centers into fewer centers
  - Reduce infrastructure complexity
  - Improve facility management
  - Reduce staffing requirements
  - Improve management costs
- Physical consolidation:
  - Consolidate many servers into fewer servers on physical resource boundaries
  - Reduce system management complexity
  - Reduce the physical footprint of servers in the data center
- Virtualization:
  - Remove physical resource boundaries
  - Increase hardware utilization
  - Allocate a less than physical boundary
  - Reduce software license costs
- Application integration:
  - Migrate many applications to fewer, more powerful server images
  - Simplify the 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 exponentially, storage systems within data centers are top candidates for energy savings. Each component of the storage system has power and cooling requirements.

Published studies show that the proportion of energy that is used by storage disk systems and storage networking products varies 15% - 25% of the overall energy consumption of the typical data center. This number significantly increases because the requirements for storage space grow continuously.

However, no matter what efficiency improvements are made, active (spinning) disk drives still use energy if they are powered on. Therefore, the most energy-intensive strategy for data storage is to keep all of the organization's data on active disks. Although this process provides the best access performance, it is not the most environmentally-friendly approach and it is not normally an absolute requirement.

Green storage technologies occupy less raw storage capacity to store the same amount of native valuable client data. Therefore, the energy consumption for each gigabyte of raw capacity falls.

The storage strategy for green data centers includes the following elements:

- Information lifecycle management (ILM)
- Consolidation and virtualization
- On-demand storage provisioning
- Hierarchical storage and storage tiering
- Compression and data deduplication

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 on the recognition that different types of information can have different values at different points in their lifecycles. Predicting storage needs and controlling costs can be especially challenging as the business grows. Although the total value of stored information increases overall, historically, not all data is created equal, and the value of that data to business operations fluctuates over time.

Figure 11-2 shows this trend, which is commonly referred to as the *data lifecycle*. The existence of the data lifecycle means that all data cannot be treated the same.

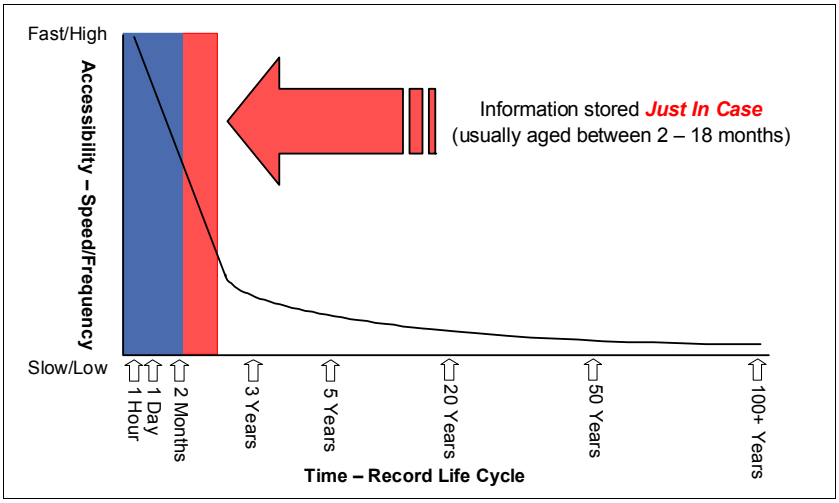


Figure 11-2 Data lifecycle

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 results in a "*buy more storage*" mentality. However, this approach increases overall operational costs, complexity, and 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% is either inactive or even obsolete (Figure 11-3).

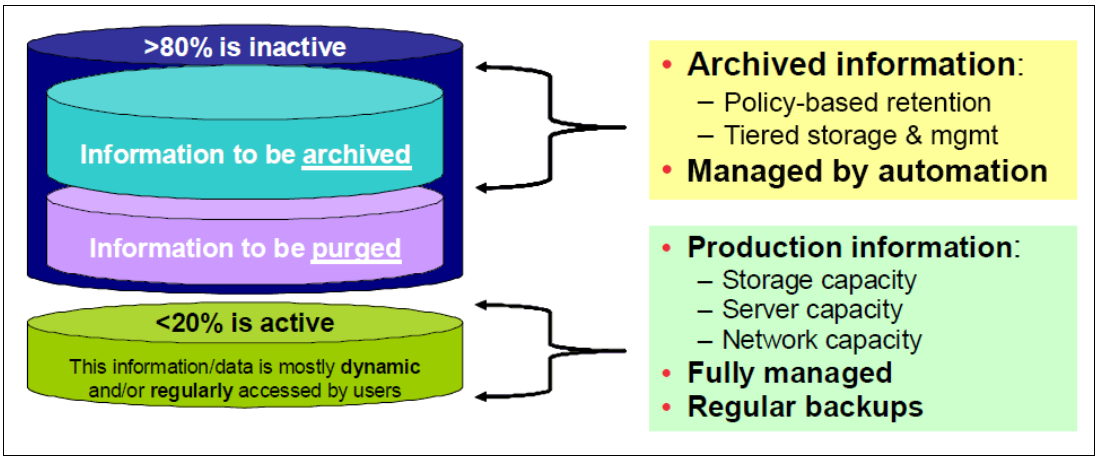


Figure 11-3 Usage of data

The automated identification of the storage resources in an infrastructure and an analysis of how effectively those resources are used are the crucial functions of ILM. File system evaluation and file level evaluation uncover 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 point in the ILM process is the automated control through policies that are customizable with actions that can include centralized alerting, distributed responsibility, and fully automated response. This process also includes data deletion.

For more information, see the *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 grow rapidly, traditional physical approaches to storage management become increasingly problematic. Physically expanding the storage environment can be costly, time-consuming, and disruptive. These drawbacks are compounded when expansion must 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 must be able to make even more rapid changes to stay competitive.

The alternative is a centralized, consolidated storage pool of disk devices that are easy to manage and are 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 shows storage consolidation.

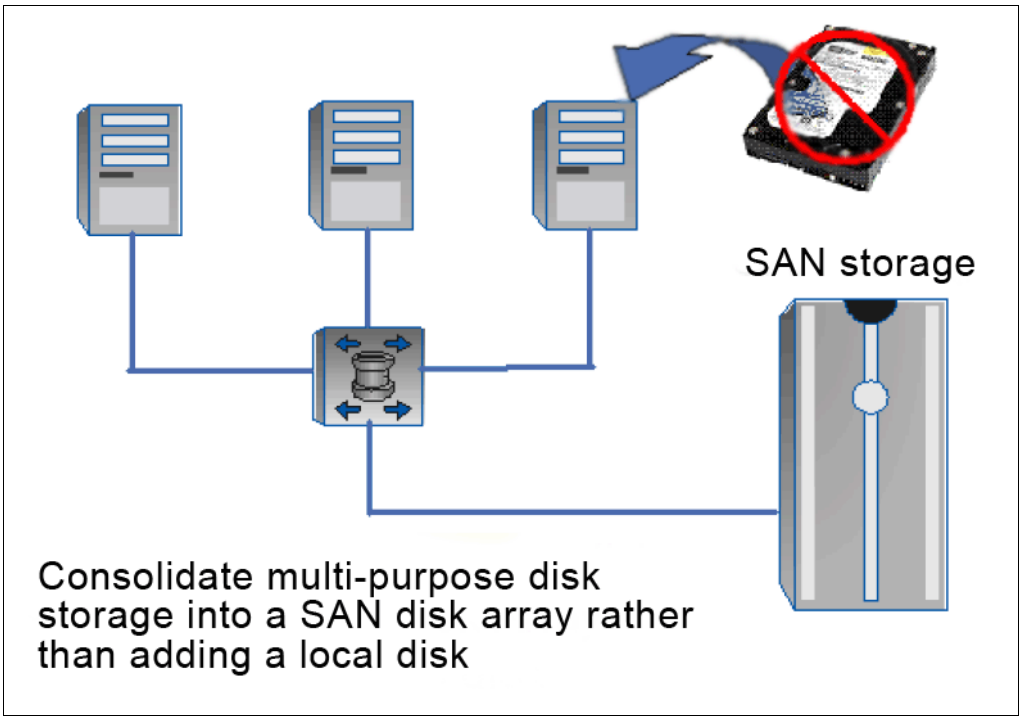


Figure 11-4 Storage consolidation

This logical view can hide undesirable characteristics of storage while it presents storage in a more convenient manner for applications. For example, storage virtualization might 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 exist virtually, regardless of where the actual physical storage resources are 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 because 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 that is 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 in the following ways:

- Reduce data center complexity and improve IT productivity by managing multiple physical resources as fewer virtual resources.
- Meet rapidly changing demands flexibly 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 the 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 and simultaneously control costs and power and cooling requirements.

The value of a virtualized infrastructure is in the increased flexibility that is created by having pools of system resources on which to draw and in the improved access to information that is afforded by a shared infrastructure. Value is also a result of the lower total cost of ownership (TCO) that comes with decreased management costs, increased asset utilization, and the ability to link infrastructure performance to specific business goals.

For more information about how IBM Storage Virtualization solutions can help your organization meet its storage challenges, see the *IBM Information Infrastructure Solutions Handbook*, SG24-7814, or see this website:

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 the ability of an IT department to respond quickly to requests to provision new capacity. However, a solution to this issue is available through automation. An on-demand storage provisioning solution monitors the current disk usage of specified target host systems and applications and allocates more disk capacity for the period of the business need.

End-to-end storage provisioning is the term that is applied to the whole set of steps that are required to provision usable storage capacity to a server. Provisioning covers the configuration of all of the elements in the chain. This process includes the steps 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 process involves a storage administrator that uses several different tools, each focused on a specific task, or the tasks are spread across several people. This spreading of tasks and tools creates many inefficiencies in the provisioning process, which affect the ability of IT departments to respond quickly to changing business demands. The resulting complexity and high degree of coordination can also lead to errors and can possibly affect the systems and application availability.

Automation of the end-to-end storage provisioning process by using workflow automation can significantly simplify this task of provisioning storage capacity. Each step is automated and the rules for preferred practices around zoning, device configuration, and path selection can be applied automatically. The benefits are increased responsiveness to business requirements, lower administrative 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 that range from high performance and high cost to high capacity and low cost. Through the deployment of SANs, many of these storage assets are now physically connected to servers that run many types of applications and create many kinds of information.

