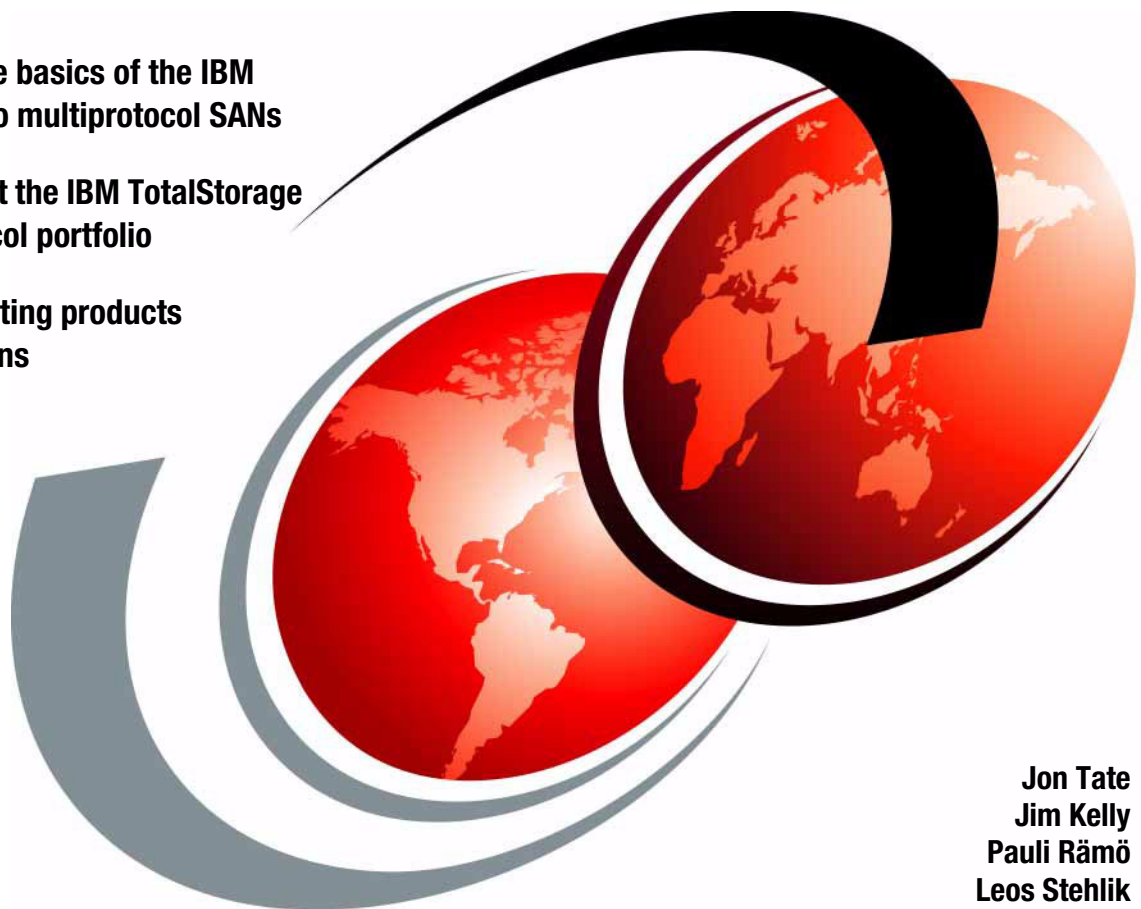


IBM TotalStorage: Introduction to SAN Routing

Uncover the basics of the IBM
approach to multiprotocol SANs

Learn about the IBM TotalStorage
multiprotocol portfolio

Explore routing products
and solutions



Jon Tate
Jim Kelly
Pauli Rämö
Leos Stehlik



International Technical Support Organization

IBM TotalStorage: Introduction to SAN Routing

May 2006

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

First Edition (May 2006)

This edition applies to the IBM TotalStorage SAN Routing portfolio described herein.

© Copyright International Business Machines Corporation 2006. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Figures	ix
Notices	xiii
Trademarks	xiv
Preface	xv
The team that wrote this redbook	xvi
Become a published author	xviii
Comments welcome	xix
Chapter 1. SAN routing	1
1.1 SAN routing definitions	2
1.1.1 Fibre Channel	2
1.1.2 Fibre Channel switching	3
1.1.3 Fibre Channel routers	3
1.1.4 Tunneling	3
1.1.5 Routers and gateways	3
1.1.6 Fibre Channel routing between physical or virtual fabrics	4
1.2 Gateway protocols	4
1.2.1 FCIP	4
1.2.2 iFCP	5
1.2.3 iSCSI	7
1.3 Routing issues	9
1.3.1 Packet size	9
1.3.2 TCP congestion control	9
1.3.3 Round-trip delay	10
1.3.4 Write acceleration	11
1.3.5 Tape acceleration	13
1.4 Multiprotocol scenarios	15
1.4.1 Dividing a fabric into sub-fabrics	15
1.4.2 Connecting a remote site over IP	16
1.4.3 Connecting hosts using iSCSI	16
Chapter 2. IBM TotalStorage b-type family routing products	19
2.1 IBM TotalStorage b-type family	20
2.2 Hardware and software	20
2.2.1 2109-A16 hardware components	21
2.2.2 Software features	23
2.2.3 Management capability	24

2.3 SAN routing terminology	24
2.4 Description of the routing solution	27
2.5 Current limitations	30
2.5.1 FC-FC routing	30
2.5.2 FCIP tunneling	31
2.5.3 iSCSI gateway	31
Chapter 3. IBM TotalStorage b-type family routing solutions	33
3.1 FC-FC routing	34
3.1.1 Local FC-FC routing	34
3.1.2 Fabric extension with FC-FC routing	36
3.2 FCIP tunneling	37
3.3 iSCSI gateway	37
Chapter 4. IBM TotalStorage b-type family routing best practices	39
4.1 Planning considerations	40
4.1.1 Piloting new technology	40
4.1.2 FC-FC routing considerations	40
4.1.3 FCIP tunneling considerations	40
4.2 Compatibility and interoperability	42
4.3 Availability	42
4.4 Security	43
4.4.1 FC-FC routing security	43
4.4.2 FCIP tunneling security	43
4.4.3 iSCSI gateway security	44
4.5 Performance	44
4.5.1 FC-FC routing performance	44
4.5.2 FCIP tunneling performance	45
4.5.3 iSCSI performance	45
4.6 IP network issues	45
4.6.1 Link bandwidth	46
4.6.2 Link latency	46
4.6.3 TCP receive window	48
4.6.4 Packet loss rate	48
4.6.5 Out-of-order packet delivery	49
Chapter 5. IBM TotalStorage b-type family real-life routing solutions	51
5.1 Backup consolidation	52
5.1.1 Customer environment and requirements	52
5.1.2 The solution	53
5.1.3 Failure scenarios	54
5.2 Migration to a new storage environment	55
5.2.1 Customer environment and requirements	55
5.2.2 The solution	56

5.3 Long distance disaster recovery over IP	58
5.3.1 Customer environment and requirements	58
5.3.2 The solution.	60
5.3.3 Normal operation.	62
5.3.4 Failure scenarios.	62
Chapter 6. Cisco family routing products	65
6.1 Overview of the Cisco MDS family	66
6.1.1 Introduction to VSAN.	66
6.2 Hardware and software	67
6.2.1 Cisco MDS 9120 and 9140 Multilayer Switches	67
6.2.2 MDS 9216A Multilayer Switch.	68
6.2.3 Cisco MDS 9216i Multilayer Switch	69
6.2.4 MDS 9506 Multilayer Director	70
6.2.5 MDS 9509 Multilayer Director	71
6.2.6 Optional modules	74
6.3 Advanced management	79
6.3.1 Fabric management	80
6.3.2 Optional licensed feature packages	84
6.4 Key features	87
6.4.1 Protocol support	87
6.4.2 Supported port types.	88
6.4.3 VSAN	91
6.4.4 Inter-VSAN Routing.	96
6.4.5 PortChanneling	98
6.4.6 Trunking	99
6.4.7 Quality of Service	100
6.4.8 Fibre Channel Congestion Control	101
6.4.9 Switch port analyzer	102
6.5 Interoperability.	105
6.5.1 Switch interoperability modes	105
6.5.2 Interoperability matrix	107
Chapter 7. Cisco solutions	109
7.1 SAN extension with FCIP	110
7.1.1 Compression.	111
7.1.2 Using Inter-VSAN Routing with FCIP	111
7.1.3 Using FCIP Write Acceleration	112
7.1.4 Using Fibre Channel tape acceleration with FCIP	113
7.2 Low-cost connection with iSCSI	113
7.2.1 iSCSI Immediate Data.	115
7.3 Isolation and interoperability using IVR.	115
7.3.1 Separating production from development	116

7.3.2 Separating corporate subsidiaries	118
7.3.3 Isolation of multivendor switches and modes	119
7.4 Managing scalability with IVR	120
7.5 Storage migration using IVR	121
Chapter 8. Cisco best practices	125
8.1 To route or not to route?	126
8.2 Piloting new technology	127
8.3 iSCSI issues	127
8.4 IP network issues	128
8.5 Interoperability	129
8.5.1 SAN Volume Controller interoperability	129
8.6 Designing for availability	130
8.6.1 Fibre Channel router hardware	130
8.6.2 Nondisruptive software upgrade	130
8.6.3 Inter-switch links	131
8.6.4 VSAN and IVR	131
8.6.5 Backup	131
8.7 Designing for security	131
8.8 Designing for performance	133
8.8.1 Hardware selection	133
8.8.2 FCIP compression and FCIP-WA	134
Chapter 9. Cisco real-life solutions	137
9.1 University ZYX	138
9.1.1 Initial growth	138
9.1.2 Lease expiration	140
9.1.3 Design and purchase of new systems	140
9.1.4 Deployment of iSCSI and FCIP	141
9.1.5 SVC synchronous replication for disaster recovery	142
9.2 Power Transmission Company ZYX	144
9.2.1 Existing systems	144
9.2.2 IT improvement objectives	145
9.2.3 Deployment of new technology and establishment of the disaster recovery site	145
9.2.4 Global Mirroring established to the disaster recovery site	147
Chapter 10. IBM TotalStorage m-type family routing products	151
10.1 Product description	152
10.1.1 IBM TotalStorage SAN04M-R	152
10.1.2 IBM TotalStorage SAN16M-R	156
10.2 SAN router architecture	160
10.2.1 SAN routing terminology	160
10.2.2 SAN routing features	161

10.2.3 SAN routing architecture	168
Chapter 11. IBM TotalStorage m-type family solutions	177
11.1 SAN fabric local FC-FC routing	178
11.2 SAN extension with iFCP	180
11.3 Low-cost connection with iSCSI	182
11.4 Isolation and interoperability using SAN routing	184
11.4.1 Separating production from development	184
11.4.2 Separating corporate subsidiaries.	184
11.4.3 Isolation of multivendor switches and modes	185
11.5 Migration of existing storage to a new environment	186
Chapter 12. IBM TotalStorage m-type family best practices	191
12.1 The planning checklist.	192
12.1.1 The installation checklist.	193
12.1.2 Running a pilot solution.	193
12.2 Fabric considerations	194
12.3 Bandwidth and capacity planning	194
12.3.1 Aspects that influence communication performance.	195
12.3.2 Throughput and efficiency.	197
12.3.3 The amount of data and link sizing	198
12.3.4 Fast Write and IBM products.	199
12.4 Planning for availability	200
12.4.1 Hardware limitations	200
12.4.2 Multiple paths and path failover on a router level	200
12.4.3 Fault isolation	201
12.5 Planning for security	201
12.5.1 Ports used by m-type SAN routers	201
12.5.2 Zoning	202
12.6 Scalability and limitations	203
Chapter 13. IBM TotalStorage m-type family real-life routing solutions	207
13.1 Backup consolidation	208
13.1.1 Customer environment and requirements.	208
13.1.2 The solution.	209
13.1.3 Failure scenarios.	210
13.2 Migration to a new storage environment.	211
13.2.1 Customer environment and requirements.	211
13.2.2 The solution.	213
13.3 Long distance disaster recovery over IP	215
13.3.1 Customer environment and requirements.	215
13.3.2 The solution.	217
13.3.3 Normal operation.	219
13.3.4 Failure scenarios.	219

Glossary 223

Related publications 243

 IBM Redbooks 243

 Other resources 244

 Referenced Web sites 244

 How to get IBM Redbooks 245

 IBM Redbooks collections..... 245

Index 247

Figures

1-1	Fibre Channel frame structure	2
1-2	FCIP encapsulates the Fibre Channel frame into IP packets	4
1-3	iFCP encapsulation and header mapping	6
1-4	iSCSI packet format	8
1-5	A standard write request	12
1-6	Write acceleration or Fast Write request	13
1-7	Tape acceleration example	14
2-1	IBM TotalStorage SAN 16B-R	20
2-2	Meta SAN with four edge fabrics connected to a backbone fabric	25
2-3	Edge fabric logical view of four EX_ports	26
2-4	FCIP tunnel	27
2-5	FC-FC routing physical layout	28
2-6	Logical view from fabric 1	29
2-7	Logical view from fabric 2	29
3-1	Local FC-FC routing between two SAN fabrics	34
3-2	Fabric extension with FC-FC routing	36
3-3	FCIP tunneling	37
3-4	iSCSI gateway	37
4-1	Direct FCIP connection	41
4-2	Routed FCIP connection	41
5-1	Current backup environment	52
5-2	New backup environment	53
5-3	Initial storage environment	55
5-4	Interim environment for migration	56
5-5	Final storage environment	58
5-6	Customer environment	60
5-7	Disaster recovery solution	61
6-1	MDS 9120 Multilayer Switch (IBM 2061-020)	68
6-2	MDS 9140 Multilayer Switch (IBM 2061-040)	68
6-3	MDS 9216A Multilayer Switch (IBM 2062-D1A) with 48 ports	68
6-4	Cisco MDS 9216i	69
6-5	MDS 9506 Multilayer Director (IBM 2062-D04)	71
6-6	MDS 9509 Multilayer Director (IBM 2062-D07)	72
6-7	MDS 9500 Series supervisor module	74
6-8	16 port switching module	75
6-9	32 port switching module	75
6-10	Cisco MDS 9000 14+2 Multiprotocol Services Module	76
6-11	The 8-port IP Services Module	77

6-12	Storage Services Module	79
6-13	Cisco MDS 9000 Fabric Manager user interface	81
6-14	Out-of-band management connection	82
6-15	In-band management connection	83
6-16	Cisco MDS 9000 family port types	91
6-17	Traditional SAN	92
6-18	Virtual SAN	93
6-19	Inter-VSAN Routing	97
6-20	PortChannels and ISLs on the Cisco MDS 9000 switches	99
6-21	Trunking and PortChanneling	100
6-22	SD_Port for ingress (incoming) traffic	103
6-23	SD_Port for egress (outgoing) traffic	104
7-1	Tunneling FCIP using transit VSANs to mitigate link bounce	112
7-2	Using iSCSI routing to provision disk storage to non-critical servers.	115
7-3	Every MDS switch is also a router: Dual director example	116
7-4	Every MDS switch is also a router: Four switch example	117
7-5	Using VSAN, IVR to isolate subsidiaries with access to shared tape	118
7-6	Using VSAN and IVR to provide multivendor isolation and integration	119
7-7	Using VSAN and IVR to manage scalability	120
7-8	An existing SAN ready for upgrade	121
7-9	IVR allows separation during migration phase.	122
7-10	Connecting servers to the new fabric and disconnecting the old SAN	123
8-1	How block size affects iSCSI performance	128
9-1	University ZYX SAN environment after four years.	139
9-2	The network after installing the DS8100, DS4300, and Cisco MDS	141
9-3	University ZYX planned network with FCIP and iSCSI	142
9-4	University ZYX network including SVC sync replication	143
9-5	Existing SAN environment at Power Transmission Company ZYX	145
9-6	Development/test break from production; DR site established	146
9-7	Input to the Async PPRC Bandwidth Sizing Estimator	147
9-8	Output from the Async PPRC Bandwidth Sizing Estimator	148
9-9	Utilization statistics for the disaster recovery DS6800 at 5000 IOPs.	149
9-10	Global Mirroring established using FCIP tunneling and IVR	150
10-1	mSAN and iSAN interconnections	161
10-2	Fibre Channel Frame to IP Datagram encapsulation.	163
10-3	SCSI write over high-latency environment without Fast Write	164
10-4	SCSI Write over high-latency environment with Fast Write	165
10-5	iSCSI protocol	166
10-6	Connecting iSCSI servers to an existing fabric using m-type router	167
10-7	SAN router internal network architecture	169
10-8	An example of iSAN architecture	174
10-9	Additional FSPF costs behind the 0x7E domain	175
11-1	Local FC-FC routing between two SAN fabrics	178

11-2	SAN extension over IP using iFCP over 700 km distance	181
11-3	Using iSCSI routing to provision disk storage to non-critical servers. .	183
11-4	Separating fabrics using two routers	184
11-5	Using m-type SAN routers to isolate subsidiaries	185
11-6	Using SAN routers to provide multivendor isolation and integration . .	186
11-7	Initial storage environment.	187
11-8	Migration interim storage environment.	188
11-9	How the environment looks after the migration	189
12-1	Path failover on router level.	201
13-1	Current backup environment	208
13-2	New backup environment	210
13-3	Initial storage environment.	212
13-4	Interim environment for migration	213
13-5	Final storage environment	215
13-6	Customer environment.	217
13-7	Disaster recovery solution	218

Notices

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

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive Armonk, NY 10504-1785 U.S.A.

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

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

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

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

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


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

COPYRIGHT LICENSE:

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

Trademarks

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

AIX®	IBM TotalStorage Proven™	Storage Tank™
BladeCenter®	IBM®	System/360™
Enterprise Storage Server®	ibm.com®	System/370™
Enterprise Systems	PowerPC®	Tivoli®
Architecture/390®	PR/SM™	TotalStorage Proven™
Everyplace®	pSeries®	TotalStorage®
ESCON®	Redbooks (logo)  ™	WebSphere®
@server®	Redbooks™	xSeries®
@server®	RS/6000®	z/Architecture™
FlashCopy®	S/360™	z/OS®
FICON®	S/370™	zSeries®
HACMP™	S/390®	

The following terms are trademarks of other companies:

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

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

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

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

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

Preface

The rapid spread and adoption of production storage area networks (SANs) has fuelled the need for multiprotocol routers. The routers provide improved scalability, security, and manageability by enabling devices in separate SAN fabrics to communicate *without* merging fabrics into a single, large SAN fabric. This capability allows clients to initially deploy separate SAN solutions at the departmental and data center levels. Then clients can consolidate these separate solutions into large enterprise SAN solutions as their experience and requirements grow and change.

Alternatively multiprotocol routers can help to connect existing enterprise SANs for a variety of reasons. For instance, the introduction of Small Computer System Interface over IP (iSCSI) provides for the connection of low-end, low-cost hosts to enterprise SANs. The use of an Internet Protocol (IP) in the Fibre Channel (FC) environment provides for resource consolidation and disaster recovery planning over long distance. And the use of FC-FC routing services provides connectivity between two or more fabrics without having to merge them into a single SAN.

This IBM® Redbook targets storage network administrators, system designers, architects and IT professionals who are engaged in the selling, designing, or administration of SANs. It introduces you to the products, concepts, and technology in the IBM TotalStorage® SAN Routing portfolio. It shows the features of each product and examples of how you can deploy and use them.

Prior to reading this Redbook, you must be familiar with SANs. If not, we recommend that you read the following IBM Redbooks™ before you start this one:

- ▶ *IBM TotalStorage: SAN Product, Design, and Optimization Guide*, SG24-6384
- ▶ *Introduction to Storage Area Networks*, SG24-5470

The team that wrote this redbook

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

Jon Tate is a Project Manager for IBM TotalStorage SAN Solutions at the ITSO, San Jose Center. Before joining the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2 support for IBM storage products. Jon has 20 years of experience in storage software and management, services, and support. He is both an IBM Certified IT Specialist and an IBM SAN Certified Specialist. He is also a Member of the British Computer Society, Chartered IT Professional (MBCS CITP).

Jim Kelly works in Storage Field Technical Sales Support for the Systems and Technology Group in IBM New Zealand. He is also a SNIA Certified Professional (SCP). Prior to joining IBM in 1999, he spent 13 years at Data General, including a brief stint with EMC. Jim spent the early part of his career working in an IBM VSE mainframe environment.

Pauli Rämö is an Advisory IT Specialist in IBM Global Services, Finland. He has 13 years of experience in working with RS/6000®, IBM @server pSeries®, AIX®, HACMP™, and Linux®. His areas of expertise include open systems storage solutions and SAP R/3 Basis. Pauli has also contributed to two other SAN-related Redbooks: *Introducing Hosts to the SAN Fabric*, SG24-6411, and *IBM TotalStorage: SAN Product, Design, and Optimization Guide*, SG24-6384.

Leos Stehlik is an IT Specialist for IBM Tivoli® Storage Manager and SAN solutions at IBM Global Services in the Czech Republic. He has eight years of experience in working with UNIX®, Microsoft® Windows® NT and 2000, and storage management. He helped to write the IBM Redbook *Using Tivoli Storage Manager in a SAN Environment*, SG24-6132, and helped to develop workshop material for IBM Tivoli Storage Network Manager.



The team (from left to right): Jon Tate, Pauli Rämö, Leos Stehlik, and Jim Kelly

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

Tom Cady
Deanna Polm
Sangam Racherla
ITSO, San Jose Center

Lisa Dorr
IBM Systems and Technology Group

Cal Blombaum
Scott Drummond
Michael Starling
Jeremy Stroup
IBM Storage Systems Group

Jim Baldyga
Brian Steffler
Brocade Communications Systems

Mark Allen
Kamal Bakshi
Reena Choudhry
Dan Hersey
Seth Mason
John McKibben
Cuong Tran
Cisco Systems

Brent Anderson
McDATA Corporation

Tom and Jenny Chang
Garden Inn Hotel, Los Gatos, California

Become a published author

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

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

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

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

ibm.com/redbooks

- ▶ Send your comments in an email to:

redbook@us.ibm.com

- ▶ Mail your comments to:

IBM Corporation, International Technical Support Organization
Dept. QXXE Building 026
5600 Cottle Road
San Jose, California 95193-0001



SAN routing

SAN routing provides new tools to manage:

- ▶ Departmental isolation and resource sharing
- ▶ Technology migration and integration
- ▶ Remote replication of disk systems
- ▶ Remote access to disk and tape systems
- ▶ Low-cost connection to SANs

This chapter introduces the terminology, technologies, and the value propositions for SAN routing.

1.1 SAN routing definitions

It is important that you clearly understand the terms and principles of Fibre Channel (FC) routing before you learn about designing routed networks.

You can find an excellent introduction in *Multiprotocol Routing for SANs* written by Brocade’s Josh Judd. Another book which addresses the topic is *IP SANs: An Introduction to iSCSI, iFCP, and FCIP Protocols for Storage Area Network*, written by McDATA’s Tom Clark. For details about locating these books, see “Related publications” on page 243.

We also recommend the two-day class “Cisco Storage Network Design Essentials” from Cisco. To learn more about this class, go to:

http://www.cisco.com/pcgi-bin/front.x/wwtraining/CELC/index.cgi?action=CourseDesc&COURSE_ID=3975

1.1.1 Fibre Channel

Fibre Channel is a set of standards for a serial input/output (I/O) bus developed through industry cooperation. A Fibre Channel frame consists of a header, payload, and 32-bit CRC bracketed by start of frame (SOF) and end of frame (EOF) delimiters. The header contains the control information necessary to route frames between N_PORTS and manage exchanges and sequences.

It is beyond the scope of this redbook to cover Fibre Channel in any great depth. For further reading, we recommend:

- *Introduction to Storage Area Networks*, SG24-5470
- *IBM TotalStorage: SAN Product, Design, and Optimization Guide*, SG24-6384

Figure 1-1 shows the layout of a Fibre Channel frame.

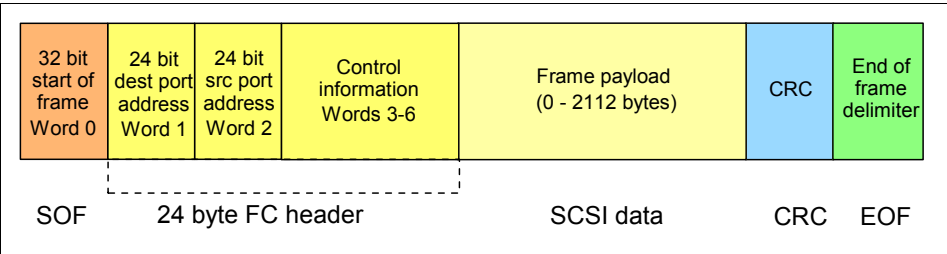


Figure 1-1 Fibre Channel frame structure

1.1.2 Fibre Channel switching

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

Important: Fibre Channel switching cannot transfer packets between fabrics.

1.1.3 Fibre Channel routers

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

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

Important: Fibre Channel routers forward packets between fabrics.

1.1.4 Tunneling

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

1.1.5 Routers and gateways

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

1.1.6 Fibre Channel routing between physical or virtual fabrics

Brocade, Cisco, and McDATA all offer FC-FC routing between separate physical fabrics. Cisco also offers Inter-VSAN Routing (IVR), which is Fibre Channel routing between separate logical (virtual) fabrics. In October 2004, the Technical Committee T11 of the International Committee for Information Technology Standards (INCITS) selected Cisco's VSAN technology for approval by the American National Standards Institute (ANSI) as the industry standard for virtual fabrics.

1.2 Gateway protocols

The topics that follow introduce the protocols that are encountered in a "routed" environment.

1.2.1 FCIP

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

Figure 1-2 shows FCIP tunneling, assuming that the Fibre Channel packet is small enough to fit inside a single IP packet.

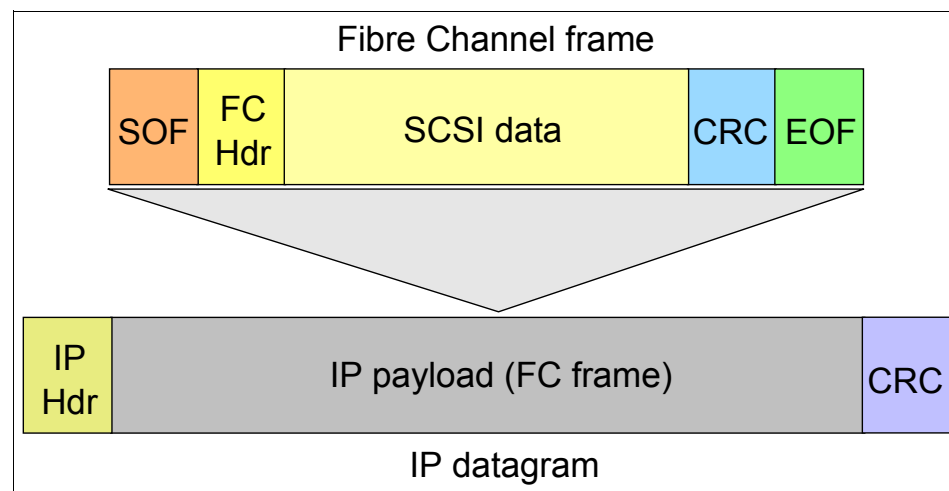


Figure 1-2 FCIP encapsulates the Fibre Channel frame into IP packets

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

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

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

Merging fabrics

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

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

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

1.2.2 iFCP

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

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

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

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

There is a popular myth that iFCP does not use encapsulation. In fact, iFCP encapsulates the Fibre Channel packet in much the same way that FCIP does. In addition, it maps the Fibre Channel header to the IP header and a TCP session, as shown in Figure 1-3.

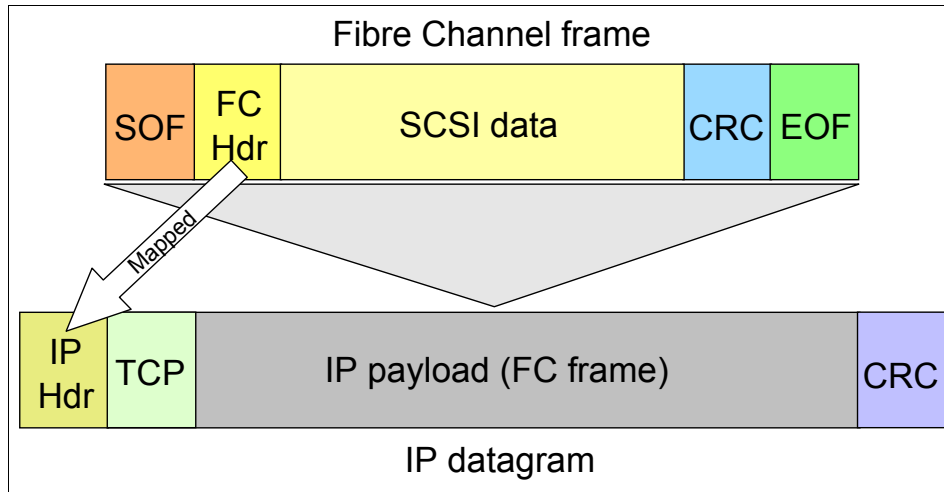


Figure 1-3 iFCP encapsulation and header mapping

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

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

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

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

1.2.3 iSCSI

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

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

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

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

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

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

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

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

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

Figure 1-4 shows the format of the iSCSI packet.

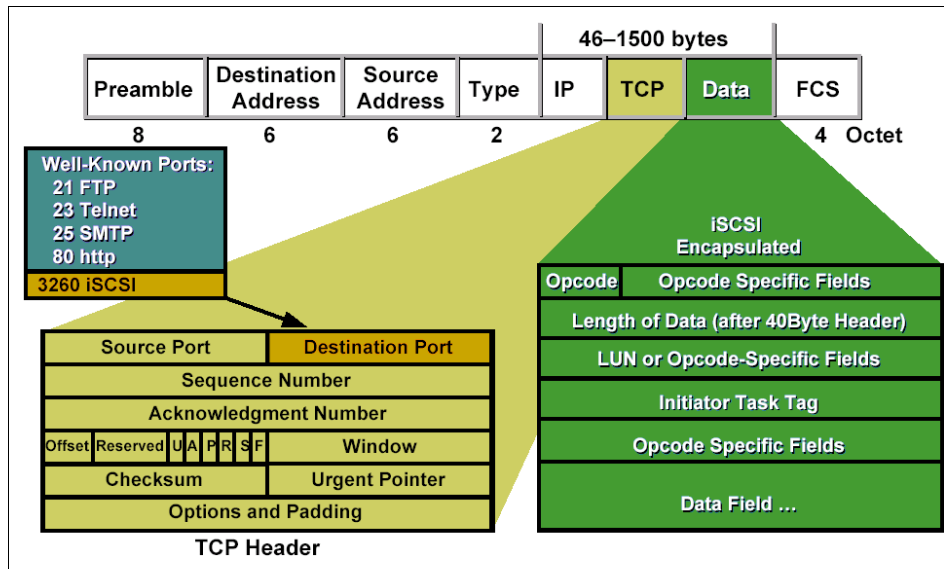


Figure 1-4 iSCSI packet format

Testing on iSCSI latency has shown a difference of up to 1 ms of additional latency for each disk I/O as compared to Fibre Channel. This does not include such factors as trying to do iSCSI I/O over a shared, congested or long-distance IP network, all of which may be tempting for some customers. iSCSI generally uses a shared 1 Gbps network. The round trip delays in “Time of frame in transit” on page 10 also apply.

iSCSI naming and discovery

There are three ways for an iSCSI initiator to understand what devices are in the network:

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

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

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

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

1.3 Routing issues

The topics that follow briefly describe some issues associated with a routed Fibre Channel environment.

1.3.1 Packet size

The standard size of a Fibre Channel packet is 2148 bytes, and the standard IP packet size is 1500 bytes (with a 1460 byte payload). When transporting Fibre Channel over IP, you can use jumbo IP packets to accommodate larger Fibre Channel packets. Keep in mind that jumbo IP packets must be turned on for the whole data path. In addition, a jumbo IP packets is not compatible with any devices in the network that do not have a jumbo IP packet enabled.

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

Each technology and each vendor may implement this differently. They all try to avoid sending small inefficient packets.

1.3.2 TCP congestion control

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

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

$$\text{minimum} = \text{maximum}/20$$

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

Other vendors take a simpler approach to achieve much the same end. Rather than introduce new algorithms, they suggest setting:

$$\text{minimum} = \text{maximum} \times 0.8$$

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

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

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

1.3.3 Round-trip delay

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

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

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

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

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

Time of frame in transit

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

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

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

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

Table 1-1 details the transmission times of this FCIP packet over some common wide area network (WAN) link speeds.

Table 1-1 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 needs to be divided into two Ethernet packets. This doubles the amount of TCP, IP, and Ethernet MAC overhead for the data transfer.

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

1.3.4 Write acceleration

Write acceleration, or Fast Write as it is sometimes called, is designed to mitigate the problem of the high latency of long distance networks. Write acceleration eliminates the time spent waiting for a target to tell the sender that it is ready to receive data. The idea is to send the data before receiving the ready signal, knowing that the ready signal will almost certainly arrive as planned. Data integrity is not jeopardized because the write is not assumed to have been successful until the final acknowledgement has been received.

Figure 1-5 shows a standard write request.

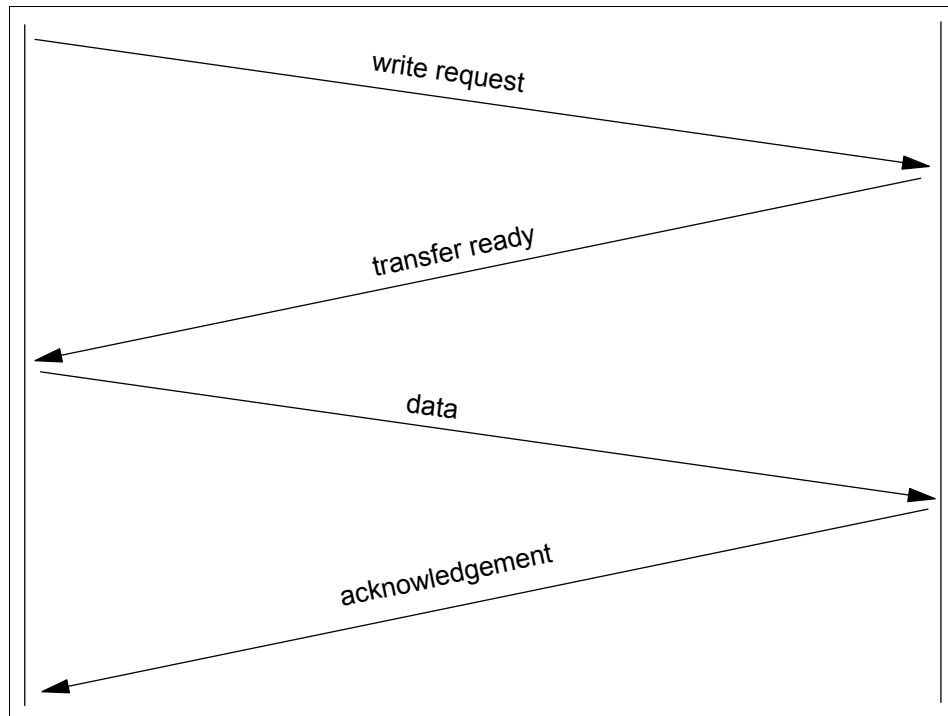


Figure 1-5 A standard write request

Figure 1-6 shows an accelerated write request.

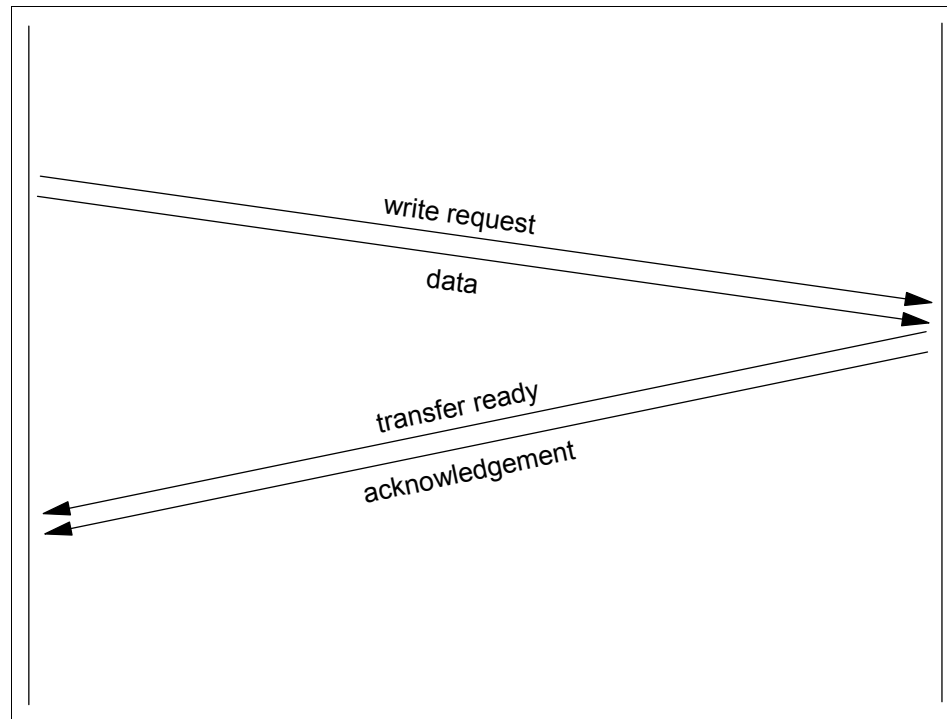


Figure 1-6 Write acceleration or Fast Write request

1.3.5 Tape acceleration

Tape acceleration (TA) takes write acceleration one step further by “spoofing” the transfer ready and the write acknowledgement. This gives the tape transfer a better chance of streaming rather than running stop and start. The risk here is that writes have been acknowledged but may not have completed successfully.

Without tape acceleration a sophisticated backup/restore application, such as Tivoli Storage Manager, can recover and restart from a broken link. However, with TA, Tivoli Storage Manager believes that any write for which it has received an acknowledgement must have completed successfully. The restart point is therefore set after that last acknowledgement. With TA, that acknowledgement was spoofed so it might not reflect the real status of that write.

Tape acceleration provides faster tape writes at the cost of recoverability. While the write acknowledgments are spoofed, the writing of the final tape mark is never spoofed. This provides some degree of integrity control when using TA.

Figure 1-7 shows how you can use tape acceleration to improve data streaming.