Finally, with the arrival of network-resident storage services for the distributed management of volumes, files, and data replication, IT managers have more control. The 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 can use expensive, fast-performance storage components to store often-accessed and mission-critical files, in contrast with using less expensive storage for less frequently used non-critical data. Tiered storage improves efficiency and saves costs. We can identify the following typical storage tiers, which are categorized based on performance and cost for each gigabyte:

- High-performance SAN-attached disk systems (solid-state drive (SSD) or serial-attached Small Computer System Interface (SCSI) (SAS))
- Medium-performance SAN-attached disks (SAS or Serial Advanced Technology Attachment (SATA))
- Network-attached storage (NAS) systems
- Tape storage and other media with sequential access

Each level of storage tier can be assigned manually by a storage administrator, or data can be moved automatically between tiers, which is based on migration policies. Figure 11-5 shows the conceptual model of storage tiering.

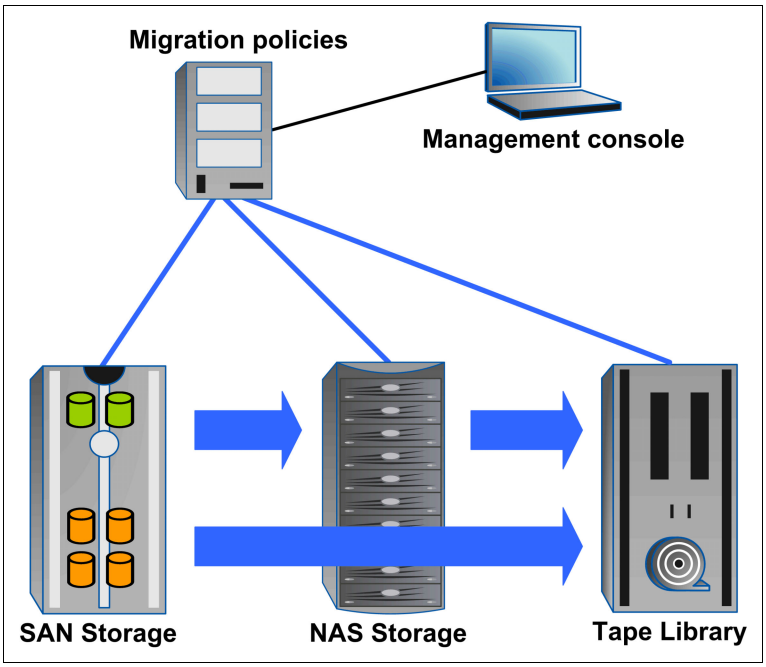


Figure 11-5 Tiered storage concept

IBM offers various tools and utilities for storage tiering and hierarchical management for different scenarios. Tools range from IBM Easy Tier, which is used in enterprise disk storage systems, up through IBM Tivoli Storage Manager for Hierarchical Storage Management for Windows and IBM Tivoli Storage Manager for Space Management for the AIX and Linux platform. For tiered management, IBM offers IBM Global Parallel File System (GPFS™) for data migration between different levels of storage.

11.3.5 Data compression and data deduplication

Business data growth rates will continue to increase rapidly. 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, the focus on improving data storage efficiencies across the information infrastructure increases.

A *data reduction* strategy can decrease the required disk storage and network bandwidth, lower the TCO for storage infrastructures, optimize the use of existing storage assets, and improve data recovery infrastructure efficiency. Compression, data deduplication, and other forms of data reduction are features that are offered in multiple components of the information infrastructure.

Compression immediately reduces the amount of required physical storage across all storage tiers. This solution, which supports the online compression of existing data, allows storage administrators to gain back free disk space in the existing storage system. You can compress data without changing administrative processes or forcing 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. Because data is stored in compressed format in the primary storage system, all other storage tiers and the transports in between realize the same benefits. Replicas, backup images, and replication links all require fewer expenditures after the implementation of compression at the source.

After compression is applied to the stored data, the required power and cooling for each unit of storage are reduced. This reduction is possible 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, fewer overall rack units are required. Figure 11-6 on page 246 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 documented compression ratios as high as 90% in certain Oracle database configurations and about 50% with PDF files. As always, compression ratios vary by data type and how the data is used.

In contrast to compression, the *data deduplication* mechanism identifies identical chunks of data within a storage container. This process keeps only one copy of each chunk. All of the other logically identical chunks point to this chunk. Various implementations of this method exist. One option is inline data deduplication and the other option is post-processing data deduplication. Each chunk of data must be identified in a way that is easily comparable. Chunks are processed by using a parity calculation or cryptographic hash function.

This processing gives the chunks shorter identifiers that are known as *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 shows the typical compression rates that can be achieved with specific IBM products.

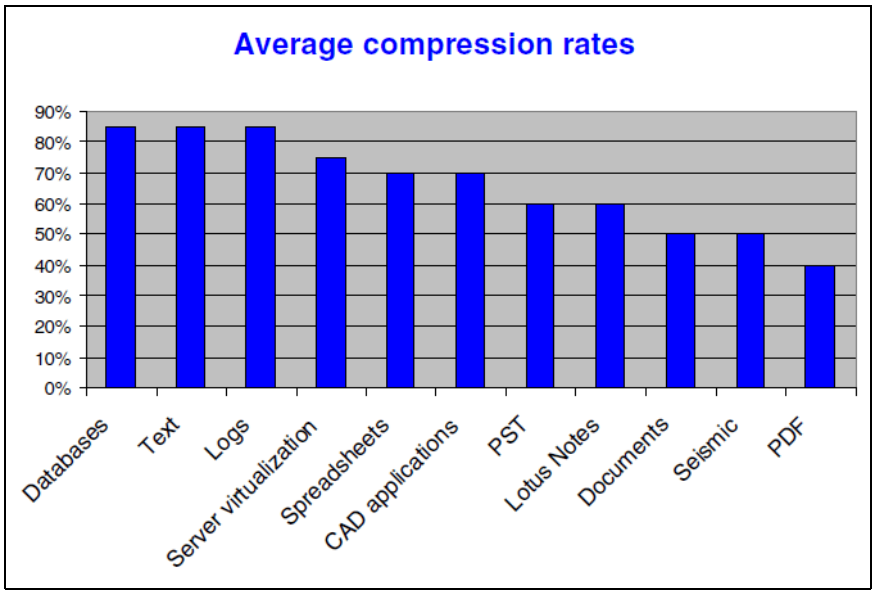


Figure 11-6 The average compression rates

Data deduplication processing can occur on the client, an infrastructure server, or the storage system. Each option has factors to consider:

- Client-based data deduplication

  This process reduces the amount of data that is transferred over the network to the storage system. But, this option can require extra CPU and disk I/O processing on the client side.

- Server-based data deduplication

  This process deduplicates the data of multiple clients at a scheduled time. But, this process requires extra CPU and disk I/O processing on the infrastructure server, for example, the IBM Tivoli Storage Manager server.

- Storage-based data deduplication

  This process occurs at the disk storage device level, where the data is stored. This type of data deduplication is generally transparent to the clients and servers. This process uses CPU and disk I/O on the storage system.

For more information about data compression, data deduplication, and concrete solutions from IBM, see *Introduction to IBM Real-time Compression Appliances*, SG24-7953, and *Implementing IBM Storage Data Deduplication Solutions*, SG24-7888.

12

IBM Fibre Channel storage area network product portfolio

This chapter guides you through the IBM storage area network (SAN) components that are offered through IBM marketing channels. The SAN products are either IBM original equipment manufacturer (OEM) products or IBM is an authorized reseller of the products.

12.1 Classification of IBM SAN products

To stay competitive in the global marketplace, the right people must have the right access to the right information at the right time 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 by using a wide range of tiered storage options, policy-based automation, and intelligent information management solutions. The IBM SAN portfolio offers the broadest range of storage solutions in the industry (including disk, tape, SAN, software, financial, and services offerings). You can use this portfolio to create long-term solutions that can be tailored to your business needs.

IBM SAN tiered disk, tape, and switch solutions provide various choices to align and move data to cost-optimized storage. This process is based on policies that match the storage solution with the service level requirements (SLAs) and the value of the data in growing environments.

You can confidently protect strategic information assets and efficiently comply with regulatory and security requirements with the unrivaled breadth of storage solutions from IBM. IBM SAN directors and routers provide metropolitan and global connectivity between sites.

IBM solutions are optimized for the unique needs of midsize organizations, large enterprises, cloud computing providers, and other businesses. You can get what you need, which saves 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 provide an in-depth analysis of all of the technical details of each product. The intention of this chapter is to introduce the principles and basic components of the SAN environments to a reasonable extent in a way that is easy to understand and follow.

The products are characterized in the following groups:

- Entry SAN switches
- Mid-range SAN switches
- Enterprise SAN directors
- Extension switches

For more information about other IBM storage products, see the *IBM System Storage Solutions Handbook*, SG24-5250.

For more information about each product and its market position, see this IBM storage website:

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

12.2 SAN Fibre Channel networking

This section describes the IBM products for Fibre Channel-based (optical) data center networking solutions, starting from entry-level SAN switches to mid-range switches up to enterprise SAN directors and multiprotocol routers.

For more information about the latest IBM SAN products, see this website:

http://www.ibm.com/systems/networking/switches/san/index.html

12.3 Entry SAN switches

Entry SAN switches represent easy-to-use preconfigured data center networking solutions for small and medium business environments. IBM offers the following products in this category:

- IBM Storage Networking SAN24B-6
- IBM System Networking SAN24B-5
- IBM System Storage SAN24B-4 Express
- Cisco MDS 9132T 32G Fabric Switch

12.3.1 IBM Storage Networking SAN24B-6

The IBM Storage Networking SAN24B-6 is an entry-level switch that combines high-performance capabilities of 4, 8, 16 and 32 Gbps with point-and-click simplicity and enterprise-class functionality. It provides small to midsized data centers with low-cost access to industry-leading Gen 6 Fibre Channel technology and the ability to start small and grow on demand—from 8 to 24 ports—to support an evolving storage environment. IBM b-type Gen 6 Fibre Channel products are designed to unleash the full potential of new storage technologies for the new high-performance application workloads.