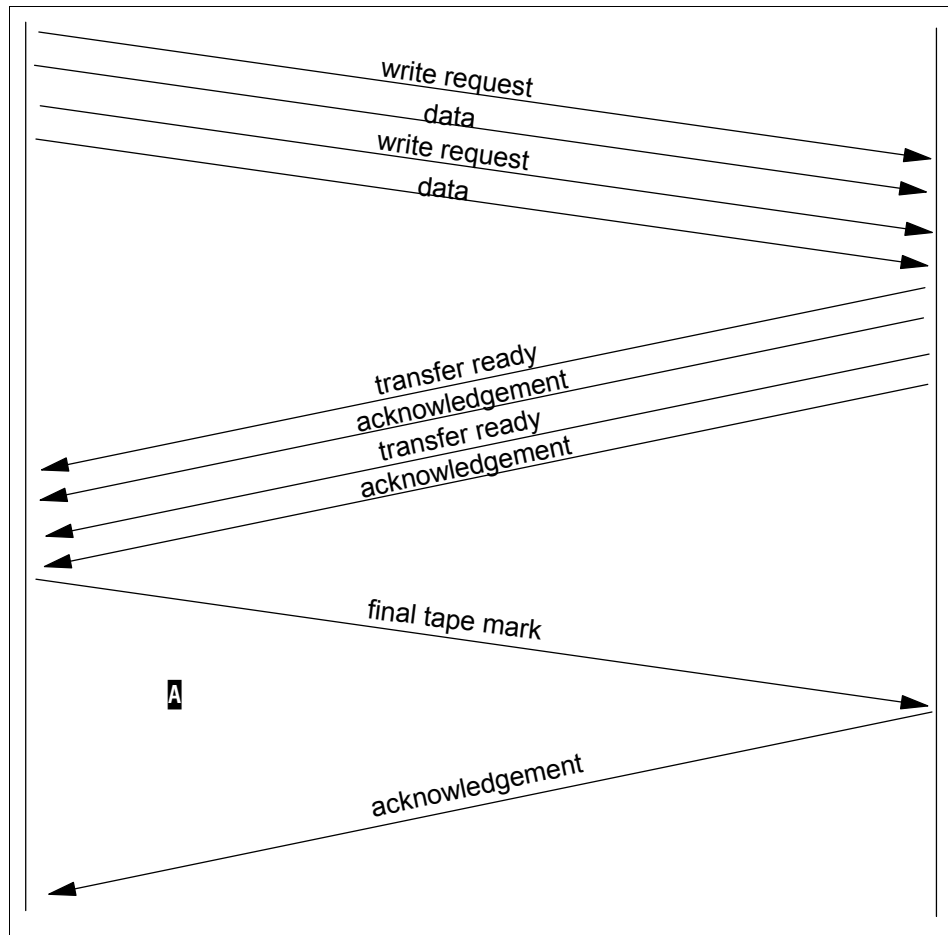


Figure 1-7 Tape acceleration example

1.4 Multiprotocol scenarios

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

1.4.1 Dividing a fabric into sub-fabrics

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

The development/test environment does not follow the same change control disciplines as the production environment. Also systems and switches can be upgraded, downgraded, or rebooted on occasions.

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

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

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

If using virtual fabrics as well (Cisco only at time of writing), then you can apply additional separation can also be applied within a physical fabric.

Note: FC-FC routing can provide departmental isolation and accommodate resource sharing.

1.4.2 Connecting a remote site over IP

Suppose you want to replicate your disk system to a remote site, perhaps 50 km away synchronously, or 500 km away asynchronously. Using FCIP tunneling or iFCP conversion, you can transmit your data to the remote disk system over a standard IP network. The router includes Fibre Channel ports to connect back-end devices or switches and IP ports to connect to a standard IP wide area network router. Standard IP networks are generally much lower in cost to provision than traditional high quality dedicated dense wavelength division multiplexing (DWDM) networks. They also often have the advantage of being well understood by internal operational staff.

Similarly you might want to provision storage volumes from your disk system to a remote site. You can do this by using FCIP tunneling (Brocade and Cisco) or iFCP protocol conversion (McDATA).

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

1.4.3 Connecting hosts using iSCSI

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

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

Note: The iSCSI router is effectively the iSCSI target. Each iSCSI initiator, such as the server, is mapped to a worldwide name (WWN) generated by the router. As far as the Fibre Channel disk system is concerned, it sees each initiator as a separate Fibre Channel attached server. Fibre Channel logical unit numbers (LUNS) are mapped to iSCSI Qualified Name (IQNs) generated by the router. As far as the iSCSI initiator is concerned, it sees iSCSI targets.

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

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

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

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



IBM TotalStorage b-type family routing products

This chapter introduces the storage area network (SAN) routing concepts in the IBM TotalStorage b-type family of SAN products, and the products involved. This chapter examines the hardware and software, SAN routing terminology, and the routing solution.

2.1 IBM TotalStorage b-type family

The IBM TotalStorage b-type family has currently one product that implements SAN routing. This is the IBM TotalStorage SAN16B-R multiprotocol router (2109-A16). The 2109-A16 is designed to provide intelligent multiprotocol routing services to help enable Internet Protocol (IP)-based global mirroring business continuity solutions. It provides Small Computer System Interface over IP (iSCSI)-based host server storage consolidation solutions. And it provides SAN routing capabilities for infrastructure simplification solutions to selectively share resources between fabrics.

Figure 2-1 shows the 2109-A16.



Figure 2-1 IBM TotalStorage SAN 16B-R

2.2 Hardware and software

The 2109-A16 provides the following hardware features:

- ▶ Eight or sixteen active multiprotocol ports with the following features:
 - Support for 1 Gbps and 2 Gbps Fibre Channel (FC) or Gigabit Ethernet
 - Automatic negotiation to the highest common speed of all Fibre Channel devices connected to port
 - Support for short wavelength (SW), long wavelength (LW), and extended long wavelength (ELW) small form-factor pluggable (SFP) transceivers
 - Support for F_Port, FL_Port, or E_Port operation
 - Support of each active Gigabit Ethernet port for either iSCSI gateway or FC over IP (FCIP) tunneling functionality
- ▶ No SFPs included; one Tri-rate SFP for each active port required
- ▶ Two 10/100 Mbps Ethernet ports and one RS-232 serial port for management
- ▶ Redundant, hot pluggable power supplies and fans
- ▶ Two U chassis

The following standard software features are included with the 2109-A16:

- ▶ Fibre Channel switch support
- ▶ Advanced WEB TOOLS
- ▶ Advanced zoning
- ▶ Exchange-based inter-switch link (ISL) and inter-fabric link (IFL) trunking
- ▶ Extended fabric capability
- ▶ iSCSI gateway functionality

The optional software features available for the router include:

- ▶ Fibre Channel routing
- ▶ FCIP
- ▶ FCIP and FC Routing Bundle

The hardware and software components are described in more detail in the following sections.

2.2.1 2109-A16 hardware components

The 2109-A16 is a multiple board design. Below the system board in the chassis is a dc power printed circuit board (PCB) that provides the required system voltages. These voltages are derived from and regulated by the 2109-A16 redundant power supply units. This regulated power is bused through a connector to the main system PCB. Mounted on top of the system board is a daughter board that contains a high-performance 800 MHz PowerPC® 745x reduced instruction set computer (RISC) processor core with SDRAM controller, PCI bus interface, peripheral local bus for external ROM and peripherals, direct memory access (DMA), I2C interface, and general purpose I/O.

The system uses four types of memory devices in the design: SDRAM, kernel flash, compact flash (user flash), and boot flash. The fabric application and switching section of the system board, the XPath per-port processing application-specific integrated circuit (ASIC) and memory chip sets, the XPath Fabric ASIC, and the SFP media are the key components that provide high-speed data manipulation and movement. The SFP media interface to external devices and support any combination of SW, LW, and ELW optical media.

Power supplies

The 2109-A16 power supply is a hot-swappable field replaceable unit (FRU), allowing 1+1 redundant configurations. The unit is a universal power supply that is capable of functioning worldwide without voltage jumpers or switches. The fully enclosed, self-contained unit has internal fans to provide cooling. The power supply provides three dc outputs (5V standby, 12V, and 48V), with a total maximum output power of 320W. An integral on/off switch, input filter, and power

indicator are provided in each power supply, as well as a serial EEPROM device that provides identifying information.

Multiprotocol ports

The 2109-A16 has 16 multiprotocol ports (numbered 0 through 15, left to right). You can purchase it with either eight or 16 active ports. If you purchase the 2109-A16 with only eight active ports, you can activate the other eight ports by purchasing and installing the Ports on Demand license key.

Each of the 16 multiprotocol ports needs to be equipped with an SFP. The SFPs are hot swappable and use industry-standard local channel connectors. Each port is supported by its own dedicated ASIC, that contains three embedded 133 MHz ARM processors.

In Fibre Channel mode, each port provides support for E_port, F_port, and FL_port modes. The ports also support automatic negotiation of both port mode and speed. With the current firmware, each port can provide up to 255 buffer credits.

Important: Since the multiprotocol ports of the 2109-A16 supports both Fibre Channel and Gigabit Ethernet, they require the use of special Tri-rate SFPs. The router only activates ports that are equipped with supported SFPs. We recommend that you always use only the SFPs delivered with the router.

Fabric ASIC

The XPath Fabric ASIC provides non-blocking connectivity between the 16 separate port ASICs.

Management ports

The 2109-A16 provides dual 10/100 BaseT Ethernet ports for management purposes. The TCP/IP address for each port can be configured via the serial port.

Serial port

An RS-232 serial port is provided on the 2109-A16. The serial port uses an RJ-45 connector. The serial port's parameters are fixed at 9600 baud, 8 data bits, and no parity, with flow control set to None. This connection is used for initial IP address configuration and for recovery of the switch to its factory default settings, should flash memory contents be lost. The serial port connection is not intended for normal administration and maintenance functions.

Cooling fans

The non-port side of the 2109-A16 includes dual hot-swappable cooling fan assemblies. Each contain three fans and system status LEDs.

2.2.2 Software features

The XPath OS is used on the 2109-A16. It provides full Fibre Channel switch capability, FC-FC routing, FCIP tunneling, and iSCSI to FC gateway.

Fibre Channel switch support

The XPath OS includes the following Fibre Channel switch features:

- ▶ Name server support
- ▶ Zone server support
- ▶ Exchange-based ISL trunking
- ▶ Extended fabric support

FC-FC routing

The FC-FC routing service provides connectivity to devices in different fabrics without merging the fabrics. FC-FC routing allows the creation of logical storage area networks (LSANs). An LSAN can span multiple fabrics, allowing Fibre Channel zones to cross physical SAN boundaries without merging the fabrics, yet maintaining access control of the zones.

FC-FC routing also allows you to share devices, such as tape drives, across multiple fabrics without the associated administrative problems that can result from merging the fabrics, including change management, network management, scalability, reliability, availability, and serviceability.

FCIP tunneling

The FCIP tunneling service enables tunneling of Fibre Channel frames through TCP/IP networks. It encapsulates them in TCP packets and then reconstructs them at the other end of the link.

Note: The XPath OS supports FCIP connection only between two 2109-A16s.

iSCSI gateway

The iSCSI gateway service provides connectivity to Fibre Channel targets for servers using iSCSI. Servers use an iSCSI adapter or an iSCSI driver and Ethernet adapter to connect to a Fibre Channel fabric over IP.

2.2.3 Management capability

There are three ways to manage the 2109-A16:

- ▶ Command line interface (CLI)
- ▶ Advanced WEB TOOLS-AP Edition
- ▶ Fabric Manager

The CLI and Advanced WEB TOOLS are similar to the same features in the IBM TotalStorage b-type switches. They are included with the 2109-A16 at no extra cost.

Fabric Manager provides a Java™-based application that can simplify management of a multiple-switch fabrics. It can be used to administer, configure, and maintain fabric switches and SANs. In addition, it has an easy-to-use wizard for creating LSANs. Fabric Manager is available as an optional feature of most IBM TotalStorage b-type SAN switches.

2.3 SAN routing terminology

In a traditional SAN fabric, we are used to the following terms:

- ▶ *E_Port* is a port type used to connect two SAN switches together.
- ▶ The link between two *E_Ports* is called an *inter-switch link*.
- ▶ The collection of SAN switches connected by ISLs is called a *SAN fabric*.
- ▶ A connectivity group within a single SAN fabric is called a *zone*.

To support the new Fibre Channel routing functionality provided by the 2109-A16, we introduce the following new terms:

- ▶ Any SAN fabric that has direct connections to hosts and storage devices is called an *edge fabric*.
- ▶ The SAN fabric that is used for routing the Fibre Channel traffic between the edge fabrics is called *backbone fabric*.
- ▶ *EX_Port* is a new port type in the router, used in the backbone fabric to connect to an *E_Port* in an edge fabric.
- ▶ The link between an *EX_Port* and an *E_Port* is called an *inter-fabric link*.
- ▶ The collection of edge fabrics and backbone fabrics connected together is called a *Meta SAN*.
- ▶ Each edge fabric has a *fabric ID (FID)* that is unique within the same Meta SAN, and configured to all *EX_Ports* connected to the edge fabric.

- A connectivity group that spans two or more edge fabrics is called an *LSAN*. It is implemented by creating a zone, whose name starts in *LSAN_*, in all edge fabrics involved.

Figure 2-2 shows a Meta SAN that consists of four edge fabrics connected to a single backbone fabric.

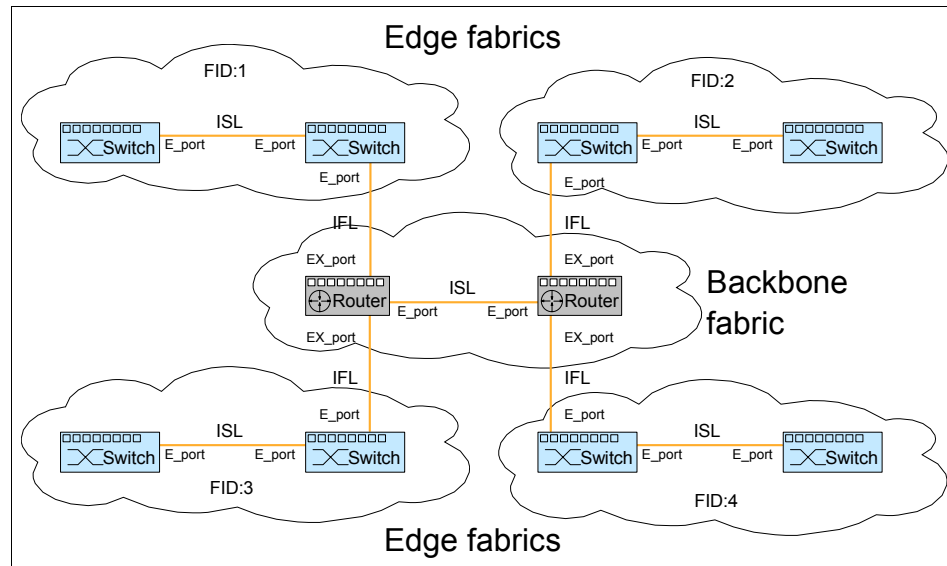


Figure 2-2 Meta SAN with four edge fabrics connected to a backbone fabric

The switches in the edge fabric treat an IFL like a normal ISL. Via the IFL, they gain access to a set of *phantom domains* that are never principal domains in the edge fabric. Two types of phantom domains are created in the edge fabric:

- Every EX_Port connected to the fabric is represented by a *front domain (fd)*.
- Every remote edge fabric that has at least one node exported to the local edge fabric is represented by a *translate domain (xlate)*.

The phantom domains can have the following connections:

- The virtual links connecting front domains to translate domains are called *phantom links*.
- The exported nodes of a remote fabric are represented by port addresses in the translate domain called *NR_Ports*.

The translate domains are used to perform Fibre Channel network address translation (FC-NAT) between the different edge fabrics. The translate domain IDs are persistent across router reboots and can be assigned manually.

The front domains are used to provide multiple paths to the translate domains via the different IFLs that are available and allow normal Fabric Shortest Path First (FSPF) routing across the paths.

When counting the hop count in the complete Meta SAN, you need to count the ISLs and IFLs as hops. You don't need to count the phantom links, since they are not physical links and do not add any delay to the data path.

The port IDs (PIDs) of the NR_Ports follow a specific format 0xAABBBB, where:

- ▶ AA is the translate domain for the remote fabric where the physical device is attached.
- ▶ BBBB is the *virtual slot* number in the range 0xf001 - 0xffff.

The virtual slot numbers can be assigned automatically, or you can assign them manually.

Note: Some operating systems, such as AIX or HP-UX, assume that the PID of any device stays constant. If you have any servers running either of these operating systems, we recommend that you define both the translate domain IDs and the virtual slot numbers manually to ensure that they remain constant.

Figure 2-3 shows the logical view of an edge fabric with four IFLs having connection to another edge fabric.

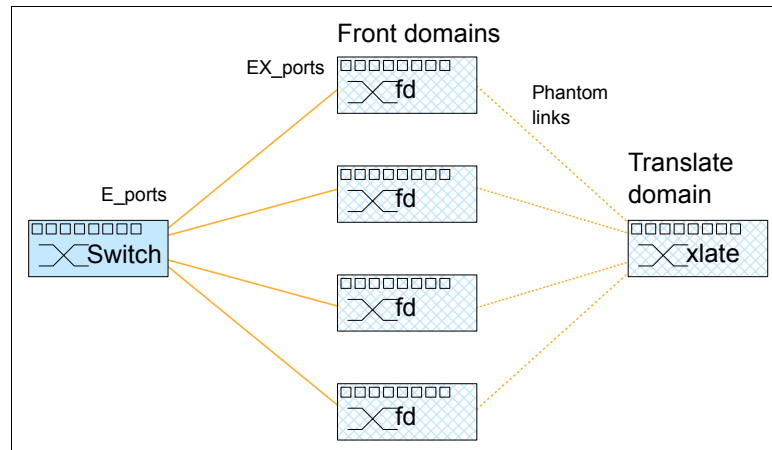


Figure 2-3 Edge fabric logical view of four EX_ports

In addition, to support FCIP tunneling, we introduce the *VE_Port*, or rather the virtual E_Port created over the FCIP tunnel. Figure 2-4 shows an FCIP tunnel between two 2109-A16s.

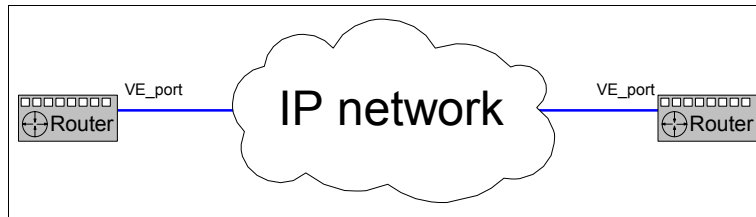


Figure 2-4 FCIP tunnel

The VE_Port functions exactly like a normal E_Port, and a normal ISL is created between the two VE_ports.

2.4 Description of the routing solution

This example has a small Meta SAN, consisting of two edge fabrics of one switch each and a single router backbone fabric. Each edge fabric has two connections to the router for redundancy.

The fabric parameters are set as follows:

- ▶ Fabric 1
 - Core PID: 1
 - Domain IDs used: 1
 - Fabric ID: 1
- ▶ Fabric 2
 - Core PID: 0
 - Domain IDs used: 1
 - Fabric ID: 2

Since we are using a different Core PID format in the fabrics, we cannot merge them into a single fabric. We are also using the same domain ID on both fabrics. Changing either of these parameters would be disruptive to the fabric involved.

We need to enable the server connected to fabric 1 access to a storage device that is connected to fabric 2. We implement this by creating an LSAN called LSAN_zone1.

Figure 2-5 shows the physical layout of our environment.