IBM b-type Gen 6 (and Gen 5) technology leverages a rich heritage of Fibre Channel innovation to deliver industry-leading reliability for the world's most demanding data centers. Using Fabric Vision technology and VM Insight, administrators can quickly identify abnormal VM behaviors to facilitate troubleshooting and fault isolation, helping to ensure maximum performance and operational stability.

Fabric Vision technology provides visibility across the network with monitoring, management and diagnostic capabilities that enable administrators to avoid problems before they impact operations. Fabric Vision technology includes VM Insight, Monitoring and Alerting Policy Suite (MAPS), Fabric Performance Impact (FPI) Monitoring, dashboards, Configuration and Operational Monitoring Policy Automation Services Suite (COMPASS), ClearLink Diagnostics, Flow Vision, FEC and Credit Loss Recovery.

SAN24B-6 delivers industry-leading Gen 6 Fibre Channel technology in a flexible and easy-to-use solution that cost-effectively scales from 8 to 24 ports with PoD. It is easy to deploy with the EZSwitchSetup wizard, featuring a simple user interface that dramatically reduces deployment and configuration times with as few as three steps.

IBM Storage Networking SAN24B-6 can be deployed as a full-fabric switch or as an Access Gateway, simplifying fabric topologies and heterogeneous fabric connectivity (the default mode setting is a switch). Access Gateway mode4 utilizes N_Port ID virtualization (NPIV) switch standards to present physical and virtual servers directly to the core of storage area network (SAN) fabrics. This makes Access Gateway transparent to the SAN fabric, greatly reducing management of the network edge.

Figure 12-1 on page 250 shows the SAN24B-6.

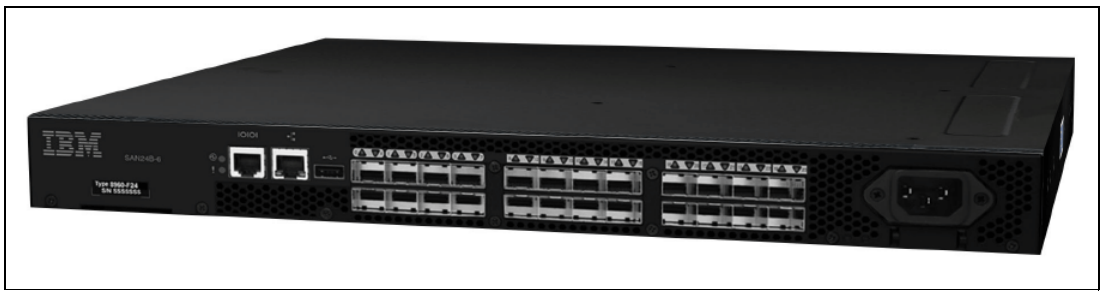


Figure 12-1 SAN24B-6

The IBM SAN24B-6 switch has the following highlights:

- Meet the high-throughput, low-latency demands of critical applications with flash-ready performance
- Scale on demand, from 8 to 24 ports, to connect additional devices as needed
- Deliver 4, 8, 16, or 32 Gbps port bandwidth for increased performance on demand
- Simplify deployment and reduce install time with a point-and-click user interface
- Automatically discover and recover from common networking problems
- Proactively monitor and optimize the health and performance of individual virtual machines (VMs)
- Simplify administration, resolve problems, increase uptime, and reduce costs by using Fabric Vision technology

12.3.2 IBM System Networking SAN24B-5

The IBM System Networking SAN24B-5 switch features Gen 5 Fibre Channel and Fabric Vision technologies, which provide outstanding price-to-performance value in an entry-level switch. The SAN24B-5 gives organizations the flexibility, simplicity, and enterprise-class functions they need to unleash the full potential of high-density server virtualization, cloud architectures, and next-generation storage.

The switch is configurable in 12 ports or 24 ports, and it supports 2 gigabits per second (Gbps), 4 Gbps, 8 Gbps, or 16 Gbps speeds in an efficiently designed 1U form factor. The switch includes a single power supply and integrated fans. A second optional power supply provides more redundancy for increased resiliency.

The SAN24B-5 (Figure 12-2) provides a critical building block for today's highly virtualized, private cloud storage environments. The SAN24B-5 can simplify server virtualization and virtual desktop infrastructure management while it meets the high-throughput demands of solid-state disks (SSDs). The SAN24B-5 supports multi-tenancy in cloud environments through quality of service (QoS) and fabric-based zoning features. It can also help minimize downtime in mission-critical environments by delivering high reliability, availability, and serviceability.

Figure 12-2 SAN24B-5

The SAN24B-5 includes the following features:

- Gen 5 Fibre Channel technology SAN switch
- Energy-efficient Fibre Channel SAN switch with 16 Gbps performance and up to 24 ports in a 1U form factor
- Ports on Demand (PoD) capability for scaling 12 ports - 24 ports in 12-port increments
- Autosensing of 2 Gbps, 4 Gbps, 8 Gbps, or 16 Gbps port speeds; 10 Gbps and optionally programmable to fixed port speed
- Sixteen Gbps optimized inter-switch links (ISLs)
- ClearLink diagnostic technology (D_ports) to identify optical and cable issues
- “Pay-as-you-grow” flexibility by using the 12-port Software Upgrade feature
- Two Gbps, 4 Gbps, 8 Gbps, 10 Gbps, or 16 Gbps speed on all ports
- Dual functionality as a full-fabric SAN switch or an N_Port ID Virtualization (NPIV)-enabled access gateway
- Ease of deployment and support of high-performance fabrics
- Innovative diagnostic, monitoring, and management capabilities through Fabric Vision technology

12.3.3 IBM System Storage SAN24B-4 Express

This system provides high-performance, scalable, and simple-to-use fabric switching by using 8 ports, 16 ports, or 24 ports that operate at 8 Gbps, 4 Gbps, 2 Gbps, or 1 Gbps (depending on which optical transceiver is used). This system is for servers that run Microsoft Windows, IBM AIX, UNIX, and Linux operating systems, server clustering, infrastructure simplification, and business continuity solutions. The SAN24B-4 Express includes an *EZSwitchSetup wizard*, which is an embedded setup tool that guides novice users through switch setup, often in less than 5 minutes. Figure 12-3 shows the front view of the SAN24B-4 switch.

Figure 12-3 Front view of the IBM System Storage SAN24B-4 Express switch

A single SAN24B-4 Express switch can serve as the cornerstone of a SAN for individuals that want to obtain the benefits of storage consolidation and implement Fibre Channel storage systems. This entry-level configuration can consist of one or two Fibre Channel links to a disk storage array or to a Linear Tape Open (LTO) tape drive. An entry-level, 8-port storage consolidation solution can support up to seven servers with a single path to disk or tape.

The Ports on Demand (PoD) feature enables a base switch to grow to 16 ports and 24 ports to support more servers and more storage devices without taking the switch offline. A high availability (HA) solution can be created with redundant switches. This capability is ideal for server clustering environments.

This configuration can support 6 - 22 servers, each with dual Fibre Channel adapters that are cross-connected to redundant SAN24B-4 Express switches. These switches are

cross-connected to a dual-controller storage system. The focus of the SAN24B-4 Express is as the foundation of small to medium-sized SANs.

However, the SAN24B-4 Express 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.

The SAN24B-4 includes the following features:

- Efficient 1U design with 8 ports, 16 ports, or 24 ports configuration on demand
- Auto-sensing 8 Gbps, 4 Gbps, 2 Gbps, or 1 Gbps ports that enable high performance and improved utilization while they provide easy installation and management
- Hot-swappable, small form-factor pluggables (SFPs)
- Inter-switch link (ISL) trunking for up to eight ports provides a total bandwidth of 128 Gbps
- Provides Fibre Channel interfaces expansion port (E_port), fabric port (F_port), fabric loop port (FL_port), and mirror port (M_port)
- Advanced, hardware-enforced zoning to protect against non-secure, unauthorized, and unauthenticated network and management access and worldwide name (WWN) spoofing
- Hot firmware activation for fast firmware upgrades that eliminate disruption to the fabric
- Compatibility with an earlier version with IBM b-type and m-type
- Optional as-needed licensed features:
  - Adaptive Networking
  - Advance Performance Monitor
  - Extended Fabric
  - Fabric Watch
  - Trunking Activation
  - Server Application Optimization (SAO)

12.3.4 Cisco MDS 9132T 32G Fabric Switch

The next-generation Cisco MDS 9132T 32G Fabric Switch for IBM® Storage Networking (Figure 12-4) provides high-speed Fibre Channel connectivity from the server rack to the SAN core. It empowers small, midsized and large enterprises that are rapidly deploying cloudscale applications using highly dense virtualized servers, by providing dual benefit of higher bandwidth and consolidation.

Figure 12-4 Cisco MDS 9132T 32G Fabric Switch

This switch has been designed to benefit both small-scale and large-scale SAN deployments. Small-scale SAN architectures can be built from the ground up using a low-cost, non-blocking, line-rate and low-latency fixed standalone SAN switch connecting both storage and host ports. Medium- to large-scale SAN architectures built with SAN core directors can

expand 32-Gbps connectivity to the server rack using these switches either in switch mode or Network Port Virtualization mode.

Additionally, investing in this switch in the server rack provides the day-one option of upgrading to 32-Gbps server connectivity using the 32-Gbps host bus adapters (HBAs) already available in the market. The Cisco MDS 9132T also provides unmatched flexibility through a unique port expansion module that provides a robust, cost-effective, field-swappable port upgrade option.

Among the main features of the Cisco MDS 9132T are high performance, high availability, pay-as-you-grow scalability and capital expenditure savings.

- High performance: MDS 9132T architecture, with chip-integrated non-blocking arbitration, provides consistent 32-Gbps low-latency performance across all traffic conditions for every Fibre Channel port on the switch.
- High availability: MDS 9132T switches continue to provide the same outstanding availability and reliability as previous generation Cisco MDS 9000 Family switches by providing optional redundancy on all major components such as the power supply and fan. Dual power supplies also facilitate redundant power grids.
- Pay-as-you-grow scalability: The MDS 9132T Fibre Channel switch provides an option to deploy as few as eight 32-Gbps Fibre Channel ports in the entry-level variant, which can grow by eight ports to 16 ports and thereafter with a port expansion module with sixteen 32-Gbps ports, to up to 32 ports. This approach results in lower initial investment and power consumption for entry-level configurations of up to 16 ports compared to a fully loaded switch. Upgrading through an expansion module also reduces the overhead of managing multiple instances of port activation licenses on the switch.
- Capital expenditure (CapEx) savings: The 32-Gbps ports allow users to deploy them on existing 16/8/4-Gbps transceivers, reducing CapEx with an option to upgrade to 32-Gbps transceivers and adapters as needed.