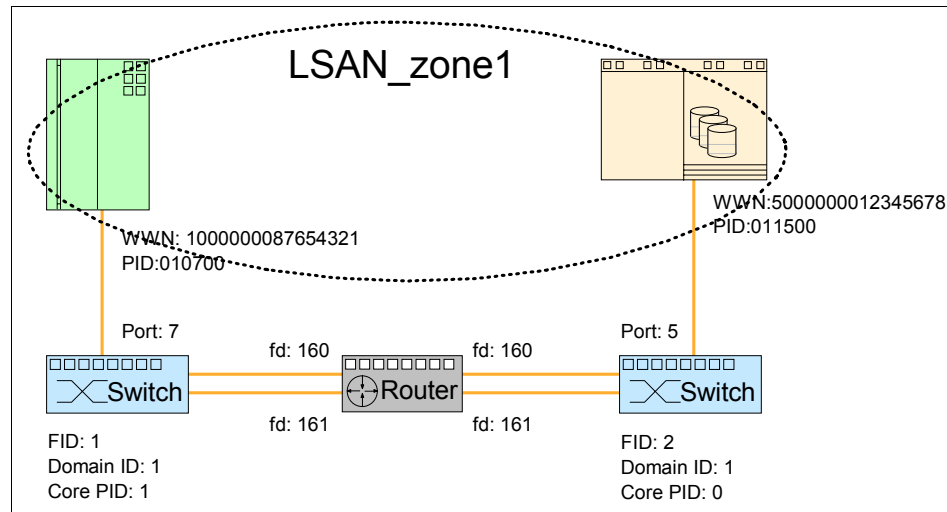


Figure 2-5 FC-FC routing physical layout

As you can see, the router automatically assigns a separate front domain for each IFL in each edge fabric, starting from domain ID 160. The front domain IDs need to be unique only within a single edge fabric, and the same router can use the same front domain ID in several different edge fabrics.

We create the LSAN by creating a zone named `LSAN_zone1` that has two members: the worldwide names (WWNs) 1000000087654321 and 5000000012345678. The router automatically intercepts any zone with a name starting with `LSAN_` and exports any LSAN members between fabrics as required.

In our case, we assume that fabric 2 is given translate domain ID 5 in fabric 1, and fabric 1 is given translate domain ID 6 in fabric 2. Any fabric 2 members exported to fabric 1 are represented in fabric 1 by `NR_ports` with PIDs starting from 05f001. And any fabric 1 members exported to fabric 2 are represented in fabric 2 by `NR_ports` with PIDs starting from 06f001.

Figure 2-6 shows the logical view seen by the server in fabric 1.

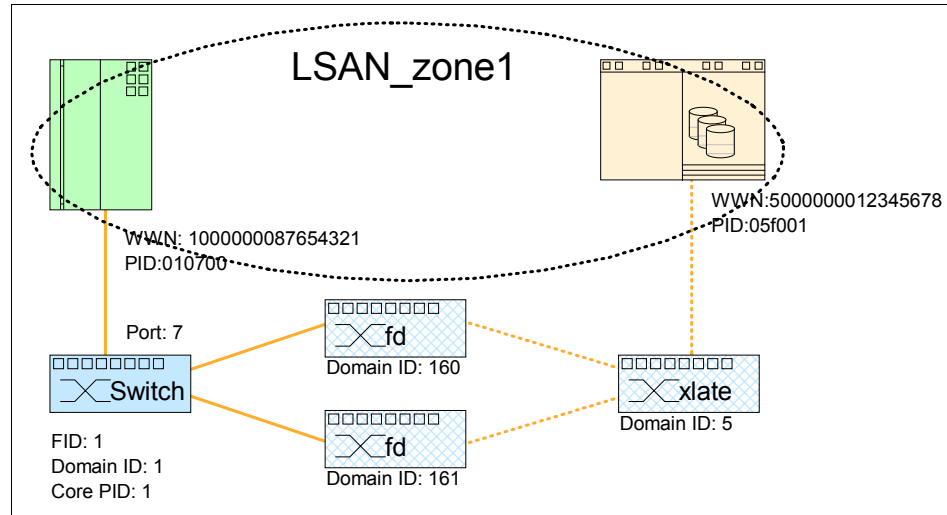


Figure 2-6 Logical view from fabric 1

Figure 2-7 shows the logical view seen by the storage device in fabric 2.

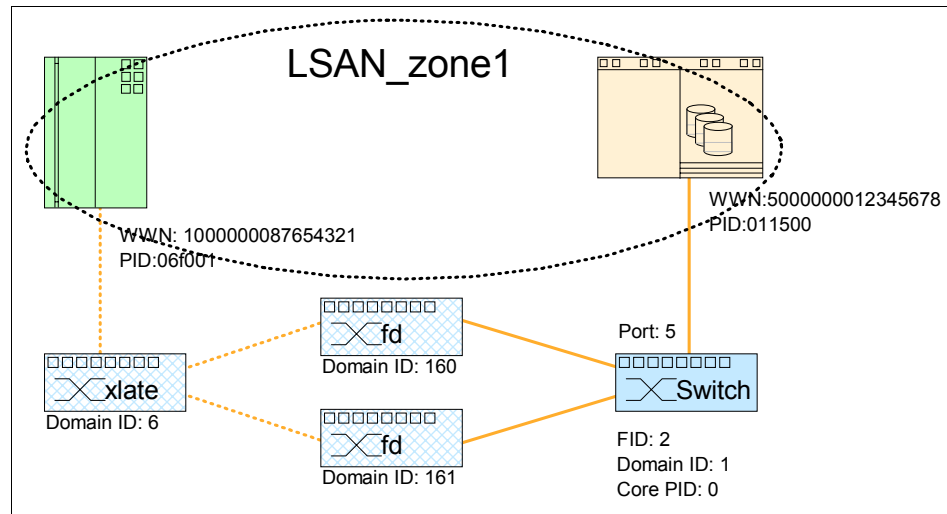


Figure 2-7 Logical view from fabric 2

The WWNs of the LSAN members are not changed by the NAT. This way you can still use the real WWN of the server for logical unit number (LUN) masking in the storage device.

2.5 Current limitations

This section examines the specific limitations of the different functions of the 2109-A16. The limitations are based on the capabilities of current hardware and XPath OS version 7.3, and are subject to change.

2.5.1 FC-FC routing

Since the backbone fabric can consist of multiple 2109-A16s and multiple FC switches, there is no practical limit on the number of edge fabrics that can be connected to the backbone fabric.

The FC-FC routing function has the following limitations:

- ▶ Edge fabrics in interoperability mode are currently not supported.
- ▶ LSAN can only contain nodes from edge fabrics, not from the backbone fabric.
- ▶ LSAN members must be specified by port WWN.
- ▶ The maximum number of hops between switches, including routers is 12.
- ▶ For each edge fabric:
 - Maximum number of front domains: 15
 - Maximum number of translate domains: 33
 - Maximum number of physical domains (switches): 32
 - Maximum total number of domains (physical + translate + front): 80
 - Maximum number of all devices (local + proxy): 1280
 - Maximum number of proxy devices: 1000
- ▶ For the complete routed fabric (Meta SAN):
 - Maximum number of edge fabrics: 34
 - Maximum number of FC routers: 14
 - Maximum number of LSAN zones: 1000
 - Maximum number of LSAN zone members: 10 000
 - Maximum number of members / LSAN zone: 200

2.5.2 FCIP tunneling

The FCIP tunneling function has the following limitations:

- ▶ One point-to-point FCIP connection is supported for each port configured for FCIP.
- ▶ The 2109-A16 does not implement compression on FCIP links.
- ▶ There is no Fast Write or write acceleration feature in the 2109-A16.

2.5.3 iSCSI gateway

The iSCSI gateway function has the following limitations:

- ▶ A maximum of 12 ports of the 2109-A16 can be configured as iSCSI portals.
- ▶ Each port can support up to eight iSCSI sessions.
- ▶ iSCSI traffic cannot exit the 2109-A16 via an EX_port or a VE_port.
- ▶ The current iSCSI client software supported is the Microsoft initiator v1.05.



IBM TotalStorage b-type family routing solutions

This chapter describes the solutions that are available using the IBM TotalStorage SAN16B-R multiprotocol router (2109-A16). It looks at the following three solutions:

- ▶ Fibre Channel to Fibre Channel (FC-FC) routing
- ▶ FC over Internet Protocol (FCIP) tunneling
- ▶ Small Computer System Interface over IP (iSCSI) gateway

3.1 FC-FC routing

This section describes the FC-FC routing functionality of the 2109-A16 and possible scenarios where the functionality can be used.

3.1.1 Local FC-FC routing

Figure 3-1 shows the local FC-FC routing solution between two storage area network (SAN) fabrics.

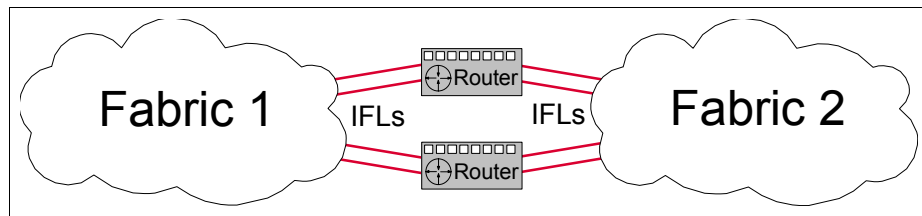


Figure 3-1 Local FC-FC routing between two SAN fabrics

This solution has two redundant 2109-A16s connected to different SAN fabrics. Both routers are connected to both fabrics using two inter-fabric link (IFLs). We can extend this configuration to span up to 8 SAN fabrics, or up to 16 fabrics by adding two more 2109-A16s.

Each edge fabric can be connected to the routers with up to four IFLs. If you have multiple switches in your fabric, we recommend that you distribute the IFL connections across them for maximum availability. If you are using a core-edge-fabric, we recommend that you connect IFLs to the core switches.

Some of the solutions that can be provided by local FC-FC routing include:

- Scalability

The Fibre Channel addressing can theoretically support up to 16 million nodes in a single fabric. However, the practical limit for the amount of nodes is much lower. This is similar to TCP/IP networks, where the network address space is divided into smaller subnets of limited number of IP addresses, and traffic is routed between them. Usually the practical limit of a fabric from both technical and management standpoint is between 250 and 1000 nodes.

FC-FC routing allows you to divide the environment into several fabrics, while providing access to shared resources between fabrics.

- Multiple SAN administrators

In many cases, enterprises have several small SAN islands that are managed by different SAN administrators, often as a result of mergers or acquisitions.

Using FC-FC routing between the fabrics allows each fabric to be managed separately from other fabrics. It prevents the propagation of any management errors to other fabrics.

The logical storage area network (LSAN) configuration is the only part of the fabric that needs to be coordinated among the SAN administrators. Since the LSAN zones need to be defined on all fabrics before they can route traffic, the devices in each fabric are protected against unplanned access from the other fabrics.

- Interoperability between storage vendors

Several storage vendors offer Brocade Silkworm SAN products, either as resellers or as OEM products. While these products are theoretically compatible with each other, each vendor usually only supports specific levels of Fabric OS, and it may be difficult to find a common, supported version among multiple storage vendors. The fabric-wide parameters and recommended zoning methodologies may also differ between vendors.

If the different vendors are each separated to their own SAN fabrics, as shown in Figure 3-1, we avoid these problems. This solution also allows each edge fabric to be supported and even managed by the corresponding storage vendor, while enabling storage access between different SAN fabrics.

- Interoperability between old and new fabrics

In many cases when implementing a new SAN fabric, you already have an existing fabric. The existing fabric may have some parameter settings that you want or need to set up differently in the new fabric. One good example is the Core PID setting.

By using FC-FC routing to connect the fabrics, you do not need to change the settings in the old fabric, and are free to choose the settings that you need for the new fabric. You can also use only a single Fabric OS level in any fabric, independent on the Fabric OS levels supported by the old hardware.

- Migration between old and new fabrics

The storage hardware is usually replaced with new hardware every three to five years. When refreshing the disk hardware, it may make sense to refresh the SAN hardware as well, especially if the new disk vendor is different than the old vendor.

FC-FC routing allows you to implement the new disk subsystems and SAN fabric in the final configuration and connect the complete new environment to the current SAN fabrics. This way you can have simultaneous access from the servers to both old and new disk subsystems, and use server-based tools, such as Logical Volume Manager (LVM), to migrate the data from the old disks to the new disks.

After you migrate any host to new disks, you can move the Fibre Channel ports of the server to the new SAN fabric as well. Since you can do this one server at a time, the outage needed is minimized.

► Storage consolidation

Many enterprises implement a separate SAN fabric for tape backups. Without FC-FC routing, this requires a separate Fibre Channel adapter in each server that needs to be connected to the backup devices, as well as the additional fiber cabling to support these adapters. If you set up FC-FC routing between the normal SAN fabrics and the backup fabric, you can share the tape devices across any adapters in any fabric, as required.

Another example of storage consolidation is implementing a single IBM TotalStorage SAN Volume Controller (SVC) cluster across multiple SAN fabrics.

3.1.2 Fabric extension with FC-FC routing

Figure 3-2 shows a simple SAN fabric extension using the 2109-A16.

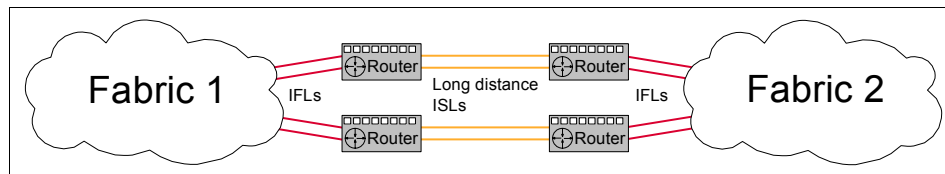


Figure 3-2 Fabric extension with FC-FC routing

This solution takes advantage of the large number of buffer credits available on each port in the 2109-A16, and the Extended Fabrics feature that is included with the 2109-A16. The long distance inter-switch links (ISLs) can be implemented with dark fiber, or with dense wavelength division multiplexing (DWDM) technology.

This solution has the advantage that the edge fabrics are separated from each other and from the long distance ISLs. This isolates any SAN fabric failure in one site to that site only, and prevents the other site from being affected. This is especially important in a disaster recovery environment. Any link failures on the long distance links are also isolated from both edge fabrics.

3.2 FCIP tunneling

In situations where dark fiber for DWDM is not available or the distances are longer than supported by DWDM, the SAN fabrics can be connected using the FCIP tunneling functionality in the 2109-A16. This solution is shown in Figure 3-3.



Figure 3-3 FCIP tunneling

We recommend that you combine FCIP tunneling with the FC-FC routing feature. This way you can isolate any SAN fabric failure in one site to that site only, and prevent the other site from being affected. Any link failures on the FCIP links are also isolated from both edge fabrics.

3.3 iSCSI gateway

The iSCSI gateway feature of the 2109-A16 allows you to connect servers with no Fibre Channel adapters to Fibre Channel attached storage devices using the iSCSI protocol. This solution is shown in Figure 3-4.

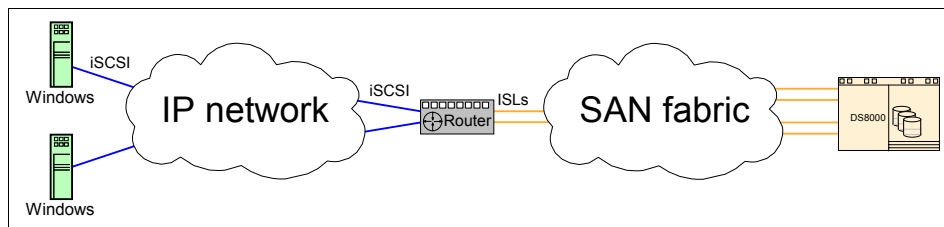


Figure 3-4 iSCSI gateway

Each 2109-A16 port configured for iSCSI supports up to eight concurrent iSCSI sessions. A total of 12 ports in a single 2109-A16 can be configured for iSCSI.

Note: In the current implementation, the iSCSI gateway traffic has to leave the 2109-A16 via an E_port. It cannot leave via an EX_Port or a VE_Port. The 2109-A16 must be a part of an edge fabric to connect to the storage devices in any edge fabric. Therefore it is usually not feasible to share a single 2109-A16 between the iSCSI function and the other functions described in this chapter.



IBM TotalStorage b-type family routing best practices

This chapter describes some of the key features of the IBM TotalStorage b-type family routing solution. It also provides information about best practices, when implementing such a solution.

Specifically the chapter examines:

- ▶ Planning considerations
- ▶ Compatibility and interoperability
- ▶ Availability
- ▶ Security
- ▶ Performance
- ▶ IP network issues

4.1 Planning considerations

This section covers the specific considerations you need to consider when planning solutions with the IBM TotalStorage SAN16B-R multiprotocol router (2109-A16).

4.1.1 Piloting new technology

Whenever you plan to use advanced features, such as Fibre Channel (FC) over IP (FCIP), Small Computer System Interface over IP (iSCSI), or FC-FC routing, it always pays to implement it initially as a pilot with the understanding that experience gained in your own environment will always be slightly unique.

When implementing leading-edge technologies, many clients prefer to avoid uncertain outcomes that a pilot implies. Instead, they secure implementation guarantees from vendors. But in fact, the outcome can never really be guaranteed and piloting allows the solution to be tailored based on lessons learned in your own environment.

4.1.2 FC-FC routing considerations

Because logical storage area networks (LSANs) are created like any other zones, apart from the name, it is possible to define all zones as LSAN zones. However, if you do this, each 2109-A16 would have to keep track of all your zones, limiting the future scalability of the Meta storage area network (SAN). Therefore we do not recommend this approach.

We recommend that you create zones that enable traffic within a single edge fabric as normal zones. Also, create a separate LSAN zone wherever you need to access a storage device from a host in different edge fabric. This way you also avoid any unnecessary access between devices.

If you are using FC-FC routing in an environment with AIX or HP-UX servers, we recommend that you define the translate domain IDs and virtual slot numbers manually. This way you can ensure that FC IDs of any devices used by the AIX or HP-UX servers remain persistent.

4.1.3 FCIP tunneling considerations

When implementing an FCIP connection, the quality of the IP link is critical. When you order the IP link from a vendor, you need to have a service-level agreement (SLA) that concerns the operation of the link. We recommend that you ensure that the link vendor takes into account the specific requirements of the storage environment in their design.

The SLA should include at least the following parameters:

- ▶ Guaranteed link bandwidth
- ▶ Round-trip latency
- ▶ Maximum packet loss rate
- ▶ Whether packet delivery can be out-of-order

You need an IP address and a subnet mask for both ends of the connection. Usually these are provided by the vendor to suit their addressing scheme. You need to configure one of the 2109-A16s to initiate the connection and the other one to listen for incoming connections.

If you have a direct connection with no Ethernet routers, the IP addresses should be in the same subnet, and the subnet mask should be same for both ends. Also, you do not need to specify any default gateway information for the link. Figure 4-1 shows an example of a direct FCIP connection.

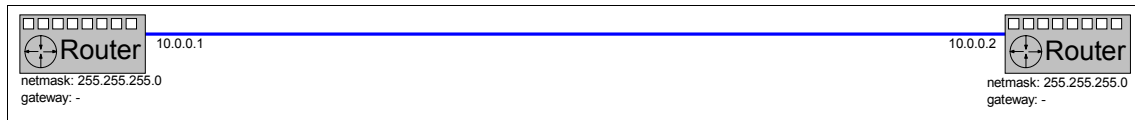


Figure 4-1 Direct FCIP connection

If you are using a routed connection, the IP addresses should be in different subnets, and you need to specify the correct default gateway address for both routers. The default gateway address is the address of the IP router in the same network with the 2109-A16. Figure 4-2 shows an example of a routed FCIP connection.

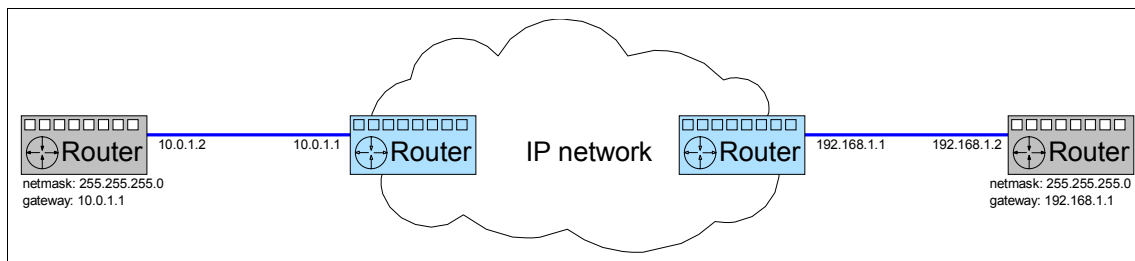


Figure 4-2 Routed FCIP connection

We also recommend that you synchronize the internal clocks of the 2109-A16s to an external NTP time server.

4.2 Compatibility and interoperability

With the current firmware version, the 2109-A16 is interoperable with most current and previous IBM b-type SAN switches, as well as the Brocade SilkWorm products. The currently supported switches and the required minimum switch Fabric OS levels are listed in Table 4-1.

Table 4-1 IBM TotalStorage SAN 16B-R interoperability

IBM products	Brocade products	Fabric OS
2109-S08, 2109-S16	SilkWorm 2000-series	v2.6.1 or later
3534-F08, 2109-F16	SilkWorm 3000, 3200, 3600, 3800	v3.1.0 or later
2109-F32, 2109-M12	SilkWorm 3900, 12000	v4.1.0x or later
2005-H08, 2005-H16, 2109-M14	SilkWorm 3250, 3850, 24000	v4.2.0x or later
2005-B32	SilkWorm 4100	v4.4.0b or later
2005-B16, 2109-M48	SilkWorm 200E and 48000	v5.0.1 or later

The 2109-A16 supports these switches in edge fabrics in their native mode, with any currently used Core PID format. The switches are also supported in the backbone fabric.

The 2109-A16 also supports interoperability with McDATA via the IBM RPQ process.

The 2109-A16 ports configured as Gigabit Ethernet are compatible with any standard Ethernet switch having fiber connection. If you need to connect the 2109-A16 to a switch that has only copper ports, use a media converter.

The FCIP implementation of the 2109-A16 only supports tunnels to another 2109-A16. The current iSCSI client supported by the 2109-A16 is the Microsoft iSCSI initiator v1.5.

4.3 Availability

The 2109-A16 has redundant, hot-swappable power supplies and fans. It currently supports hot code load, but not hot code activation.

For FC-FC routing and FCIP tunneling, we recommend that you install two or more 2109-A16s to ensure availability of the backbone fabric.

For the iSCSI gateway function, the 2109-A16 can be configured to support iSCSI failover between two routers. However, all iSCSI clients do not support such a configuration.

4.4 Security

The 2109-A16 can be used in three different roles. Therefore the following sections describe the security features used in each role.

4.4.1 FC-FC routing security

The FC-FC routing security is based on the LSAN concept. Each LSAN is a zone that contains multiple worldwide names (WWNs) from separate edge fabrics. It is implemented by defining the zone in each of the edge fabrics involved. The LSAN zones are separated from all zones by having their name start with LSAN_. The letters in this context are not case sensitive. When the zones are defined, the 2109-A16 automatically exports any shared nodes across the fabrics.

An LSAN zone can only contain members from edge fabrics, and not from the backbone fabric. Therefore, the nodes in the backbone fabric are inherently secure from any nodes in edge fabrics.

Since the LSAN zone needs to be separately defined in both edge fabrics, a zoning change in one edge fabric cannot enable unauthorized access to any node in any other edge fabric. This is especially important in cases where the edge fabrics are managed by different SAN administrators.

4.4.2 FCIP tunneling security

When you configure an FCIP tunnel, you need to choose which of the 2109-A16 routers you want to initiate the connection. The other 2109-A16 is configured to listen for connections.

The 2109-A16 configured to initiate the connection opens a connection to the FCIP well-known Transmission Control Protocol (TCP) port 3225 in the 2109-A16 that is configured to listen for the connection. If you have one or more firewalls between the 2109-A16s, you need to allow this traffic in the firewall. We recommend that you also allow **ping** between the devices to make problem determination easier.

To enhance security, you can configure each end of the FCIP tunnel with the WWN of the 2109-A16 in the remote end. This prevents any other 2109-A16 from connecting to the port.

4.4.3 iSCSI gateway security

When opening the iSCSI session, the iSCSI initiator connects to the iSCSI well-known TCP port 3260 in the router. If you have a firewall between the iSCSI client and the 2109-A16, you need to allow this traffic in the firewall.

The iSCSI gateway function of the 2109-A16 supports the optional use of Challenge Handshake Authentication Protocol (CHAP) in either one-way or two-way configurations.

In a one-way CHAP configuration, the iSCSI initiator is authenticated by the 2109-A16. The iSCSI initiator can open an iSCSI session only if the CHAP secret of the initiator matches the CHAP secret configured for the initiator in the 2109-A16. This way the iSCSI connection is protected against any other iSCSI initiator trying to use the same iSCSI Qualified Name (IQN).

In a two-way CHAP configuration, the iSCSI initiator is first authenticated by the 2109-A16, and the 2109-A16 is then authenticated with the iSCSI initiator. The iSCSI initiator can open an iSCSI session only if the CHAP secret of the initiator matches the CHAP secret configured for the initiator in the 2109-A16, and the CHAP secret of the 2109-A16 matches the CHAP secret configured for the 2109-A16 in the iSCSI initiator. This way the iSCSI initiator is also protected against any other device imitating the 2109-A16 iSCSI portal.

We recommend that you always use at least one-way CHAP configuration.

4.5 Performance

This section looks at the performance of the 2109-A16, when it is used in different solutions.

4.5.1 FC-FC routing performance

The 2109-A16 is capable of routing traffic at the full 2 Gbps line rate on all ports. Therefore, in an FC-FC routing scenario, the performance of an inter-fabric link (IFL) is practically the same as the performance of any 2 Gbps inter-switch link (ISL).

Due to having a separate application-specific integrated circuit (ASIC) for each port, the 2109-A16 supports exchange level trunking, instead of the frame level trunking available in the switches in the IBM TotalStorage b-type SAN family.

4.5.2 FCIP tunneling performance

FCIP performance is a complex issue, since it is affected by many separate factors, such as performance of the FCIP links and latency caused by the FCIP encapsulation itself.

In general, an FCIP link has significantly lower performance and higher latency than the same link would if it was a native Fibre Channel. Therefore we recommend that you use FCIP tunneling only when a native Fibre Channel link cannot be implemented for technical or financial reasons.

The 2109-A16 implements the traffic shaping feature that allows you to limit the amount of bandwidth used for the FCIP link. This way you can avoid overloading a wide area network (WAN) link slower than the Gigabit Ethernet interface used in the 2109-A16. The 2109-A16 also supports jumbo frames and exchange level trunking between up to four FCIP links.

You can learn more about FCIP link performance in 4.6, “IP network issues” on page 45.

4.5.3 iSCSI performance

iSCSI gateway functionality allows small servers, which do not have high performance or availability requirements for their storage connection, access to Fibre Channel-based storage systems.

While it is possible to share the same network interface between normal TCP/IP traffic and iSCSI traffic, we recommend that you implement a separate network for the iSCSI traffic. The iSCSI network should have low latency, less than 15 ms, and minimal packet loss. Due to the network latency requirements, we do not recommend using iSCSI over long distance connections.

Each iSCSI portal in the 2109-A16 is capable of providing up to 450 Mbps of throughput. We recommend that, if you have iSCSI clients using 100 Mbps Ethernet connections, you separate those clients to a different iSCSI portal than the clients that are using 1 Gbps Ethernet.

4.6 IP network issues

Several factors affect the performance of an FCIP link:

- ▶ Link bandwidth
- ▶ Link latency
- ▶ TCP receive window

- ▶ Packet loss rate
- ▶ Out-of-order packet delivery

This section discusses these factors in more detail.

Ask your telecommunications company about high quality Quality of Service (QoS) managed links, including information about such offerings as IP over SONET/SDH. Work closely with your telecommunications company so that they clearly understand your network quality expectations and you clearly understand the cost and management implications of any decisions you make.

4.6.1 Link bandwidth

The link bandwidth is the most obvious factor affecting performance. It is also one of the key metrics used when provisioning the link. In storage environments, the link bandwidth used should always be the *guaranteed bandwidth* from the service provider.

If the guaranteed bandwidth is anything less than one full Gigabit Ethernet, it needs to be configured into the routers in both ends of the link to use the traffic shaping feature of the 2109-A16 and avoid overrunning the link. We recommend that you set the maximum allowed speed of the FCIP port to 96% of the guaranteed bandwidth of the link on both ends of the link.

4.6.2 Link latency

Link latency is a metric of the round-trip time (RTT) it takes for a packet to cross the link. The key factors contributing to the link latency include:

- ▶ Distance
- ▶ Router and firewall latencies
- ▶ Time of frame in transit

Distance

The speed of light in optical fiber is approximately 208 000 km/s. Therefore the delay caused by a fiber connection is approximately 4.8 μ s/km. To calculate the round trip latency, we have to count this delay both ways.

For example, for a 100 km link, the round trip latency is approximately:

$$100 \text{ km} \times 4.8 \text{ } \mu\text{s/km} \times 2 = 960 \text{ } \mu\text{s}$$

Similarly, for a 1000 km link, the round trip latency is 9600 μ s, or 9.6 ms.

Router and firewall latencies

Any delay caused by routers and firewalls along the network connection needs to be added to the total latency. The latency varies a lot depending on the routers or firewalls and the traffic load. It can range from a few microseconds to several milliseconds.

You also need to remember that the traffic generally passes the same routers both ways. For round trip latency, you need to count the one-way latency twice.

If you are purchasing the routers or firewalls yourself, we recommend that you include the latency of any particular product among the criteria 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 SLA.

Time of frame in transit

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

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

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

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

Table 4-2 details the transmission times of this FCIP packet over some common 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 μ s	0.6 μ s
OC-12	622.08 Mbps	28 μ s	1.2 μ s
OC-3	155.52 Mbps	113 μ s	4.7 μ s
T3	44.736 Mbps	394 μ s	16.5 μ s
E1	2.048 Mbps	8600 μ s	359 μ s
T1	1.544 Mbps	11 400 μ s	477 μ s

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

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

4.6.3 TCP receive window

In addition to the line speed available, the maximum throughput available on any given TCP connection is determined by the TCP receive window of the receiving device. This is similar to the buffer-to-buffer-credit flow control used in Fibre Channel networks. To enable the full utilization of a given FCIP link, the size of TCP window allocated for the link needs to be large enough for all FCIP packets that are being transferred on the line.

4.6.4 Packet loss rate

In traditional TCP/IP networks, packet loss is a normal and accepted behavior. The built-in retransmission mechanism in the protocols handle retransmitting any dropped packet. Most protocols used in TCP/IP networks can easily handle high packet loss rates, such as 1%, without significant performance degradation.

Since FCIP uses a TCP connection for data transfer, it also uses the same mechanism. However, since latency is usually critical in storage applications, the storage networks do not cope well with retransmissions. Therefore networks used for storage traffic need a much lower packet loss rate. Even an IP network

with a packet loss rate of 0.01% is considered a low quality network, compared to the baseline of zero frame loss in Fibre Channel networks.

4.6.5 Out-of-order packet delivery

Fibre Channel networks rely on in-order delivery of Fibre Channel frames. The 2109-A16 receiving FCIP traffic needs to ensure that the Fibre Channel frame order is retained.

If IP packets are received out of order, as is often the case with shared IP networks using Multiprotocol Label Switching (MPLS), the 2109-A16 needs to buffer them until it receives the complete sequence of packets.

Usually the only effect of out-of-order packet delivery is slightly increased latency. However, in extreme cases, the receiving 2109-A16 may run out of buffer space and have to drop packets.



IBM TotalStorage b-type family real-life routing solutions

This chapter presents some real-life solutions implemented with the IBM TotalStorage b-type family routing products. It discusses the following solutions:

- ▶ Backup consolidation
- ▶ Migration to a new storage environment
- ▶ Long distance disaster recovery over IP

Important: The solutions and sizing estimates that we discuss or make in this chapter are unique. Make no assumptions that they will be supported or apply to each environment. We recommend that you engage IBM to discuss any proposal.

5.1 Backup consolidation

This scenario presents a solution to consolidate local area network (LAN)-free tape backups from two separate storage area network (SAN) fabrics.

5.1.1 Customer environment and requirements

The customer has two existing SAN fabrics and is currently using ArcServe software to back up the Windows servers in the SAN fabrics to tape. The customer also has several application servers that do not have SAN attachment. The customer environment is shown in Figure 5-1.

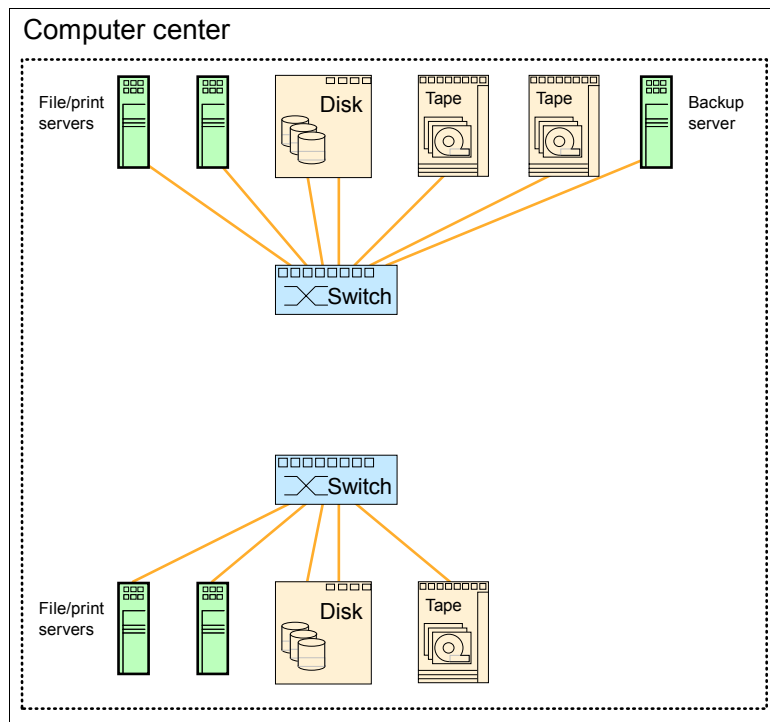


Figure 5-1 Current backup environment

The customer has the following requirements for the new solution:

- ▶ Consolidate the tape backups to a single Tivoli Storage Manager environment
- ▶ Provide for LAN-free backups from both current SAN fabrics
- ▶ Implement the new backup system to a separate location from the computer center
- ▶ Leverage the existing investment in SAN hardware

In the first SAN fabric, the customer currently has 160 GB of disk space. This space is projected to grow to 630 GB in the near future. In the second SAN fabric, the customer has 100 GB of disk space.

5.1.2 The solution

Our solution has the following new components:

- ▶ IBM @server pSeries server for Tivoli Storage Manager
- ▶ IBM 3583-L72 tape library with four Fibre Channel (FC) drives
- ▶ IBM TotalStorage SAN32B-2 switch for the backup environment
- ▶ IBM TotalStorage SAN16B-R router, with FC-FC routing enabled

The new backup environment is shown in Figure 5-2.

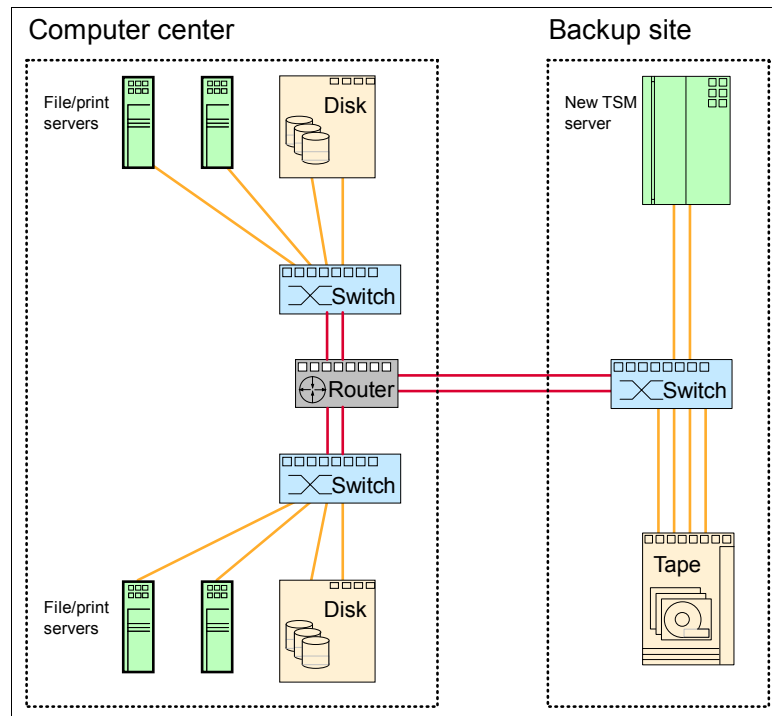


Figure 5-2 New backup environment

We locate the router in the computer center to minimize the need of fiber connections between the computer center and the backup site. All other components are located in a single rack at the backup site.

We connect each of the current fabrics, and the new backup switch, to the router with two inter-fabric links (IFLs) for redundancy. The customer provides the two

long wave fiber connections required between the computer center and the backup site.

The Tivoli Storage Manager server will use its internal disks for both Tivoli Storage Manager databases and disk storage pools. Therefore, it does not need any access to the existing customer SAN fabrics. The tape drives are divided evenly to the two Fibre Channel adapters in the Tivoli Storage Manager server.

We create a separate logical SAN (LSAN) zone for each server in any SAN fabric that needs access to the tape drives. The LSAN zone will contain the worldwide name (WWN) of the host bus adapter (HBA) of the server and the WWNs of all the tape drives. We also ask the customer to create the same LSANs in the existing SAN fabrics.

Since our new environment is used only for daily backups, it does not have the high availability requirements that SAN fabrics use for disk access. Therefore, it is adequate to have a single backup switch and a single router in the solution.

The application servers that are not connected to any SAN fabric are backed up to the Tivoli Storage Manager server over a LAN connection.

5.1.3 Failure scenarios

This section explains how the failure of different components affects the operation of our solution.

- ▶ **Power failure**

The Tivoli Storage Manager server, the tape library, and all SAN fabric components in the environment have dual redundant power supplies connected to different power circuits. Therefore a power failure in one circuit does not have any effect on operation.

- ▶ **IFL failure**

If an IFL fails, the system remains operational, but the maximum bandwidth available is reduced by 50%.

- ▶ **Router failure**

If the SAN router fails, it is impossible to run LAN-free backups. In this situation, the Tivoli Storage Manager client automatically uses a LAN-based method for any backups and restores. The Tivoli Storage Manager server and the servers that are not using LAN-free backups are not affected.

- ▶ **Backup switch or Tivoli Storage Manager server failure**

The failure of either the backup switch or the Tivoli Storage Manager server prevents any backup and restore activity.

5.2 Migration to a new storage environment

This scenario presents a solution to migrate the customer's current storage environment to a new environment.

5.2.1 Customer environment and requirements

The customer has a Hewlett-Packard (HP) XP512 storage system that is shared between AIX, HP-UX, and Windows servers. Due to historical reasons, each server platform has its own SAN fabrics and connections to the XP512.

Each SAN fabric consists of a single 16-port, 1 Gbps Brocade 2800 switch. Since the lease period of the environment expires within a few months, the customer needs a new solution to replace the current environment. The initial environment is shown in Figure 5-3. For clarity, we show only some of the servers.