The new 32-Gbps fabric switches address the requirement for highly scalable, virtualized, intelligent SAN infrastructure in current-generation data center environments. The industry is already poised to transition to 32-Gbps fixed switches with the availability of 32-Gbps HBAs and storage arrays from vendors. Additionally, as low-latency flash arrays and highly dense virtualization deployments become more pervasive and as storage ports become 32-Gbps capable, fixed switches will need to provide 32-Gbps connectivity to the SAN core.

This solution offers several important benefits:

- Server port consolidation: The demand for 32-Gbps fabric switches is driven by hyperscale virtualizations that will significantly increase the virtual machine density per rack, and this growth will push the need for higher bandwidth HBA ports per rack of blade or standalone servers. One way to meet this demand is for 32-Gbps HBA ports to consolidate the current 16-Gbps HBA installed base to meet future needs to grow the number of ports. As a result, the MDS 9132T, with its lower port density, can be a preferred solution and its flexibility to grow can be an added advantage.
- Simplification: Through consolidation, a SAN administrator can reduce complexity and simplify management.
- Multiprotocol convergence: 32-Gbps links benefit from lower latency compared to lower-bandwidth links, bringing better-performing storage workloads to your storage array. Higher bandwidth also helps ensure less inter-switch link (ISL) congestion for newer storage protocols that are expected to be available on externally attached storage arrays; for instance: Fibre Channel Non-Volatile Memory Express (NVMe) can co-exist on the same link as existing SCSI workloads.
- Scale and performance: This small form-factor switch supports the performance and scale required to deploy a dedicated and standalone Fibre Channel SAN connecting both initiators and targets, without requiring any other switching infrastructure.

For more information about entry SAN switches, see this website:

https://www.ibm.com/storage/san#71626

12.4 Mid-range SAN switches

The IBM Mid-range SAN switches provide scalable and affordable small and medium business and enterprise solutions for storage networking.

The category of mid-range SAN switches includes the following products:

- Cisco MDS 9396S 16G Multilayer Fabric Switch
- IBM System Networking SAN96B-5
- IBM Storage Networking SAN64B-6
- IBM System Storage SAN48B-5
- Cisco MDS 9148S 16G Multilayer Fabric Switch for IBM System Storage

12.4.1 Cisco MDS 9396S 16G Multilayer Fabric Switch

The Cisco MDS 9396S 16G Multilayer Fabric Switch for IBM System Storage is the latest generation of the highly powerful, dense, and reliable Cisco MDS Series switches. This switch combines high performance with outstanding flexibility and cost-effectiveness. This robust, compact two rack-unit (2RU) switch scales from 48 to 96 line-rate 16 Gbps Fibre Channel ports.

The Cisco MDS 9396S is excellent for the following functions:

- A stand-alone SAN in large departmental storage environment
- A middle-of-the-row or top-of-the-rack (ToR) switch in medium-sized redundant fabrics
- An edge switch in enterprise data center core-edge topologies

Powered by Cisco NX-OS Software and Cisco Data Center Network Manager (DCNM) software, the Cisco MDS 9396S delivers advanced storage networking features and functions that combine with ease of management and compatibility with the entire Cisco MDS 9000 Family portfolio for reliable end-to-end connectivity. Figure 12-5 shows the Cisco MDS 9396S.

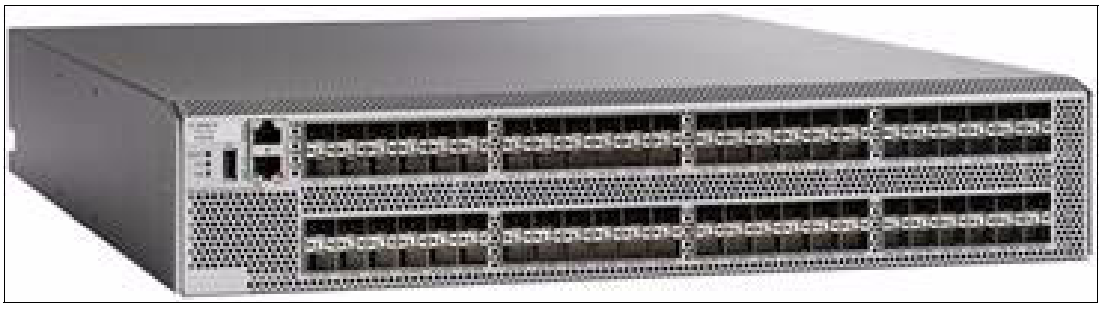


Figure 12-5 Cisco MDS 9396S

The Cisco MDS 9396S includes the following features:

- “Pay-as-you-grow” scalability in a high-density switch that supports up to ninety-six 16 Gbps Fibre Channel ports in a compact, two rack-unit (2RU) form factor.
- Autosensing Fibre Channel ports are provided that deliver up to 16 Gbps of high-speed, dedicated bandwidth for each port.

- Availability boost with In-Service Software Upgrades (ISSU), which enable the switch to be upgraded without affecting network traffic.
- Built-in storage network management and SAN plug-and-play capabilities.
- Virtual SAN (VSAN) technology that is used for hardware-enforced, isolated environments within a physical fabric.
- Extensive set of innovative and powerful security features that are provided in the optional Cisco MDS 9300 Family Enterprise Package.
- Up to 96 autosensing Fibre Channel ports can handle speeds of 2 Gbps, 4 Gbps, 8 Gbps, and 16 Gbps, with 16 Gbps of dedicated bandwidth for each port.
- Base switch includes 48 ports that are enabled.
- Dual-redundant hot-swappable power supplies and fan trays, PortChannels, and F_port channeling.

12.4.2 IBM System Networking SAN96B-5

The IBM System Networking SAN96B-5 switch is a high-density, purpose-built, foundational building block for large and growing SAN infrastructures. This switch provides highly resilient, scalable, and simplified network infrastructure for storage. By delivering market-leading, Gen 5 Fibre Channel technology and capabilities with 16 Gbps performance, the SAN96B-5 meets the demands of growing, dynamic workloads; evolving, virtualized data centers; and highly virtualized private and hybrid cloud storage environments (Figure 12-6).

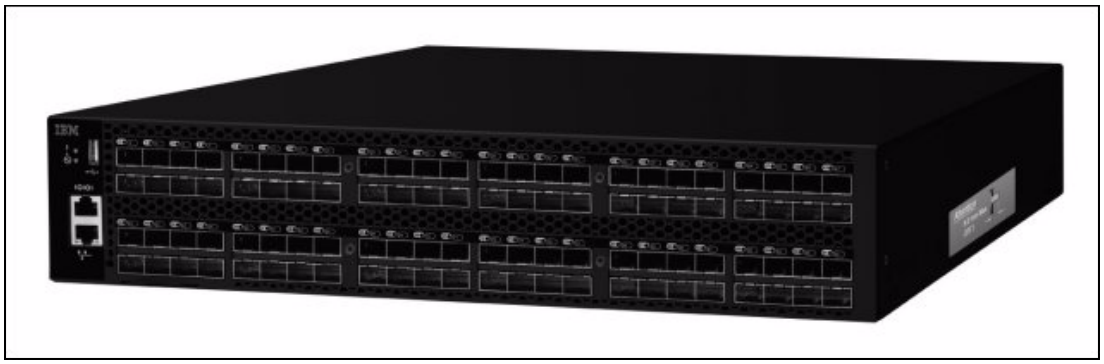


Figure 12-6 SAN96B-5

The SAN96B-5 includes the following features:

- Supports virtualized private and hybrid cloud storage environments and data center consolidation with high scalability in an ultra-dense, Gen 5 Fibre Channel 96-port switch
- Enables “pay-as-you-grow” flexibility by using the Ports on Demand (PoD) feature with speeds up to 16 Gbps
- Simplifies management with Fabric Vision technology, which helps reduce operational costs and optimize application performance
- Provides data center-to-data center security and bandwidth savings with up to eight in-flight encryption and compression ports
- Helps maximize application uptime and performance while it helps reduce expenses with ClearLink diagnostic technology (D_ports)
- Optimizes link and bandwidth utilization with ISL Trunking and Dynamic Path Selection
- Accelerates deployment and troubleshooting with Dynamic Fabric Provisioning, monitoring, and advanced diagnostics

- Reduces operational costs and complexity, simplifying and centralizing management with IBM Network Advisor
- Supports highly virtualized, private, and hybrid cloud storage and data center consolidation with high scalability in an ultra-dense Gen 5 Fibre Channel 96-port switch
- Offers 2 Gbps, 4 Gbps, 8 Gbps, 10 Gbps, or 16 Gbps speed on all ports
- Provides innovative diagnostic, monitoring, and management capabilities through Fabric Vision technology
- Offers data center-to-data center security and bandwidth savings
- Supports the latest hot aisle/cold aisle configurations
- Optimizes link and bandwidth utilization
- Provides comprehensive management of data center fabrics, including configuration, monitoring, and management of IBM b-type backbones, switches, and adapters
- Offers 16 Gbps performance with up to 96 ports in a 2U, enterprise-class Fibre Channel SAN switch
- Provides Ports on Demand (PoD) feature capability and can scale 48 ports - 96 ports in 24-port increments
- Autosenses 2 Gbps, 4 Gbps, 8 Gbps, or 16 Gbps port speeds; 10 Gbps and optionally programmable to a fixed port speed
- Offers flexible, high-speed 16 Gbps and 8 Gbps optics
- Supports up to eight in-flight encryption and compression ports
- Provides two models with different airflow options (2498-F96 delivers non-port-side to port-side airflow; 2498-N96 delivers port-side to non-port-side airflow)
- Offers ISL Trunking and Dynamic Path Selection
- Supports IBM Network Advisor

12.4.3 IBM Storage Networking SAN64B-6

The IBM Storage Networking SAN64B-6 switch meets the demands of hyper-scale virtualization, larger cloud infrastructures, and growing flash-based storage environments by delivering market-leading Gen 6 Fibre Channel technology and capabilities.