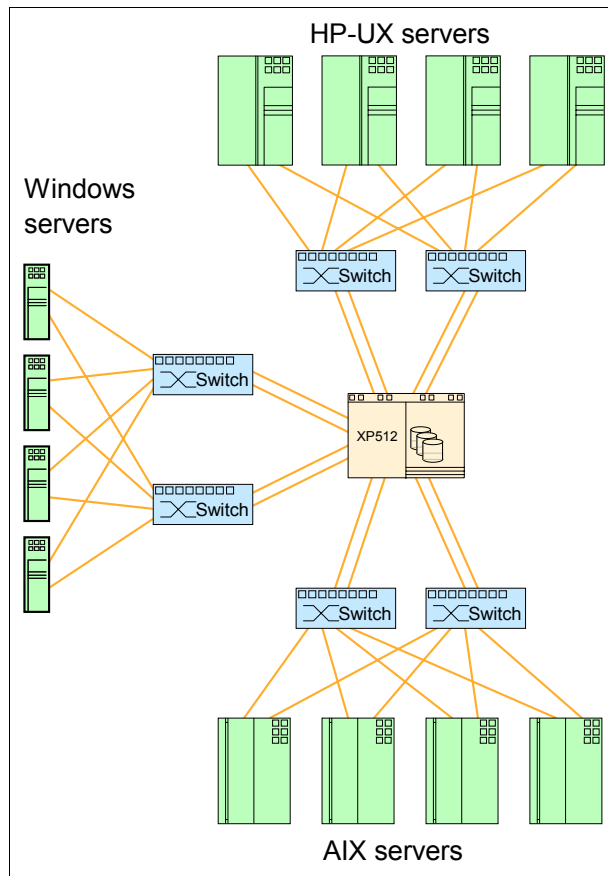


Figure 5-3 Initial storage environment

The customer has the following requirements for the new solution:

- ▶ New hardware to replace the current disk system and SAN fabric
- ▶ Flexibility in allocating ports between different platforms
- ▶ Scalability to support future applications
- ▶ Minimize the amount of downtime of servers due to migration

5.2.2 The solution

Our solution has the following new components:

- ▶ IBM TotalStorage DS8100 disk subsystem
- ▶ Two IBM TotalStorage SAN256B directors with 64 ports each
- ▶ Two IBM TotalStorage SAN16B-R routers with the FC-FC routing feature

We install the components of the new storage environment, and connect the environment to the old environment with IFLs, as shown in Figure 5-4.

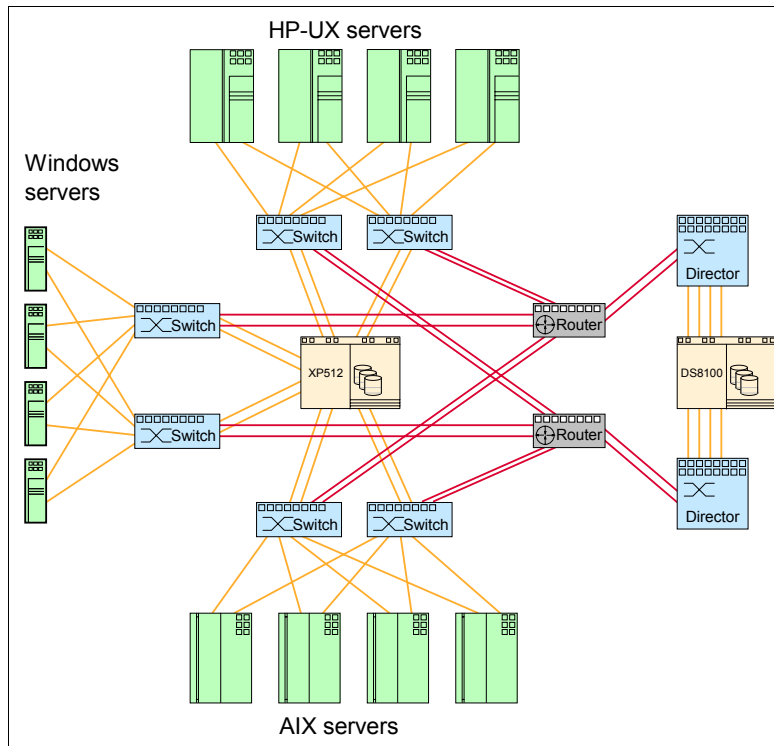


Figure 5-4 Interim environment for migration

In the new environment, all TotalStorage DS8100 ports are shared among all servers. Since we are only migrating a few servers at the same time, using a limited number of IFLs does not cause any performance degradation to the servers.

When the new storage environment is completely installed, we start migrating the servers one server or a group of servers at a time, using the following procedure:

1. Create LSANs to allow the server to access TotalStorage DS8100.
2. Install the IBM Subsystem Device Driver (SDD) package and any other TotalStorage DS8100 specific software on the server.
3. Allocate new storage in TotalStorage DS8100 to the servers.
4. Migrate all server data from the old storage to the new storage using the following operating system-based tools:
 - Native Logical Volume Manager (LVM) for AIX
 - PVLinks for HP-UX
 - Veritas Volume Manager for Windows
5. Create non-LSAN zones to allow the server to access the storage from the new SAN fabrics.
6. Disconnect the server from the old switches and move it to the new directors.
7. Delete the LSANs created in step 1.

The only step that requires server downtime in the procedure is step 5. If the new cabling is prepared beforehand, this step should take little time.

After the migration of all servers is complete, no servers should be connected to the old switches and the XP512 should be idle. At this time, we can remove the old storage hardware from the environment. The IBM TotalStorage SAN16B-R routers are also freed and can be used for other purposes, such as SAN extension over FC over IP (FCIP).

Figure 5-5 shows the final storage environment.

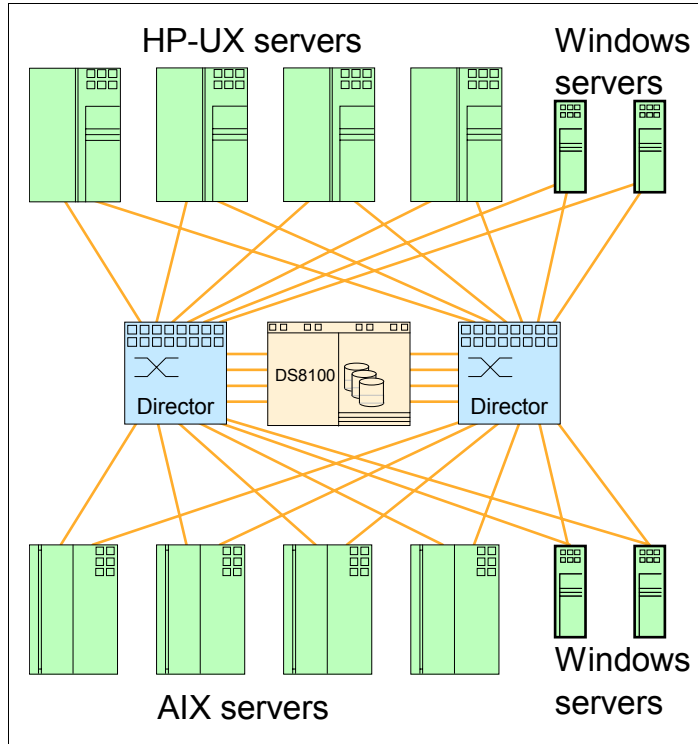


Figure 5-5 Final storage environment

5.3 Long distance disaster recovery over IP

This scenario presents a solution that allows for long distance disaster recovery over an IP connection.

5.3.1 Customer environment and requirements

The customer has three different SAN islands that need to be connected:

- ▶ Development SAN in the primary site
- ▶ Production SAN in the primary site
- ▶ Disaster recovery SAN in the disaster recovery site

The distance between the primary site and the disaster recovery site is 600 km. The amount of data in the productive environments is expected to grow to 5 TB within two years, and we expect that 3% of the data is changing in the peak hour.

The customer has the following requirements for the solution:

- ▶ Provide asynchronous replication for production data from the primary site to the disaster recovery site, with a 5 minute recovery point objective (RPO) and a 5 minute recovery time objective (RTO)
- ▶ Keep the dual fabrics of each SAN both physically and logically separate
- ▶ Provide access to a point-in-time copy of productive data from the test environment at the development SAN
- ▶ Provide for LAN-free backup from the development network to the tape library in productive network

The current environment contains the following components.

- ▶ Production environment at the primary site
 - Dual SAN fabrics, based on IBM TotalStorage SAN 256B directors
 - IBM TotalStorage DS8100 disk subsystem with eight Fibre Channel ports
 - IBM TotalStorage 3584 tape library with six IBM 3592 tape drives
 - Eight pSeries servers, with dual Fibre Channel adapters
 - Sixteen IBM @server xSeries® servers, with dual Fibre Channel adapters
- ▶ Development environment at the primary site
 - Dual SAN fabrics, based on IBM TotalStorage SAN 32B-2 switches
 - IBM TotalStorage DS6800 disk subsystem with four Fibre Channel ports
 - Eight pSeries servers, with dual Fibre Channel adapters
 - Sixteen xSeries servers, with dual Fibre Channel adapters
- ▶ Disaster recovery environment at the disaster recovery site
 - Dual SAN fabrics, based on IBM TotalStorage SAN 256B directors
 - IBM TotalStorage DS8100 disk subsystem with eight Fibre Channel ports
 - IBM TotalStorage 3584 tape library with six IBM 3592 tape drives
 - Eight pSeries servers, with dual Fibre Channel adapters
 - Sixteen xSeries servers, with dual Fibre Channel adapters

The environment is shown in Figure 5-6. For clarity, we show only some of the servers and connections.

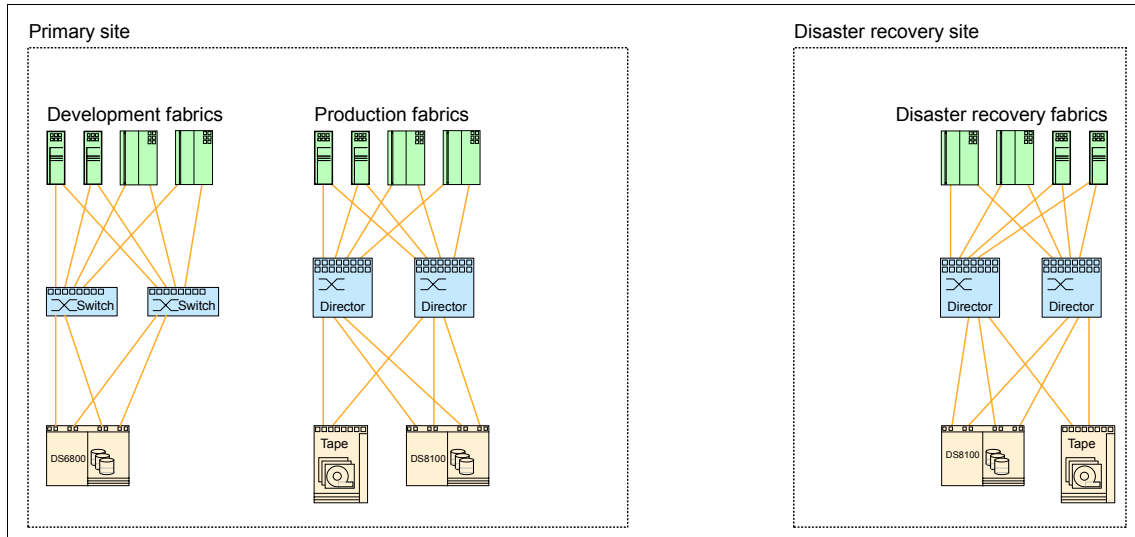


Figure 5-6 Customer environment

5.3.2 The solution

Our solution has the following components:

- ▶ TotalStorage DS8100 Global Mirroring feature for asynchronous replication
- ▶ Four IBM TotalStorage SAN16B-R routers (2109-A16), with FC-FC routing and FCIP tunneling features activated
- ▶ Four IP links between the 2109-A16 routers from the primary site to the disaster recovery site
- ▶ IBM enterprise Remote Copy Management Facility (eRCMF) software to provide automatic fail over of both pSeries and xSeries servers

Figure 5-7 shows the complete solution.

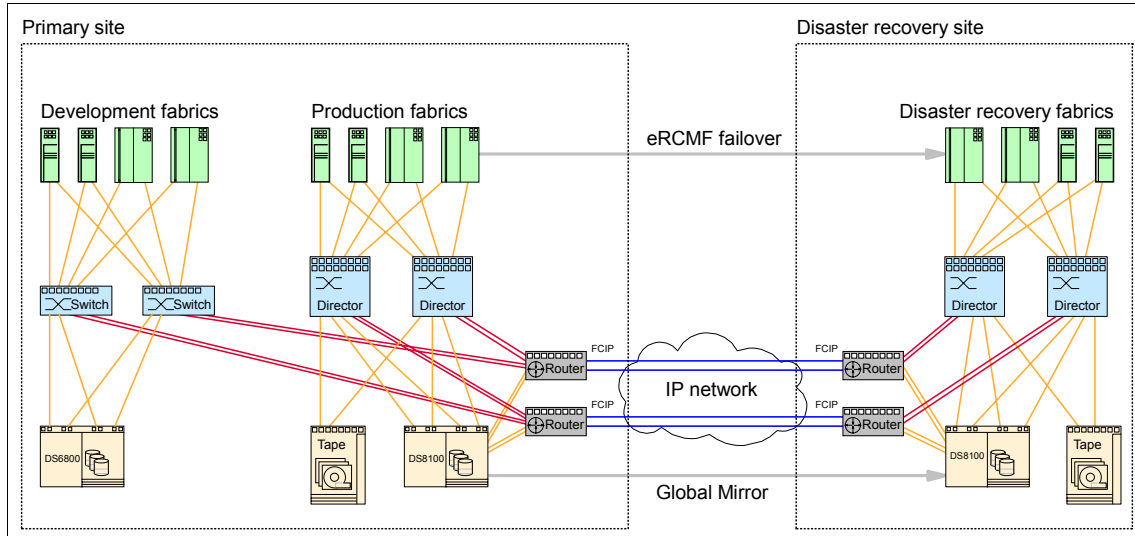


Figure 5-7 Disaster recovery solution

FCIP link sizing

Since we are using only the FCIP links for Global Mirror between the TotalStorage DS8100 systems, we only need to take into account any changes to the data when sizing the links.

Based on the customer requirements, the amount of data changing during the peak hour is 3% of 5 TB, or 150 GB. If we assume that the changes are evenly divided over the hour, the changes are 2.5 GB/min, or approximately 42 MB/s, or 336 Mb/s. We use this number as the basis for our link sizing.

If we divide the amount evenly across four links, we get a traffic of 84 Mb/s over each link. However, to allow the loss of one link or any peaks in the traffic, we divide the traffic only across three links, for 33% extra bandwidth and 112 Mb/s traffic over each link. We also plan to have a maximum of 90% utilization on the link, so the minimum link speed we need is 125 Mb/s.

Each link can be implemented over an OC-3 line, that has the capacity of 155 Mb/s. An alternative is to use a Multiprotocol Label Switching (MPLS)-based, shared connection. However, due to possible router latency issues, we prefer the private OC-3 -based connection.

The most significant part of the OC-3 link latency is the propagation time of the light within the fiber. For a 600 km connection, with 1200 km round trip, it is $1200 \times 4.8 \mu\text{s}$, or 5.8 ms. We round this to 6 ms to account for the packet transmission time over the 155 Mbps OC-3 link.

5.3.3 Normal operation

In normal operation, the production servers use only TotalStorage DS8100 disks in the primary site. The disaster recovery servers are connected to TotalStorage DS8100 in the disaster recovery site, but do not have the disk mounted or any applications running.

The development servers are using TotalStorage DS6800 disks in the primary site, and some capacity from TotalStorage DS8100 in the primary site.

For TotalStorage DS8100 disk subsystems, four of the eight ports are used for host attachments and the remaining four are used for Global Mirroring. The ports used for Global Mirroring are directly connected to the routers.

In addition to normal zoning, we define the following LSANs in our environment:

- ▶ Separate LSANs for the HBAs of any server in the development fabric that needs access to TotalStorage DS8100, containing:
 - The HBA of the server
 - Both Fibre Channel ports of TotalStorage DS8100 used for host attachment in the fabric
- ▶ Separate LSANs for the HBAs of any server in the development fabric that needs access to LAN-free backup, containing:
 - The HBA of the server
 - All Fibre Channel ports of the tape drives in the primary site connected to the fabric

In addition, we define zones for each Global Mirror connection in the backbone fabric.

5.3.4 Failure scenarios

This section explains how the failure of different components affects the operation of our solution.

- ▶ Power failure

All of the SAN fabric components in the environment have dual redundant power supplies connected to different power circuits. Therefore a power failure in one circuit does not have any effect on operation.

- ▶ FCIP link failure

The failure of a single FCIP link reduces the available bandwidth between the sites by 25%. However, since we assumed three available links in our sizing, the performance of the system will remain adequate.

► Development fabric switch failure

The failure of a switch in development fabric reduces the Fibre Channel bandwidth available for development and test servers by 50%. The traffic is automatically routed via the remaining paths by the SDD. The production environment is not affected.

► Primary site router failure

If the router at the primary site fails, the capacity of the Global Mirror connection will be reduced by 50%. However, since we rounded up our link speed, we still have about 300 Mbps or about 90% of the peak hour capacity available. In addition, it reduces the Fibre Channel bandwidth available between development and test servers, and the storage in the production fabrics, by 50%.

► Primary site director failure

Director failure at the primary site reduces the Fibre Channel bandwidth available for production servers by 50%. It also reduces the Fibre Channel bandwidth available between development and test servers, and the storage in the production fabrics, by 50%.

► Disaster recovery site router failure

If the router at the disaster recovery site fails, the capacity of the Global Mirror connection is reduced by 50%. However, since we rounded up our link speed, we still have about 300 Mbps or about 90% of the peak hour capacity available.

► Disaster recovery site director failure

Director failure at the disaster recovery site reduces the Fibre Channel bandwidth available for disaster recovery servers by 50%. However, in normal situations, those servers are idle, so this reduction affects the system only in the case where the production workload is already running at the disaster recovery site.

► Primary site TotalStorage DS8100 port failure

If a port used for host access in TotalStorage DS8100 at the primary site fails, the Fibre Channel bandwidth available for host access is reduced by 25%.

If a port that is used for Global Mirror in TotalStorage DS8100 at the primary site fails, the remaining Fibre Channel ports can sustain the full Global Mirror performance.

► Primary site TotalStorage DS8100 failure

If TotalStorage DS8100 at the primary site fails, all hosts lose access to it. This event can be promoted to site failure, and production can resume at the disaster recovery site.

- ▶ Primary site TotalStorage DS8100 port failure

If a port used for host access in TotalStorage DS8100 at the disaster recovery site fails, the Fibre Channel bandwidth available for host access is reduced by 25%. However, in normal operation, those servers are idle, so this reduction affects the system only in the case where the production workload is already running at the disaster recovery site.

If a port that is used for Global Mirror in TotalStorage DS8100 at the primary site fails, the remaining Fibre Channel ports can sustain the full Global Mirror performance.

- ▶ Disaster recovery site TotalStorage DS8100 failure

If TotalStorage DS8100 at the disaster recovery site fails, the Global Mirror connections change to a *suspended* state. TotalStorage DS8100 at the primary site will accumulate changes to the data, and copy the changed data over to the disaster recovery site, when TotalStorage DS8100 becomes available.

- ▶ Primary site failure

If the complete primary site fails, the IBM eRCMF software starts the production at the disaster recovery site automatically. While manual failover is also possible, it is difficult to manually reach the RTO target.



Cisco family routing products

This chapter provides information about the routing functions of the Cisco MDS 9000 Multilayer switches including:

- ▶ Hardware and software
- ▶ Advanced management
- ▶ Key features
- ▶ Interoperability

6.1 Overview of the Cisco MDS family

The Cisco MDS family is fundamentally different from products designed by other vendors. Routing features are inherent in every one of Cisco's MDS switches, so switch and routing functions do not need to be separated into multiple devices.