Figure 12-7 shows the IBM Storage Networking SAN64B-6 switch.

Figure 12-7 IBM Storage Networking SAN64B-6 switch

The SAN64B-6 switch includes the following features:

- Gen 5 Fibre Channel switch with 16 Gbps performance and up to 48 ports in an energy-efficient, 1U form factor
- 2, 4, 8, 10, or 16 Gbps speed on all ports
- PoD capability for scaling from 24 to 48 ports in 12-port increments

- 16 Gbps optimized Inter-Switch Links (ISLs)
- 128 Gbps high-performance and resilient frame-based trunking
- Fabric Vision technology capabilities
- Improved flexibility and investment protection
- Aggregate of 768 Gbps full-duplex throughput
- Meets the demands of hyper-scale private or hybrid cloud storage environments
- Helps maximize uptime, simplifies SAN management, and provides outstanding visibility and insight across the storage network
- Multitenancy in cloud environments through Virtual Fabrics, Integrated Routing, quality of service (QoS), and fabric-based zoning features
- 10 Gbps Fibre Channel integration on the same port
- In-flight data compression and encryption

12.4.4 IBM System Storage SAN48B-5

The SAN48B-5 switch meets 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, the SAN48B-5 can address demanding reliability, availability, and serviceability (RAS) requirements to help minimize downtime to support mission-critical cloud environments.

Figure 12-8 shows the front view of the IBM System Storage SAN48B-5 16 Gbps Fibre Channel switch.

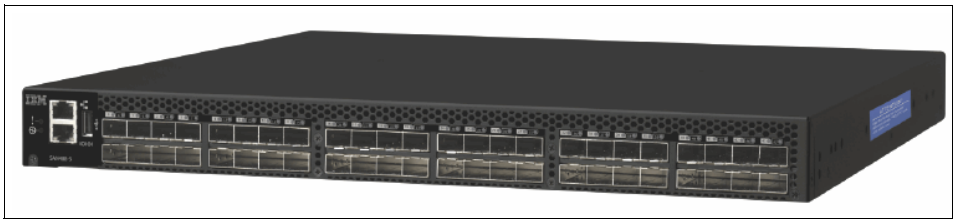


Figure 12-8 IBM System Storage SAN48B-5 fabric switch

The SAN48B-5 includes the following features:

- Performance of 16 Gbps with up to 48 ports in an energy-efficient, 1U enclosure
- Speed of 2 Gbps, 4 Gbps, 8 Gbps, 10 Gbps, or 16 Gbps on all ports that produce an aggregate 768 Gbps full-duplex throughput
- High-performance of 128 Gbps and resilient frame-based trunking
- Ten Gbps Fibre Channel integration on the same port for dense wavelength division multiplexing (DWDM) metropolitan connectivity on the same switch
- In-flight data compression and encryption for efficient link utilization and security
- Redundant, hot-swap components and nondisruptive 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, QoS, and fabric-based zoning features

12.4.5 Cisco MDS 9148S 16G Multilayer Fabric Switch for IBM System Storage

The Cisco MDS 9148S 16G Multilayer Fabric Switch for IBM System Storage (Figure 12-9 on page 258) is the latest generation of the highly reliable, flexible, and low-cost Cisco MDS 9100 Series switches. It combines high performance with exceptional flexibility and cost-effectiveness. This powerful, compact one rack-unit (1RU) switch scales from 12 to 48 line-rate 16 Gbps Fibre Channel ports.

The Cisco MDS 9148S is excellent for the following functions:

- A stand-alone SAN in small departmental storage environments
- A top-of-the-rack switch in medium-sized redundant fabrics
- An edge switch in enterprise data center core-edge topologies

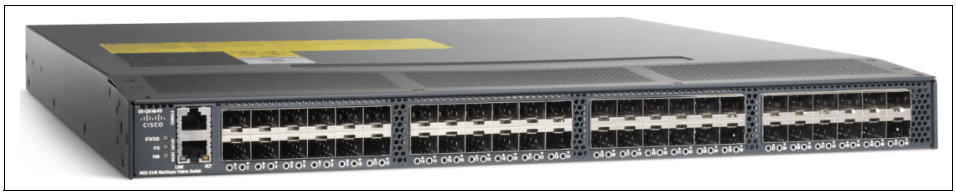


Figure 12-9 Cisco MDS 9148S for IBM System Storage

The Cisco MDS 9148S includes the following features:

- High performance and flexibility
- High-availability platform
- Up to 48 autosensing Fibre Channel ports that are capable of speeds of 2 Gbps, 4 Gbps, 8 Gbps, and 16 Gbps, with 16 Gbps of dedicated bandwidth for each port
- “Pay-as-you-grow” scalability with 12 - 48 high-performance Fibre Channel ports in a single 1RU switch
- Dual redundant hot-swappable power supplies and fan trays, PortChannels, and F_port channeling

For more information about IBM mid-range SAN switches, see this website:

`https://www.ibm.com/storage/san#71516`

12.5 Enterprise SAN directors

IBM Enterprise SAN directors provide the data center networking infrastructure with enterprise solutions for the highest availability and scalability.

IBM offers the following enterprise SAN directors through its marketing channels:

- IBM Storage Networking SAN512B-6 and SAN256B-6
- Cisco MDS 9718 Multilayer Director
- Cisco MDS 9710 Multilayer Director
- IBM System Storage SAN768B-2 and SAN384B-2
- Cisco MDS 9706 Multilayer Director for IBM System Storage

12.5.1 IBM Storage Networking SAN512B-6 and SAN256B-6

Designed to meet relentless growth and mission-critical application demands, the IBM Storage Networking SAN512B-6 and SAN256B-6 b-type Gen 6 directors are ideal platforms for large enterprise environments that require increased capacity, greater throughput, and higher levels of resiliency. SAN512B-6 and SAN256B-6 directors with Fabric Vision technology are modular building blocks that combine innovative hardware, software, and built-in instrumentation to ensure high levels of operational stability and redefine application performance. Fabric Vision technology enhances visibility into the health of storage environments, delivering greater control and insight to quickly identify problems and achieve critical SLAs.

Breakthrough 32/128 Gbps performance shatters application performance barriers and provides support for more than 1 billion I/O operations per second (IOPS) for flash-based storage workloads.

Figure 12-10 on page 259 shows the SAN512B-6.

Figure 12-10 SAN512B-6

Figure 12-11 shows the SAN256B-6.

Figure 12-11 SAN256B-6

The SAN512B-6 and SAN256B-6 directors include the following features:

- Enhance operational stability, maximize application performance, and increase business agility with enterprise-class b-type Gen 6 directors
- Shatter application performance barriers across 32 Gbps links, and support up to 1 billion IOPS without oversubscription
- Consolidate infrastructure with high-density solutions built with 128 Gbps UltraScale Inter-Chassis Link (ICLs) connectivity for simpler, flatter, low-latency fabrics
- Optional 128 Gbps shortwave QSFP and 4 x 32Gbps 2 km QSFP to deliver additional Inter-Chassis Links (ICL) connectivity options
- Simplify end-to-end management of large-scale environments by automating monitoring and diagnostics
- Detect degraded application or device performance with built-in monitoring
- Extend distance and replication with a highly scalable, multiprotocol extension solution
- Simplify configuration automation and enable integrated advanced services across the fabric with standard Representational State Transfer (REST) application programming interfaces (APIs)
- Integrate next-generation flash storage based on non-volatile memory express (NVMe) flash memory with current and future b-type Gen 6 Fibre Channel networks
- Mitigate risk with compatibility with earlier versions

The IBM b-type Gen 6 Extension Blade includes the following features:

- Accelerates data replication across data centers to meet recovery objectives and secure data flows over distance
- Consolidates Fibre Channel and IP storage replication traffic within a single blade with flexible multiprotocol port connectivity
- Connects more Fibre Channel and IP devices with industry-leading port density and scale-as-you-grow flexibility
- Centralizes management of Fibre Channel and IP extension for storage traffic while extending Fabric Vision technology over distance for greater control and insight

- Extends proactive monitoring and alerting between data centers to automatically detect WAN anomalies and simplify troubleshooting of end-to-end I/O flows over distance, avoiding unplanned downtime
- Provides load balancing and network resilience with Extension Trunking and Adaptive Rate Limiting to increase WAN utilization and protect against WAN link failures
- Achieves always-on business operations with nondisruptive firmware upgrades and maximizes availability with redundant, hot-pluggable chassis components

12.5.2 Cisco MDS 9718 Multilayer Director

Cisco MDS 9718 Multilayer Director for IBM Storage Networking addresses the mounting storage requirements of today's large virtualized data centers. It has the industry's highest port density for a SAN director, featuring 768 line-rate 16 Gbps Fibre Channel ports. Designed to support multiprotocol workloads, MDS 9718 enables SAN consolidation and collapsed-core solutions for large enterprises, thereby reducing the number of managed switches and leading to easy-to-manage deployments. By reducing the number of front-panel ports used on ISLs, it also offers room for future growth.

As a director-class SAN switch, MDS 9718 uses the same operating system and management interface as other Cisco data center switches. It brings intelligent capabilities to a high-performance, protocol-independent switch fabric—delivering uncompromising availability, security, scalability, simplified management, and the flexibility to integrate new technologies. MDS 9718 lets you transparently deploy unified fabrics with Fibre Channel and FCoE connectivity to achieve low total cost of ownership (TCO).

For mission-critical enterprise storage networks that require secure, robust, cost-effective business-continuance services, the Fibre Channel over IP (FCIP) extension module delivers outstanding SAN extension performance. It provides features that reduce latency for disk and tape operations with FCIP acceleration, including FCIP write acceleration and FCIP tape write and read acceleration.

Figure 12-12 shows the MDS 9718.