Cisco was the first to implement virtual storage area network (VSAN), which has now been adopted as an official standard by the industry. In addition, Cisco is the first to implement Inter-VSAN Routing (IVR) on every port and to support Small Computer System Interface over IP (iSCSI) and Fibre Channel (FC) over IP (FCIP) over Gigabit Ethernet.

The Cisco Systems MDS 9020 (2061-420), MDS 9120 (2061-020), MDS 9140 (2061-040), MDS 9216x (2062-D1A/D1H) Multilayer Fabric Switches, MDS 9506 (2062- D04/T04), and MDS 9509 (2062-D07/T07) Multilayer Directors are available from IBM and authorized IBM Business Partners.

Note: Every Fibre Channel port on a Cisco switch is an FC to FC routing port.

6.1.1 Introduction to VSAN

VSAN technology is designed to enable efficient storage area network (SAN) use by dividing a physical fabric into multiple logical fabrics. Each VSAN can be zoned as a typical SAN and maintains its own fabric services for added scalability and resilience.

VSAN is a standard feature of all Cisco MDS switches. Transmitting data between VSANs requires the IVR feature, which is included with the Enterprise license option. The Cisco MDS family also supports routing on every port, including routing to IBM m-type and IBM b-type switches.

SAN-OS 2.1 (released in March 2005) provides support for network address translation (NAT). NAT allows routing between switches with the same domain identifier and routing between VSANs with the same VSAN identifier. SAN-OS 2.1 includes performance improvements for iSCSI and FCIP transmissions and some management enhancements.

SAN-OS 2.1 also provides support for the new Storage Services Module (SSM), which implements the SANTap protocol, Fabric Application Interface Standard (FAIS), and Network Accelerated Serverless Backup (NASB). Aside from Fibre Channel Fast Write, the SSM requires layered applications from third-party vendors to deliver end-user functionality.

The following sections discuss the Cisco products that are available as part of the IBM Storage Networking solutions portfolio.

6.2 Hardware and software

The following section describes the software and hardware of the Cisco MDS 9000 Multilayer Switch family.

Important: Not all features are included with the hardware. It is likely that you need to purchase additional software licenses depending on your requirements. Consult your IBM representative for more details.

6.2.1 Cisco MDS 9120 and 9140 Multilayer Switches

The Cisco MDS 9120 Multilayer Fabric Switch (IBM 2061-020) and Cisco MDS 9140 Multilayer Fabric Switch (IBM 2061-040) are one rack-unit (RU) fabric switches that can support 20 or 40 shortwave or longwave small form-factor pluggable (SFP) fiber optic transceivers. Some of these ports operate with a 3.2 to one (3.2:1) over-subscription (fanout) and are referred to as *host optimized ports*.

The MDS 9120 has a total of 20 ports. The first group of four ports on the left side are full bandwidth ports and are identified by a white border. The remaining four groups of ports are host optimized port groups.

Cisco MDS cooling and airflow: MDS 9120 and MDS 9140 switches use what Cisco calls *front-to-rear airflow for cooling*. Be careful because the *front* is where the Fibre Channel cables are. Only the power cables are in the back. If you install the switches with the ports facing the back for ease of server cabling, then the switches will suck in hot air from the servers and overheat.

If you mount the switches with the ports to the front, as Cisco recommends, you may need to plan cable management carefully since cables need to connect from the front of the rack to the back of the rack where the server ports are. Alternatives include mounting the switches in a separate communications rack, or mounting the switches with the ports facing the back at the *bottom* of the server rack. However, this may be less convenient for access and does not comply with the best practice of mounting the heaviest devices at the bottom.

By way of contrast, the MDS 92xx and MDS 95xx use *right-to-left cooling*, looking from the front (which is the ports side).

Figure 6-1 shows the MDS 9120 switch.



Figure 6-1 MDS 9120 Multilayer Switch (IBM 2061-020)

The MDS 9140 has a total of 40 ports. The first eight ports on the left side are full bandwidth ports and are identified by a white border. The remaining eight groups of ports are host optimized port groups. Figure 6-2 shows the MDS 9140 switch.



Figure 6-2 MDS 9140 Multilayer Switch (IBM 2061-040)

The switches are configured with dual redundant power supplies either of which can supply power for the whole switch. They include a hot-swappable fan tray to manage the cooling and airflow for the entire switch.

The 91n0 switches share a common firmware architecture with the Cisco MDS 9500 series of multilayer directors, making them intelligent and flexible fabric switches.

Note: The MDS9120 and MDS9140 also both support optional coarse wavelength division multiplexing (CWDM) SFPs to provide aggregation of multiple links onto a single optical fiber through a passive optical mux. This is a unique Cisco feature, which allows architects to design relatively low cost CWDM solutions around Cisco equipment.

6.2.2 MDS 9216A Multilayer Switch

The Cisco MDS 9216A Model D01 (IBM 2062-D1A) is a three RU, 2-slot fabric switch that can support from 16 to 48 shortwave or longwave SFP fiber optic transceivers. These ports fully support either 1 Gbps or 2 Gbps Fibre Channel and are auto-sensing. Figure 6-3 shows the MDS 9216A switch.



Figure 6-3 MDS 9216A Multilayer Switch (IBM 2062-D1A) with 48 ports

The chassis consists of two slots. The first slot contains the supervisor module. This provides the control and management functions for the 9216A and includes 16 standard Fibre Channel ports. It contains 2 GB of DRAM and has one internal CompactFlash card that provides 256 MB of storage for the firmware images.

The second slot can contain any one of the modules described in 6.2.6, “Optional modules” on page 74.

The MDS9216A also supports optional CWDM SFPs to provide aggregation of multiple links onto a single optical fiber through a passive optical mux.

6.2.3 Cisco MDS 9216i Multilayer Switch

The Cisco MDS 9216i uses the same backplane as the MDS 9216A. However, the MDS 9216i includes a fixed 14+2 supervisor module to provide 14 full capability target-optimized Fibre Channel ports and two Gigabit Ethernet interfaces. The Gigabit Ethernet interfaces support iSCSI initiators connecting to Fibre Channel disk systems. They also support FCIP, which was previously licensed separately, but is now included with the base unit. The MDS 9216i accepts any MDS optional modules into its second slot.

Note: FCIP and IVR are now both included in the base functionality of the MDS 9216i for FCIP ports only, without needing to purchase the Enterprise licensing package.

FCIP can help to simplify data protection and business continuance strategies by enabling backup, remote replication, and other disaster recovery services over wide area network (WAN) distances using open-standard FCIP tunneling. Figure 6-4 shows the MDS 9216i.



Figure 6-4 Cisco MDS 9216i

Both models accommodate expansion with the full line of optional switching modules and IP multiprotocol switching modules.

MDS 9216i supports the following features:

- ▶ Integrated IP and Fibre Channel SAN solutions
- ▶ Simplified large storage network management and improved SAN fabric utilization helping to reduce total cost of ownership
- ▶ Throughput of up to 2 Gbps per port and up to 32 Gbps with each PortChannel inter-switch link (ISL) connection
- ▶ Scalability
- ▶ Gigabit Ethernet ports for iSCSI or FCIP connectivity
- ▶ Modular design with excellent availability capabilities
- ▶ Intelligent network services that help simplify SAN management and reduce total cost
- ▶ Assistance with security for large enterprise SANs
- ▶ VSAN capability for creating separate logical fabrics within a single physical fabric
- ▶ Compatibility with a broad range of IBM servers, as well as disk and tape storage devices
- ▶ Hardware-based encryption and compression to ensure secure IP links, as well as reducing the bandwidth requirements (and thus cost) on the IP link
- ▶ Up to 255 buffer credits per port

The MDS 9216i also supports optional CWDM SFPs to provide aggregation of multiple links onto a single optical fiber through a passive optical mux.

6.2.4 MDS 9506 Multilayer Director

The Cisco MDS 9506 (IBM 2062-D04) is a seven RU Fibre Channel director that can support from 32 to 128 shortwave or longwave SFP fiber optic transceivers. These ports fully support either 1 Gbps or 2 Gbps Fibre Channel and are auto-sensing.

The chassis has six slots, two of which are reserved for dual, redundant supervisor modules. The dual supervisor modules provide the logic control for the director. They also provide high availability and traffic load balancing capabilities across the director. Either supervisor module can control the whole director, with the standby supervisor module providing full redundancy in the event of an active supervisor failure.

The remaining four slots can contain a mixture of switching modules which provide either 16 or 32 ports per module, 4 or 8 port IP storage services modules, and virtualization Caching Services Modules.

The director is configured with dual, redundant power supplies, either of which can supply power for the whole chassis. It also includes a hot-swappable fan tray that manages the cooling and right to left (looking from the SFP side) airflow for the entire director.

Figure 6-5 shows the MDS 9506 Multilayer Director.



Figure 6-5 MDS 9506 Multilayer Director (IBM 2062-D04)

The IBM 2062-T04 product is designed for the telecommunications industry and ships with -48 to -60V dc fed 1900W power supplies. This is the only difference when compared to the 2062-D04.

The MDS 9506 also supports optional CWDM SFPs to provide aggregation of multiple links onto a single optical fiber through a passive optical mux.

6.2.5 MDS 9509 Multilayer Director

The Cisco MDS 9509 Model D07 (IBM 2062-D07) is a fourteen RU Fibre Channel director that can support from 32 to 224 shortwave or longwave SFP fiber optic transceivers. These ports fully support either 1 Gbps or 2 Gbps Fibre Channel and are auto-sensing.

Figure 6-6 shows the MDS 9509 Multilayer Director. The chassis has nine slots, two of which are reserved for dual, redundant supervisor modules. The dual supervisor modules provide the logic control for the director and provide high availability and traffic load balancing capabilities across the director. Either supervisor module can control the whole director, with the standby supervisor module providing full redundancy in the event of an active supervisor failure.



Figure 6-6 MDS 9509 Multilayer Director (IBM 2062-D07)

The backplane of the 9509 provides the connectivity for two supervisor modules and up to seven switching modules. In addition to the supervisor and switching modules, the redundant power supplies and the redundant, dual clock modules also plug directly into the backplane. If one clock module fails, the remaining clock module takes over operation of the director.

Note: Although there are dual redundant clock modules in the Cisco MDS950x directors, if one clock module needs to be replaced, a director outage is required because these modules are not hot-pluggable.

The remaining seven slots can contain a mixture of switching modules which provide either 16 or 32 ports per module, 4 or 8-port IP storage services modules, and advanced modules for virtualization and replication services.

The IBM 2062-T07 is designed for the telecommunications industry. It ships with -48 to -60V dc fed 2500W power supplies. This is the only difference when compared to the 2062-D07.

Supervisor module

The supervisor module is the heart of the 9500 series directors because it provides the control and management functions for the director, as well as an integrated crossbar switching fabric. The crossbar fabric provides up to 720 Gbps full duplex switching capacity.

The MDS 9500 comes standard with two supervisor modules for redundancy and availability. In the event of a supervisor module failing, the surviving module becomes active, taking over the operation of the director.

Note: The MDS 9216A uses a different supervisor module than the one used by the 9509, which integrates 16 target optimized ports. The function provided by the MDS 9216x supervisor is the same as the one described in this section.

Control and management

The supervisor module provides the following control and management features:

- ▶ Multiple paths avoid a single point of failure
- ▶ A redundant central arbiter provides traffic control and access fairness.
- ▶ It performs a nondisruptive restart of a single failing process on the same supervisor. A kernel service running on the supervisor module keeps track of the high availability policy of each process and issues a restart when a process fails. The type of restart issued is based on the process's capability:
 - Warm or stateful (state is preserved)
 - Cold or stateless (state is not preserved)

If the kernel service is unable to perform a warm restart of the process, it issues a cold restart.

- ▶ It performs a nondisruptive switchover from the active supervisor to a redundant standby without loss of traffic.

If the supervisor module has to be restarted, then the secondary supervisor (continuously monitoring the primary) takes over. Switchover is non-revertive. After a switchover has occurred and the failed supervisor is replaced or restarted, the operation does not switch back to the original primary supervisor, unless it is forced to switch back or unless another failure occurs.

Crossbar switching fabric

The MDS 9500 supervisor module provides a crossbar switching fabric that connects all the modules. A single crossbar provides 720 Gbps full-duplex speed allowing 80 Gbps bandwidth per switching module.

Dual supervisor configurations provide 1.4 Tbps throughput with a 160-Gbps bandwidth per switching module.

Figure 6-7 shows the 9500 series supervisor module.



Figure 6-7 MDS 9500 Series supervisor module

6.2.6 Optional modules

The MDS 9200 and 9500 families allow for optional modules to provide additional port connectivity, IP services, or storage virtualization functionality into empty expansion slots.

The MDS 9216x can accept one optional module, while the MDS 9506 can accept four. The MDS 9509 supports up to a maximum of seven optional modules.

The 16-port switching module

The 16-port switching module provides up to 64 Gbps of continuous aggregate bandwidth. Autosensing 1 Gbps and 2 Gbps target-optimized ports deliver 200 MB/s and 255 buffer credits per port.

Note: The 64-Gbps, continuous, aggregate bandwidth is based on 2 Gbps per port in full duplex mode, that is:

16 ports at 2 Gbps (or 213 MB/s) in both directions = 64 Gbps

The 16-port module is designed for attaching high-performance servers and storage subsystems, and for connecting to other switches using ISL connections. This module also supports optional CWDM SFPs to provide aggregation of multiple links onto a single optical fiber through a passive optical mux.

Figure 6-8 shows the 16-port switching module for the Cisco MDS 9000 family.



Figure 6-8 16 port switching module

The 32-port switching module

The 32-port switching module is designed to deliver an optimal balance of performance and port density. This module provides high line-card port density along with 64 Gbps of total bandwidth and 12 buffer-to-buffer credits per port. Bandwidth is allocated across eight 4-port groups, with each port group sharing 2.5 Gbps, making it an aggregate bandwidth of approximately 5 Gbps full-duplex. This module provides a low-cost means to attach lower performance servers and storage subsystems to high-performance crossbar switches without requiring ISLs.

By combining 16- and 32-port switching modules in a single, modular chassis, administrators can configure price and performance-optimized storage networks for a wide range of application environments.

The 32-port switching module also supports optional CWDM SFPs to provide aggregation of multiple links onto a single optical fiber through a passive optical mux.

Switching modules are designed to be interchanged or shared between all Cisco MDS 9200 switches and 9500 directors. Figure 6-9 shows the 32-port switching module for the Cisco MDS 9000 family.

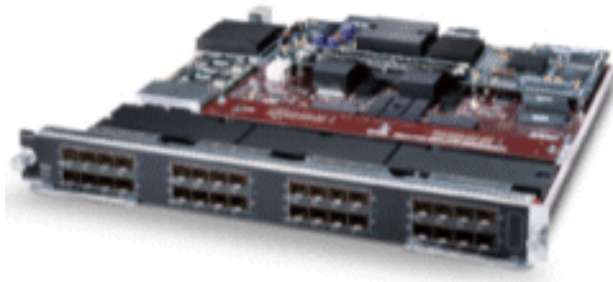


Figure 6-9 32 port switching module

Cisco MDS 9000 14+2 Multiprotocol Services Module

The Cisco MDS 9000 14+2 Multiprotocol Services Module is designed to provide fourteen Fibre Channel and two IP storage interfaces. The 14 Fibre Channel ports are based around the same full rate target optimized ports as the 16-port module, providing all the same operating modes. In addition the 14+2 card can be configured with high buffer credits on one Fibre Channel port, to support longer distance FC to FC connections.

The two IP storage interfaces are similar to the IP Services Module, including hardware compression and security.

Restriction: The two Ethernet ports on the 14+2 Multiprotocol Services Module *cannot* be combined into a single EtherChannel. However, PortChannel can be used.

Figure 6-10 shows the Cisco MDS 9000 14+2 Multiprotocol Services Module.

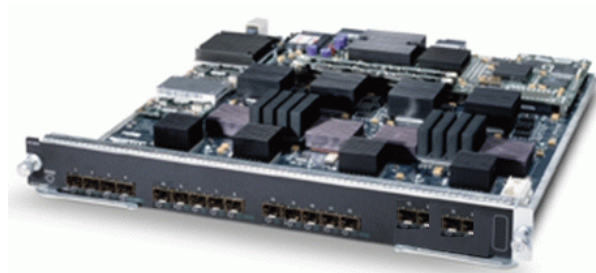


Figure 6-10 Cisco MDS 9000 14+2 Multiprotocol Services Module

This module also supports optional CWDM SFPs to provide aggregation of multiple links onto a single optical fiber through a passive optical mux.

IP Services Module

The IP Services (IPS) Module is available in two versions, IPS-4 or IPS-8, and provides four or eight Gigabit Ethernet ports that can support iSCSI and FCIP protocols simultaneously. Because the bit rate of Gigabit Ethernet is different from the bit rate of Fibre Channel, the card requires tri-rate SFPs.

Figure 6-11 shows the 8-port IP Services Module.

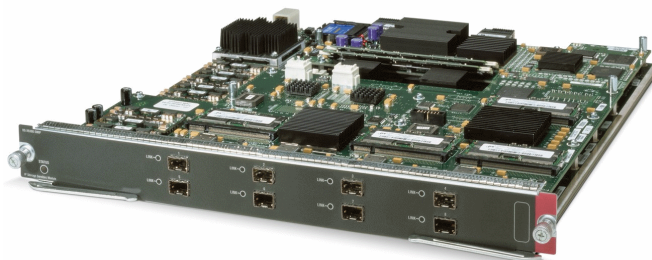


Figure 6-11 The 8-port IP Services Module

Note: Two Ethernet ports on the IPS modules *can* be combined into a single EtherChannel, but only between ports that share the same application-specific integrated circuit (ASIC). However, PortChannel can be used.

Ports configured to run FCIP

The ports configured for FCIP can support up to three, virtual ISL connections (FCIP tunnels). This way you can transport Fibre Channel traffic transparently, except for latency, over an IP network between two FCIP-capable switches. Each virtual ISL connection acts as a normal Fibre Channel ISL or extended ISL (EISL).

To use FCIP, you need to purchase the FCIP Activation for 8-port IP Services Line Card feature for every 8-port IP line card that needs to support FCIP.

Ports configured to run iSCSI

Ports configured to run iSCSI work as a gateway between iSCSI hosts and Fibre Channel attached targets. The module terminates iSCSI commands and issues new Fibre Channel commands to the targets.

The Cisco Fabric Manager is used to discover and display iSCSI hosts. These iSCSI hosts are bound to assigned worldwide names (WWNs) and create a static relationship that enables:

- ▶ Zoning of iSCSI initiators
- ▶ Accounting against iSCSI initiators
- ▶ Topology mapping of iSCSI initiators

In SAN-OS 2.1, iSCSI connections for each IPS port have the following theoretical limitations:

- ▶ 200 simultaneous connections (one per client)
- ▶ 2000 configured initiators
- ▶ 5000 simultaneous connections per switch/director

Important: The 200 simultaneous iSCSI connections per port is a theoretical limit. The practical limit is generally much lower. Architects should calculate the peak traffic volumes and overheads for each specific customer environment.

Storage Services Module

The SSM is based on the 32-port Fibre Channel Switching Module and provides intelligent storage services in addition to 1 Gbps and 2 Gbps Fibre Channel switching. The SSM uses eight IBM PowerPC processors for SCSI data-path processing. It can be combined with the optional Cisco MDS 9000 Enterprise Package to enable Fibre Channel Write Acceleration (FC-WA).

FC-WA can help improve the performance of remote mirroring applications over extended distances by reducing the effect of transport latency when completing a SCSI operation over distance. This supports longer distances between primary and secondary data centers and can help improve disk replication performance.

The optional Storage Systems Enabler Package Bundle can enable independent software vendors (ISVs) to develop intelligent fabric applications that can be hosted on the SSM through an application programming interface (API).