Figure 12-12 MDS 9718

The MDS 9718 includes the following features:

- Enables up to 1.5 Tbps of Fibre Channel front-panel bandwidth per module and 768 full line-rate (2/4/8/16 Gbps) autosensing Fibre Channel ports in a single chassis
- Provides up to 768 full line-rate (10 Gbps) autosensing Fibre Channel over Ethernet (FCoE) ports in a single chassis
- Provides up to 384 full line-rate (40 Gbps) FCoE ports in a single chassis
- Provides 16 Gbps line rate, non-blocking, and predictable performance across all traffic conditions for every port in the chassis
- Enables large and scalable deployment of SAN extension solutions through the SAN extension module
- Offers nondisruptive software upgrades, stateful process restart and failover, and full redundancy of major components
- Enables migration from SAN islands to enterprise-wide storage networks through virtual SAN (VSAN) technology, access control lists for hardware-based intelligent frame processing, and fabric-wide QoS
- Provides virtual machine transparency
- Provides deterministic hardware performance and a comprehensive feature set that allows virtual machines to have the same SAN attributes as a physical server
- Cisco Data Center Network Manager (DCNM) for SAN provides end-to-end visibility from the virtual machine down to storage

- Supports a comprehensive security framework
- Provides unified SAN management
- Includes built-in storage network management with all features available through a command-line interface (CLI) or DCNM
- Provides sophisticated diagnostics
- Provides intelligent diagnostics, protocol decoding, and network analysis tools; and it provides integrated Call Home capability for added reliability, faster problem resolution, and reduced service costs
- Reduces total cost of ownership
- Accommodates future mission-critical applications, massive amounts of data, and cloud environments
- Features multiprotocol intelligence
- Enables a consistent feature set over a protocol-independent switch fabric

12.5.3 Cisco MDS 9710 Multilayer Director

The Cisco MDS 9710 Multilayer Director for IBM SystemNetworking (9710-E08) is a director-class storage area network (SAN) switch that is designed for deployment in large-scale storage networks to enable enterprise clouds and business transformation by adding enterprise connectivity options that support IBM Fibre Connection (IBM FICON®) connectivity.

MDS 9710 delivers a high performing and reliable FICON infrastructure that is designed to support fast and scalable IBM z Systems™ servers.

With the Cisco MDS 9700 48-Port 10-Gbps Fibre Channel over Ethernet (FCoE) Module and the Cisco MDS 9700 24-port 40-Gbps FCoE Module, the MDS 9700 platforms provide multiprotocol flexibility for SANs that delivers 16 Gbps FC and 10/40 Gbps FCoE capability and 1/10 FCIP capabilities.

Layering a comprehensive set of intelligent features onto a high-performance, protocol-independent switch fabric, the MDS 9710 addresses the stringent requirements of large virtualized data center storage environments: High availability, security, scalability, ease of management, and transparent integration of new technologies for flexible data center SAN solutions. Sharing the operating system and management interface with other Cisco data center switches, the MDS 9710 enables seamless deployment of fabrics with high-performance Fibre Channel, IBM FICON, FCoE, and Fibre Channel over IP (FCIP) connectivity to achieve low total cost of ownership (TCO).

For mission-critical enterprise storage networks that require secure, robust, cost-effective business-continuance services, the FCIP extension module is designed to deliver outstanding SAN extension performance, reducing latency for disk and tape operations with FCIP acceleration features, including FCIP write acceleration and FCIP tape write and read acceleration. (see Figure 12-13 on page 264).

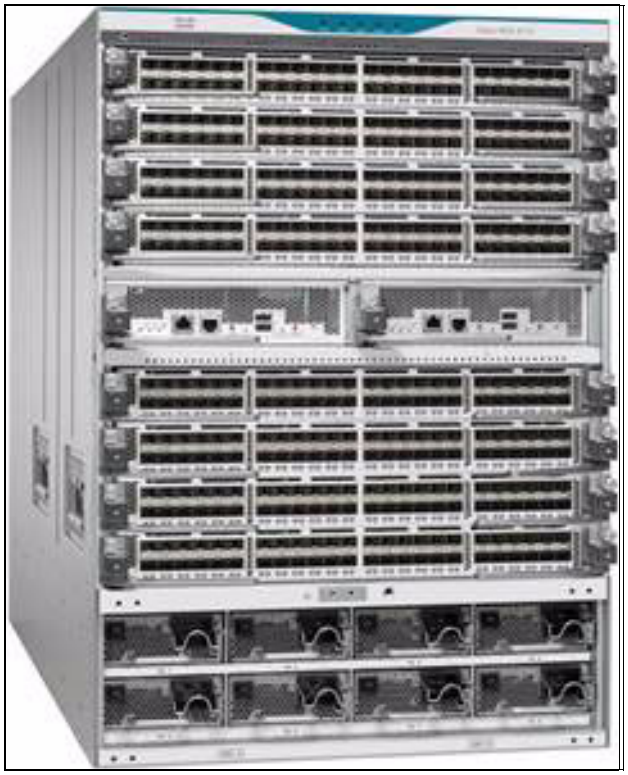


Figure 12-13 MDS 9710

The MDS 9710 includes the following features:

- Up to 24-Tbps front-panel Fibre Channel switching capacity
- 1.5-Tbps front-panel Fibre Channel performance per slot
- Industry-leading port densities of up to 384 32/16/10/8/4/2-Gbps autosensing line-rate Fibre Channel, or 384 10-Gbps FcoE, or 192 40-Gbps FCoE ports per chassis
- Up to 64 10-Gbps FCIP ports and 16 40-Gbps FCIP ports
- Integrated hardware-based analytics support with 32G module
- 48-port Fibre Channel module supporting 4/8/16/32-Gbps autosensing (optionally configurable) Fibre Channel ports
- 48-port FC module supporting 2/4/8/16-Gbps autosensing (optionally configurable) and 10-Gbps fixed rate FC ports
- 48-port FCoE module with 10-Gbps FCoE ports
- 24-port FCoE module with 40-Gbps FCoE ports
- 24/10 SAN extension module with 24 ports at 2/4/8/10/16 Gigabit Ethernet Fibre Channel, 8 ports at 1/10 GE, and 2 ports at 40 GE for FCIP
- NVMe over Fibre Channel support on all ports
- Virtual SAN (VSAN)
- Inter-VSAN Routing (IVR)
- PortChannel with multipath load balancing
- Flow- and zone-based quality of service (QoS)
- N_Port ID Virtualization (NPIV)
- Integrated Analytics
- Online, nondisruptive software upgrades
- Stateful, nondisruptive supervisor module failover
- Hot-swappable switching modules, supervisor modules, fans, power supplies, and small form-factor pluggables
- Front-to-back airflow

12.5.4 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 for evolving enterprise data centers.

Each system combines breakthrough performance, scalability, and energy efficiency with long-term investment protection.

These systems support open systems and IBM z Systems environments and address data growth and server virtualization challenges to achieve the following benefits:

- 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 512 sixteen-Gbps FC device ports. The SAN384B-2 is ideal for midsize core or edge deployments. The SAN384B-2 fabric backbone provides four horizontal blade slots and up to 256 sixteen-Gbps FC device ports. The flexible blade architecture also supports FCoE, in-flight data compression and encryption, SAN extension advanced functionality for high-performance servers, I/O consolidation, data protection, and disaster recovery solutions.

- The benefits that are associated with higher throughput are compelling. The IBM System Storage Gen 5 Fibre Channel b-type SAN products offer additional advantages to meet new and evolving requirements:
- Low latency and high I/O operations per second (IOPS) performance maximize the number of virtual hosts for each physical server.
- Data center-proven, purpose-built architectures minimize the risk and fault domains of high-density server virtualization.
- Nonstop networking and automated management minimize operational cost and complexity.
- Integrated advanced diagnostics, monitoring, and reliability, availability, and serviceability (RAS) capabilities simplify management and increase resiliency.
- Integrated ISL data compression and encryption offer bandwidth optimization and data protection.
- Compatibility with existing infrastructure minimizes the need to “rip out and replace” equipment.
- Low overhead and low latency eliminate I/O bottlenecks and unleash the full performance of Flash, SSD, and 16 Gbps-capable storage.

In addition, Brocade Gen 5 Fibre Channel platforms offer breakthrough technologies that dramatically simplify SAN deployment and management and drive down operational costs:

- UltraScale chassis connectivity enables higher density and simpler fabrics that reduce network complexity and cost.
- Fabric Vision technology maximizes uptime, optimizes application performance, and simplifies SAN management through innovative diagnostic, monitoring, and management technology.

The SAN768B-2 and SAN384B-2 are efficient at reducing power consumption, cooling, and the carbon footprint in data centers. Although these switches provide exceptional performance and scale, these networking backbones use less than 0.2 watts/Gbps.

As members of the IBM System Storage family of b-type SAN products, the SAN768B-2 and the SAN384B-2 participate in fabrics that contain other b-type and m-type devices that are 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.

Figure 12-14 shows the directors.

Figure 12-14 IBM System Storage SAN384B-2 (left) and SAN768B-2 (right)

The following blades are available for both models:

- Sixteen Gbps 32-port, 48-port, and 64-port FC blades (Feature Code (FC) 3632, FC 3648, and FC 3664)
- Eight Gbps 64-port FC blades (FC 3864)
- FCoE blade that supports twenty-four 10 Gbps Converged Enhanced Ethernet (CEE)/FCoE ports (FC 3880)
- Enhanced Extension blade that supports twelve 8-Gbps FC ports and ten 1-Gbps Gigabit Ethernet (GbE) ports, or two optional 10-Gbps GbE ports (FC 3891)

The SAN768B-2 and SAN384B-2 include the following features:

- Redundant hot-swappable control processor modules, power supplies, and cooling.
- Autosensing 16 Gbps, 8 Gbps, 4 Gbps, 2 Gbps E_port, F_port, FL_port, EX_port, and 10 Gbps E_port Fibre Channel interfaces and 10 Gbps converged Ethernet ports.
- Sixteen Gbps, 8 Gbps, 4 Gbps, and 2 Gbps FICON interfaces, FICON with control unit port (CUP) capability (FC 7893).
- One GbE and 10 GbE Fibre Channel over Internet Protocol (FCIP) Ethernet interfaces for Channel Extension.
- Advanced Extension support for FCIP Trunking and Adaptive Rate Limiting (FC 7891).
- FICON Acceleration for XRC (Extended Remote Copy) and Tape R/W pipelining over FCIP (FC 7893).
- Integrated Fibre Channel routing (FCR) (FC 7890).
- In-Flight Encryption and Compression at line rate for both FC and FICON E_port links.

- UltraScale Inter-Chassis Links (ICL) that provide up to 128 sixteen-Gbps Chassis-to-Chassis Optical backplane ports. The SAN768B-2 has 32 x Quad small form-factor pluggable (QSFP) (4 x 16 Gbps) ports and the SAN384B-2 has 16 QSFP (4 x 16 Gbps) ICL ports. QSFPs are available for 100 m (328.08 ft) over OM4 and 2 km (1.24 miles) over SMF.
- Throughput of up to 10.2 Tbps per chassis with 512 sixteen-Gb FC device ports, plus 128 sixteen-Gb ICL ports.
- Cut-through FC routing for lowest latency.
- Local switching on each FC blade is supported, providing switched latency of 700 ns.
- Blade-to-blade latency 2.1 microseconds.
- Other Base System features include the following features:
  - Advanced Diagnostic tools
  - Bottleneck Detection
  - Buffer Credit recovery
  - ClearLink Optical path diagnostics
  - Dynamic Fabric provisioning
  - ISL trunking
  - Server Application Optimization (SAO)
  - Virtual Fabrics
  - Ingress Rate Limiting
  - Traffic Isolation
  - QoS
- Fabric Vision incorporates MAPS to provide these functions:
  - Policy-based threshold monitoring and alerting
  - Fabric Performance Impact (FPI) monitoring to automatically detect and alert administrators to severe levels of latency and to identify slow drain devices
  - Flow Vision to identify, monitor, and analyze specific application flows, including the ability to automatically learn flows
  - Flow Generation to provide a built-in traffic generator for pre-testing and validating the data center infrastructure at full line-rate before you deploy applications
- Management protocols that include Simple Network Management Protocol (SNMP) v1/v3 (FE Management Information Base (MIB) and FC Management MIB), Storage Management Interface Specification (SMI-S), and OpenStack FC zone management.
- Support for security:
  - Advanced Encryption Standard (AES)-GCM-256 encryption on ISLs
  - Diffie-Hellman (D-H) Challenge Handshake Authentication Protocol (DH-CHAP) (between switches and end devices)
  - Fault, configuration, accounting, performance, security (FCAP) switch authentication
  - Federal Information Processing Standard (FIPS) 140-2 L2-compliant
  - Hypertext Transfer Protocol Secure (HTTPS)
  - Internet Protocol Security (IPsec)
  - Internet Protocol (IP) filtering
  - Lightweight Directory Access Protocol (LDAP) with IPv6
  - OpenLDAP
  - Port Binding

- Remote authentication dial-in user service (RADIUS)
- User-defined role-based access control (RBAC)
- Secure Copy Protocol (SCP)
- Secure Remote Procedure Call (RPC)
- Secure Shell (SSH) File Transfer Protocol (also Secure File Transfer Protocol) SFTP
- SSHv2
- Secure Sockets Layer (SSL)
- Switch Binding
- Terminal Access Controller Access Control System+ (TACACS+)
- Trusted Switch

- Management and operational software:
  - Auditing
  - Syslog
  - Command-line interface (CLI)
  - Browser-based Web Tools
  - IBM Network Advisor
- Full compatibility with an earlier version with IBM System Storage and IBM TotalStorage b-type and m-type SAN directors, switches, and routers; other directors, switches, and routers that are manufactured by Brocade.

12.5.5 Cisco MDS 9706 Multilayer Director for IBM System Storage

Cisco MDS 9706 Multilayer Director for IBM System Storage is a director-class SAN switch designed for deployment in small to midsized storage networks that can support enterprise clouds and business transformation (see Figure 12-15 on page 269). It layers a comprehensive set of intelligent features onto a high-performance, protocol-independent switch fabric.

MDS 9706 addresses the stringent requirements of large virtualized data center storage environments. It delivers uncompromising availability, security, scalability, ease of management, and transparent integration of new technologies for extremely flexible data center SAN solutions. It shares the same operating system and management interface with other Cisco data center switches. MDS 9706 lets you transparently deploy unified fabrics with Fibre Channel, FICON, FCoE, and Fibre Channel over IP (FCIP) connectivity for low total cost of ownership (TCO).

For mission-critical enterprise storage networks that require secure, robust, cost-effective business-continuance services, the FCIP extension module delivers outstanding SAN extension performance. It reduces latency for disk and tape operations with FCIP acceleration features, including FCIP write acceleration and FCIP tape write and read acceleration.

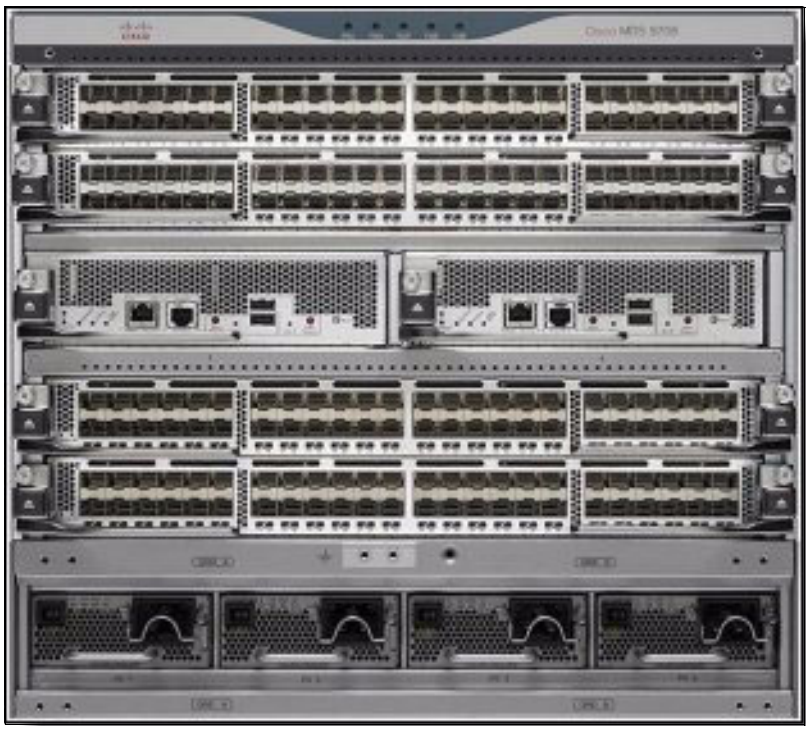


Figure 12-15 MDS 9706

The Cisco MDS 9706 includes the following features:

- Delivery of up to 12 Tbps front-panel Fibre Channel switching bandwidth
- 16 Gbps line-rate, non-blocking and predictable performance across all traffic conditions for every Fibre Channel and FCoE port in chassis
- Integrated mainframe support
- Deterministic hardware performance and features that allow virtual machines to have the same SAN attributes as physical servers
- Enablement of outstanding levels of availability and reliability
- Modular, multilayer, highly available, dual supervisor modules with six fabric slots and four module slots (9RU)
- 192 full line-rate (2/4/8, 4/8/16 Gbps and 10 Gbps) autosensing Fibre Channel ports in single chassis for deployment in open systems
- 192 full line-rate (10 Gbps) autosensing FCoE ports in a single chassis
- 96 full line-rate (40 Gbps) autosensing FCoE ports in a single chassis
- 48-port 16 Gbps Fibre Channel switching module
- 48-port 10 Gbps FCoE switching module
- 24-port 40-Gbps FCoE switching module
- 24/10-port SAN extension module
- Hot-swappable switching modules, supervisor modules, fans, power supplies and small form-factor pluggables
- Front-to-back airflow

For more information, see this website:

https://www.ibm.com/storage/san#71396

12.6 Extension switches

The following IBM extension switches are available:

- IBM System Storage SAN42B-R extension switch
- Cisco MDS 9250i Multiservice Fabric Switch for IBM System Storage
- IBM System Storage SAN06B-R multiprotocol router

12.6.1 IBM System Storage SAN42B-R

IT organizations continually face two key issues: The unrelenting growth of data being transferred between data centers and the changes driven by virtualized application workloads within Fibre Channel and FICON storage environments. Also faced with rising SLA requirements and recovery expectations, enterprise data centers need their disaster-recovery infrastructure to help ensure fast, continuous, and simple access to mission-critical data from anywhere in the world.

The IBM System Storage SAN42B-R extension switch (Figure 12-16) is a purpose-built extension solution that securely moves more data faster over distance, while minimizing the impact of disruptions. With enhanced extension capability and Fabric Vision technology, SAN42B-R delivers outstanding performance, strong security, continuous availability, and simplified management. These features enable it to handle the unrelenting growth of data traffic between data centers in Gen 5 and Gen 6 Fibre Channel and FICON storage environments.

In addition, SAN42B-R helps storage administrators replicate and back up large amounts of data over wide area network (WAN) quickly, securely, reliably, and simply while minimizing operational and capital expenses.

Figure 12-16 SAN42B-R

The SAN42B-R includes the following features:

- Leading throughput performance and port density required for maximum application throughput over wide area network (WAN) links
- Support for Fibre Channel, IBM FICON and IP storage traffic over IP links allowing hardware consolidation for open-systems, mainframe, storage, and tape extension solutions
- “Pay-as-you-grow” scalability with capacity-on-demand upgrades
- Extension Trunking for improving load balancing and network resilience against WAN link failure

- Non-disruptive firmware upgrade
- Advanced performance and network optimization features including data compression, disk protocol acceleration, QoS and storage optimized TCP
- Outstanding insight and visibility across the storage extension network through Fabric Vision technology
- Twenty-four 16 Gbps Fibre Channel ports
- Sixteen 1/10-Gigabit Ethernet (GbE) ports
- Two 40-Gigabit Ethernet (GbE) ports
- Base configuration including a comprehensive set of advanced services and IBM Network Advisor

12.6.2 Cisco MDS 9250i Multiservice Fabric Switch

The Cisco MDS 9250i Multiservice Fabric Switch for IBM System Storage is an optimized platform for deploying high-performance SAN extension solutions, distributed intelligent fabric services, and cost-effective multiprotocol connectivity for open systems and mainframe environments. With a compact form factor and advanced capabilities that are normally available on director-class switches only, the MDS 9250i is an ideal solution for departmental and remote branch-office SANs and large-scale SANs with the Cisco MDS 9710 Multilayer Director.

The MDS 9250i (Figure 12-17 on page 271) offers up to forty 16 Gbps Fibre Channel ports, two 1/10 Gigabit Ethernet IP storage services ports, and eight 10-Gigabit Ethernet FCoE ports in a fixed, two-rack-unit (2RU) form factor. The MDS 9250i connects to native Fibre Channel networks, which protects your current investments in storage networks.

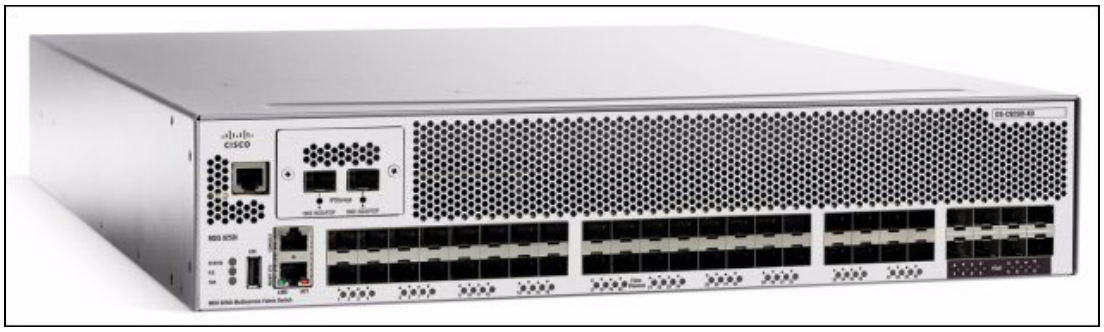


Figure 12-17 MDS 9250i

The MDS 9250i includes the following features:

- Hardware-based virtual fabric isolation with VSANs and FC routing with Inter-VSAN Routing (IVR).
- Enterprise-class availability features for departmental SANs.
- Intelligent network services and advanced traffic management capabilities.
- Industry-leading SAN security to support compliance and regulatory requirements.
- Fixed configuration with up to 40 ports of 16-Gbps Fibre Channel and 10 ports of 1/10-Gigabit Ethernet.
- Four-Gbps, 8-Gbps, and 16-Gbps autosensing, optionally configurable port speed.

- Up to 256 buffer-to-buffer credits can be assigned to a single Fibre Channel port to extend storage networks over long distances.
- Hot-swappable, 2+1 redundant power supplies and fans.

12.6.3 IBM System Storage SAN06B-R

The IBM System Storage SAN06B-R extension switch uses advanced Fibre Channel and Fibre Channel over IP (FCIP) technology to provide a fast, highly reliable, cost-effective network infrastructure for remote data replication, backup, and migration. Whether it is configured for simple point-to-point or comprehensive multi-site SAN extension, the SAN06B-R extension switch addresses the most demanding business continuity, compliance, and global data access requirements.

Up to sixteen 8-Gbps Fibre Channel ports and six 1-Gigabit Ethernet (GbE) ports provide the Fibre Channel and FCIP bandwidth, port density, and throughput that are required to help maximize application performance over WAN links (Figure 12-18).

Figure 12-18 SAN06B-R

The SAN06B-R includes the following features:

- High performance design with up to 8 Gbps Fibre Channel ports and hardware-assisted traffic processing for up to 1 Gbps line-rate performance across GbE.
- Intranet, IP-based metropolitan area network (MAN), or WAN infrastructures for metro and global SAN extensions can be used for business continuity solutions.
- Integrated IBM System Storage SAN b-type switch management helps simplify installation and administration and helps provide fabric investment protection.
- One-unit (1U) 19-inch packaging that is designed for rack mount or tabletop.
- Up to 8 Gbps Fibre Channel ports and up to 1 GbE port.
- Support for 8-Gbps, 4-Gbps, and 2-Gbps Fibre Channel link speeds or 4-Gbps, 2-Gbps, and 1-Gbps Fibre Channel link speeds.
- Speed autosensing capabilities provide compatibility with an earlier version with 4-Gbps, 2-Gbps, and 1-Gbps Fibre Channel links.
- Shortwave and longwave small form-factor pluggables can be intermixed in the same router to meet your unique requirements.
- Full Fabric operation and Universal Port Operation are available on all ports.
- Optional inter-switch link (ISL) trunking is supported on all active ports.
- Hardware-based compression.
- Extensive buffering.
- Integrated routing services.
- SAN isolation from Internet, WAN, or MAN failures.

- Scalable remote-site fan-in.
- Hardware-based encryption.
- Internet Protocol Security (IPSec).
- Write acceleration for fast replication performance.
- Optional FICON with control unit port (CUP) and FICON Accelerator enable support for enterprise-class environments.

For more information about these products, see this website:

https://www.ibm.com/storage/san#71666

13

Certification

In this chapter, we provide an insight into several professional certifications that relate to the topics in this book.

13.1 The importance of certification

Why make the effort to certify when you have more than enough work to do anyway? The following benefits to an individual can be realized:

- Validates your skills and knowledge
- Gains peer recognition
- Offers you the potential to become more valuable to your company and in the marketplace
- Ratifies your skills as an industry professional

To an employer, the following benefits can be seen:

- Offers a great way of benchmarking the skill level of the employee
- Gives confidence about the employee's ability to support storage networks
- Demonstrates standards-based, non-proprietary, and vendor-neutral storage concepts

13.2 IBM Professional Certification Program

Today's marketplace is both crowded and complex. Individuals and businesses that do not 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. The extensive IBM portfolio of integrated certifications includes servers, software, application, and solution skills. The certification process is designed to prepare you and your company to meet business initiatives with real solutions.

The IBM Professional Certification Program helps you lay the groundwork for your personal journey to become a world-class resource to your clients, colleagues, and company. The program provides you with the correct skills and accreditation that are needed to succeed.

13.2.1 About the program

The IBM Professional Certification Program is both a journey and a destination. It is a business solution: a way for skilled IT professionals to demonstrate their expertise to the world. The certification validates your skills and demonstrates your proficiency in the latest IBM technology and solutions.

The certification requirements can be tough. It is a rigorous process that differentiates you from everyone else.

The following list provides the mission of the IBM Professional Certification Program:

- To provide a reliable, valid, and fair method of assessing skills and knowledge.
- To provide IBM with a method of building and validating the skills of individuals and organizations.
- To develop a loyal community of highly skilled certified professionals who recommend, sell, service, support, or use IBM products and solutions.

For more information about the IBM Professional Certification Program, see this website:

http://www.ibm.com/certify/index.shtml

13.2.2 Certifications by product

IBM has various certification courses that cover software, hardware, and products. For more information about the courses that are available, see this website:

http://www.ibm.com/certify/certs/index.shtml

13.2.3 Mastery tests

Mastery tests are used to verify the mastery of knowledge that is covered in a course or a defined set of learning materials. They are not certification tests, which are designed to validate skills that are needed in a specific job role. Rather, mastery tests help to assure that an individual achieved a foundation of knowledge and understanding of a subject matter.

Mastery tests supplement certifications as a method that is used by IBM to evaluate the knowledge of IBM sales and technical professionals. As with certifications, the successful completion of a mastery test might be required for participation in certain IBM Business Partner activities.

For more information about the courses that are available, see this website:

http://www.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. Different certification options are available within the SNIA. The certification program is called the *Storage Networking Certification Program* (*SNCP*).

The SNCP provides a strong foundation of vendor-neutral, systems-level credentials that integrate with and complement individual vendor certifications.

The structure of the SNCP is enhanced to reflect the advancement and growth of storage networking technologies over the past few years. And the structure is refined 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, no single standard existed 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.

For an up-to-date list of the possible certifications, see this website:

http://www.snia.org/education/certification

13.4 Brocade certification

Brocade has a large track of certification exams. Brocade certification exams are designed to validate your knowledge and expertise. The questions require knowledge and demonstrated expertise in the various areas that are tested.

For more information about the available certifications, see this website:

http://www.brocade.com/en/education/certification.html

13.5 Cisco certification

Cisco has various certifications for different product categories. This section focuses on storage area networking (SAN) and system networking. Cisco has five levels of general IT certification: Entry, Associate, Professional, Expert, and Architect.

For more information about the available certifications, see this website:

http://www.cisco.com/web/learning/certifications/index.html#~Cert

13.6 Open Group certification

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.

For more information about the certifications, see this website:

http://www.opengroup.org/certifications

Related publications

The publications that are listed in this section are considered particularly suitable for a more detailed discussion of the topics that are covered in this book.

IBM Redbooks

The following IBM Redbooks domains and associated publications provide additional information about the topics in this document. Note that some publications might be available in softcopy only:

IBM Storage Networking Redbooks:

http://www.redbooks.ibm.com/Redbooks.nsf/domains/san?Open

IBM Flash storage Redbooks

http://www.redbooks.ibm.com/Redbooks.nsf/domains/flash?Open

IBM Software Defined Storage Redbooks

http://www.redbooks.ibm.com/Redbooks.nsf/domains/sds?Open

IBM Disk storage Redbooks

http://www.redbooks.ibm.com/Redbooks.nsf/domains/disk?Open

IBM Storage Solutions Redbooks

http://www.redbooks.ibm.com/Redbooks.nsf/domains/storagesolutions?Open

IBM Tape storage Redbooks

http://www.redbooks.ibm.com/Redbooks.nsf/domains/tape?Open

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

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

Online resources

The following websites are also relevant as further information sources:

- IBM Storage hardware, software, and solutions:

  http://www.storage.ibm.com
- IBM System Storage storage area networks:

  http://www.ibm.com/systems/storage/san/
- Broadcom:

  http://www.broadcom.com
- Brocade:

  http://www.brocade.com

- Cisco:

  http://www.cisco.com
- QLogic:

  http://www.qlogic.com
- Emulex:

  http://www.emulex.com
- Finisar:

  http://www.finisar.com
- IBM Tivoli software:

  http://www.ibm.com/software/tivoli
- 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 Portal and downloads:

https://www.ibm.com/support/entry/portal/support

IBM Global Technology Services:

http://ibm.co/1lyI24R

Redbooks

Introduction to Storage Area Networks

SG24-5470-08

ISBN DocISBN

IBM

(0.5” spine)
0.475”<->0.873”
250 <-> 459 pages

SG24-5470-08

ISBN DocISBN

Printed in U.S.A.