ISVs may use the API to offer the following applications:

- ▶ Network-accelerated storage applications, such as serverless backup
- ▶ Network-assisted appliance-based storage applications using Cisco MDS 9000 SANTap Service, such as global data replication
- ▶ Network-hosted storage applications based upon proposed Fabric Application Interface Standard (FAIS) APIs offered by ISVs

Note: IBM support for these ISV applications is limited to IBM TotalStorage Proven™ solutions. For the most current IBM TotalStorage Proven information, go to:

<http://www.ibm.com/storage/proven>

Figure 6-12 shows the Storage Services Module.



Figure 6-12 Storage Services Module

Buffer credits

Buffer credits affect the number of input/output (I/O) that can be sent before an acknowledgement is received. In extended Fibre Channel networks, you need more buffer credits to keep the “pipe” filled because the latency has increased.

Each target optimized port supports 255 buffer credits, and host-optimized ports support 12 buffer credits per port. On the 14+2 line card, up to 3500 buffer credits can be assigned to a single port if you are willing to sacrifice buffers on other ports and shut down three ports on the quad controlled by that ASIC. A maximum of 1500 buffer credits can be configured if the additional three ports are left enabled.

6.3 Advanced management

The Cisco SAN-OS is the operating system running within the MDS 9000 supervisor and modules to enable the multilayer functionality of the products. Cisco SAN-OS provides a rich suite of management tools.

While the SAN-OS installable files are specific to each MDS 9000 platform (9100, 9200, and 9500), the standard features provided by the SAN-OS are common to all, although some features are applicable only to switches with Ethernet ports. Features include support for:

- ▶ Fibre Channel Protocol (FCP)
- ▶ iSCSI
- ▶ VSANs
- ▶ Zoning
- ▶ FCC
- ▶ Virtual Output Queuing
- ▶ Diagnostics (SPAN, RSPAN, and so on)
- ▶ SNMPv3
- ▶ SSH
- ▶ SFTP
- ▶ RBAC

- ▶ Radius
- ▶ High Availability
- ▶ PortChannels
- ▶ RMON
- ▶ Call home
- ▶ TACACS+
- ▶ FDMI
- ▶ SMI-S (XML-CIM)
- ▶ iSNS Client
- ▶ iSNS
- ▶ IPS ACLs
- ▶ Fabric Manager

The SAN OS 2.1 includes the following new features:

- ▶ Heterogeneous IVR
- ▶ WWN-based VSANs
- ▶ Zone-based Quality of Service (QoS)
- ▶ Auto-creation of PortChannels
- ▶ Enhanced zoning (locking)
- ▶ Cisco fabric services (lock and apply changes across the fabrics)

6.3.1 Fabric management

The Cisco MDS 9000 family provides three modes of management:

- ▶ The MDS 9000 family command line interface (CLI) presents the user with a consistent, logical CLI, which adheres to the syntax of the widely known Cisco IOS CLI. This is an easy-to-use command interface which has broad functionality.
- ▶ The Cisco Fabric Manager is a Java application that simplifies management across multiple switches and fabrics. It enables administrators to perform such tasks as topology discovery, fabric configuration and verification, provisioning, monitoring, and fault resolution. All functions are available through a remote management interface.
- ▶ Cisco also provides an API for integration with third-party and user developed management tools.

Cisco MDS 9000 Fabric Manager

The Cisco Fabric Manager is included with the Cisco MDS 9000 family of switches and is a Java and Simple Network Management Protocol (SNMP)-based network fabric and device management tool. It provides a GUI that displays real-time views of your SAN fabric and installed devices. The Cisco Fabric Manager provides three views for managing your network fabric:

- ▶ The Device View displays a continuously updated physical picture of device configuration and performance conditions for a single switch.
- ▶ The Fabric View displays a view of your network fabric, including multiple switches.
- ▶ The Summary View presents a summary view of switches, hosts, storage subsystems, and VSANs.
- ▶ The Cisco Fabric Manager provides an alternative to the CLI for most switch configuration commands.

The Cisco Fabric Manager is included with each switch in the Cisco MDS 9000 family. Figure 6-13 shows the Fabric Manager user interface.

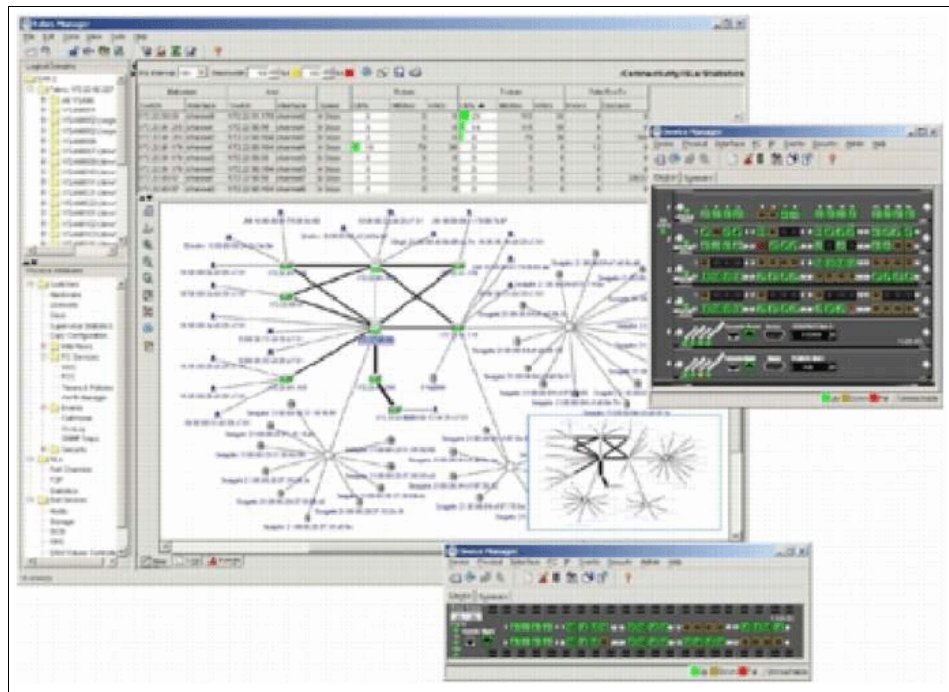


Figure 6-13 Cisco MDS 9000 Fabric Manager user interface

In-band management and out-of-band management

The Cisco Fabric Manager requires an out-of-band (Ethernet) connection to at least one Cisco MDS 9000 family switch to enable it to discover and manage the fabric.

The interface used for an out-of-band management connection is a 10/100 Mbps Ethernet interface on the supervisor module, labeled mgmt0. The mgmt0

connection can be connected to a management network to access the switch through IP over Ethernet.

Ethernet connectivity is required to at least one Cisco MDS 9000 family switch. This connection is then used to manage the other switches using in-band (Fibre Channel) connectivity. Otherwise, you need to connect the mgmt0 port on each switch to your Ethernet network.

Each supervisor module has its own Ethernet connection, However, the two connections in a redundant supervisor system operate in active or standby mode. The active supervisor module also hosts the active mgmt0 connection. When a failover event occurs to the standby supervisor module, the IP address and medium access control (MAC) address of the active Ethernet connection are moved to the standby Ethernet connection. This eliminates any need for the management stations to relearn the location of the switch.

Figure 6-14 shows an example of an out-of-band management solution.

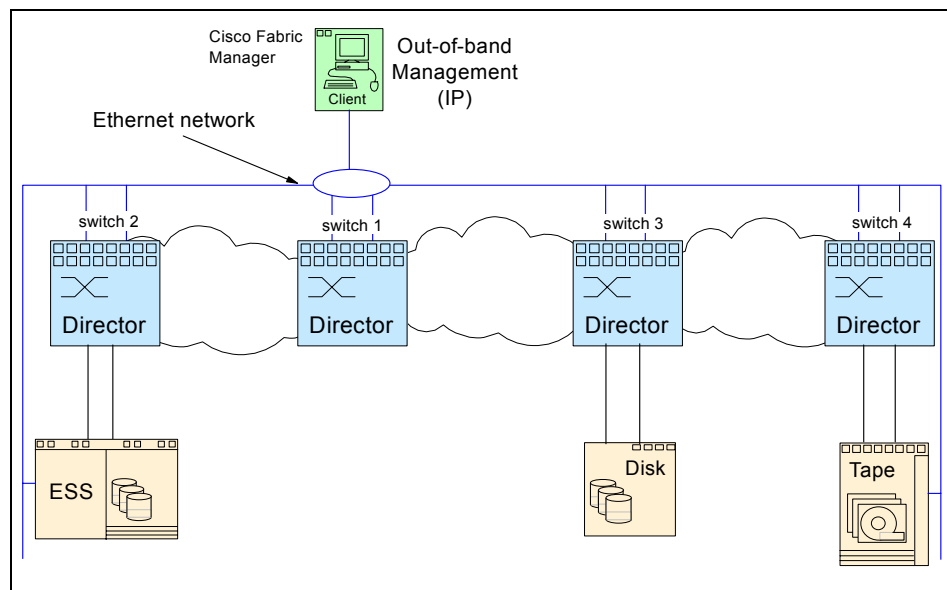


Figure 6-14 Out-of-band management connection

You can also manage switches on a Fibre Channel network using an in-band connection to the supervisor module. This in-band connection supports either management protocols over Fibre Channel or IP embedded within Fibre Channel. The Cisco MDS 9000 family supports RFC 2625 IP over Fibre Channel (IPFC), which allows IP to be transported between Fibre Channel devices over the Fibre Channel protocol, as shown in Figure 6-15.

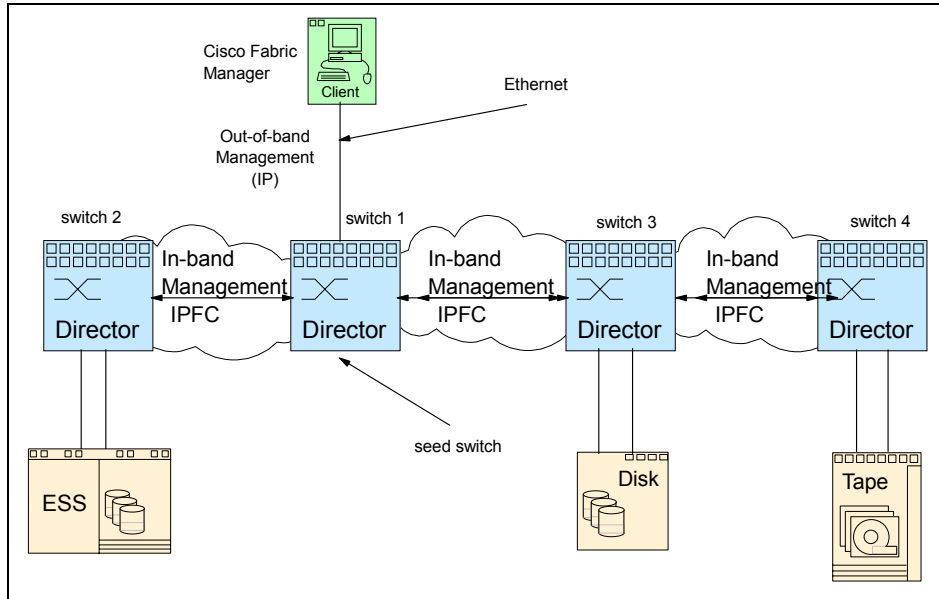


Figure 6-15 In-band management connection

IPFC encapsulates IP packets into Fibre Channel frames so that management information can cross the Fibre Channel network without requiring a dedicated Ethernet connection to each switch. IP addresses are resolved to the Fibre Channel address through the Address Resolution Protocol (ARP). With host bus adapters (HBAs) that support IP drivers, this capability allows for a completely in-band management network. The switch also uses the in-band interface to discover its own environment, including directly connected and fabric-wide elements.

Cisco now also provide the capability to assign an IP address to each VSAN and manage each VSAN in-band or out-of-band.

Note: When initially setting up a Cisco MDS Multilayer switch, all ports are by default in VSAN1, and the speed and type are set to *autosense*.

Role-based management

The Cisco MDS 9000 family switches support role-based management access with the CLI or the Cisco Fabric Manager. This lets you assign specific management privileges to particular roles and then assign one or more users to each role.

Roles can also be assigned on a per-VSAN basis. For example, the administrator of one VSAN does not need to be given administrator privileges on other VSANs.

6.3.2 Optional licensed feature packages

In addition to the standard Fabric Manager features provided in the SAN-OS 2.1, there are four optional licensed feature packages to address different enterprise requirements. The four optional licensed packages are:

- ▶ FCIP Activation (one FCIP license included with each MDS 9216i)
- ▶ Enterprise Package
- ▶ Fabric Manager Server (FMS)
- ▶ Mainframe Package

These packages have per-switch licensing, except for FCIP Activation which is per-line-card licensing.

SAN Extension over IP Package for SAN-OS 2.1

This package contains the following features and benefits:

- ▶ FCIP can be used to connect Fibre Channel across a distance using IP networks. Each Cisco MDS 9000 Family Gigabit Ethernet port is capable of managing up to three FCIP tunnels. Each 8-port IP Storage Services Module supports 24 simultaneous FCIP tunnels (12 simultaneous tunnels per 4-port module).
- ▶ FCIP Compression in the Cisco MDS 9000 Family SAN-OS increases the effective WAN bandwidth. Gigabit Ethernet ports for IP Storage Services can theoretically achieve up to a thirty to one (30:1) compression ratio, but typical ratios of less than two to one (2:1) are more likely to be achieved in the field.
- ▶ IVR for FCIP allows selective transfer of data traffic between specific initiators and targets on different VSANs without merging VSANs into a single logical fabric. IVR can be used in conjunction with FCIP to increase the resiliency of SAN Extension over IP networks and create more efficient business continuity and disaster recovery solutions. To use IVR for Fibre Channel, the Enterprise package is also required.
- ▶ FCIP Write Acceleration can significantly improve application performance when storage traffic is routed over WANs using FCIP. When FCIP Write Acceleration is enabled, write I/O latency is decreased by minimizing the impact of WAN latencies.
- ▶ FCIP Tape Acceleration allows servers to transfer data across a WAN to streaming tape drives, which require a continuous flow of data to avoid write data underruns (dramatically reduce write throughput). Without FCIP Tape

Acceleration, the effective WAN throughput for remote tape backup decreases exponentially as the WAN latency increases.

- ▶ Cisco's SAN Extension Tuner helps optimize FCIP performance. The SAN Extension Tuner generates SCSI I/O commands that are directed to a specific virtual target. It reports I/Os per second and I/O latency results.

Note: The *SAN Extension over IP Package* is usually referred to by IBM as *FCIP Activation*.

Enterprise Package for SAN-OS 2.1

The Enterprise Package optional license enables the following features:

- ▶ Inter-VSAN Routing
- ▶ Zone-based QoS as well as port-based, Fibre Channel ID-based and VSAN-based
- ▶ Extended buffer credits on the 14+2 Multiprotocol card and 9216i
- ▶ FC-WA when used with the Storage Services Module
- ▶ SCSI flow statistics when used with the Storage Services Module
- ▶ Switch/Switch and Host/Switch authentication using Fibre Channel Security Protocol
- ▶ LUN zoning
- ▶ Read-only zones
- ▶ Port security
- ▶ VSAN-based access control
- ▶ IP security protocol (IPsec)

Note: To enable FC-WA, you must also purchase the Storage Services Module and the Storage Services Enablement package bundle.

Fabric Manager Server for SAN-OS 2.1

The standard Cisco Fabric Manager software that is included at no charge with the Cisco MDS 9000 Family multilayer switches provides basic switch configuration and troubleshooting capabilities. The Cisco MDS 9000 Family FMS package extends standard Cisco Fabric Manager by providing historical performance monitoring for network traffic hotspot analysis, centralized management services, and advanced application integration.

The FMS license enables the following additional features:

- ▶ Fibre Channel Statistics Monitoring provides continuous performance statistics Fibre Channel connections.
- ▶ Performance Thresholds allow the administrator to set two different event thresholds for each throughput statistic monitored by the Cisco FMS.

Threshold values can be set with user-specified levels or with baseline values automatically calculated from performance history.

- ▶ Reporting and Graphing provides historical performance reports and graphs over daily, weekly, monthly, and yearly intervals for network hotspot analysis. Top 10 and daily summary reports for all ISLs, hosts, storage connections, and flows provide fabric-wide statistics.
- ▶ Intelligent Setup Wizards are provided to quickly select information to monitor, set up flows, and estimate performance-database storage requirements. Statistics are associated with host and storage devices, allowing physical connections to switches to be changed without losing historical statistics.
- ▶ Performance Database provides a compact round robin database (RRD) maintained at a constant size by rolling up information to reduce the number of discrete samples for the oldest data points. Therefore, it requires no manual storage-space maintenance.
- ▶ Web-Based Operational View provides a Web-browser interface to historical performance statistics, SAN inventory, and fabric event information needed for day-to-day operations.
- ▶ Multiple Fabrics Management allows for multiple Fibre Channel fabrics to be monitored by each management server.
- ▶ Continuous Health and Event Monitoring is enabled via SNMP traps and polling, instead of only when the application user interface is open.
- ▶ Common Discovery runs a centralized background discovery of Fibre Channel HBAs, storage devices, and switches.
- ▶ Roaming User Profiles allow user preference settings and topology-map layout changes to be applied whenever the Cisco Fabric Manager client is opened. It maintains a consistent interface regardless of which computer is used for management.
- ▶ FMS Proxy Services help isolate a private IP network used for Cisco MDS management from the LAN or WAN used for remote connectivity.
- ▶ Cisco Traffic Analyzer Integration provides an easy drill down to SCSI I/O or Fibre Channel frame-level details.
- ▶ Management Server allows a server to be set up to continuously run Cisco FMA. Up to 16 remote Cisco Fabric Manager user interface clients can access this management server concurrently.

Mainframe package

The Cisco MDS 9000 Family Mainframe package is a collection of features required for using the Cisco MDS 9000 Family switches in mainframe storage networks. IBM Fibre Connection (FICON®) is an architecture for high-speed

connectivity between mainframe systems and I/O devices. With the Mainframe package, the Cisco MDS 9000 Family has the capability to simultaneously support the FCP, iSCSI, FCIP, and FICON protocols.

Applying the Mainframe Package optional license enables all FICON requirements with a single license key. The mainframe package optional license enables the following features:

- ▶ FICON Control Unit Port (CUP) for in-band management of the switch from FICON hosts
- ▶ The Fabric Binding feature to help ensure that ISLs are enabled only between switches that have been authorized in the fabric binding configuration
This feature helps prevent unauthorized switches from joining the fabric or disrupting current fabric operations.
- ▶ The Switch Cascading feature to support FICON hosts accessing devices that are connected through ISLs
- ▶ VSAN support of FICON and FCP intermixed environments to provide separation of FCP and FICON traffic and to protect the mainframe environment from instability or excessive control traffic
 - Qualified with IBM TotalStorage Virtual Tape Server (VTS) and IBM TotalStorage Peer-to-Peer Virtual Tape Server
 - Qualified with IBM TotalStorage Extended Remote Copy (XRC) for z/OS®
- ▶ FICON Native Mode and Native Mode Channel-to-Channel operation
- ▶ Persistent FICON Fibre Channel ID (FCID) assignment
- ▶ Port swapping for host-channel cable connections

Note: A license is required for each switch that participates in a FICON-cascaded fabric.

6.4 Key features

The following sections discuss some of the features of Cisco's MDS 9000 multilayer switches, which are important in a routing environment.

6.4.1 Protocol support

The MDS 9000 family supports the following attachment types:

- ▶ FICON
- ▶ FCP
- ▶ FC_AL (including public and private loop support)
- ▶ FCIP over Gigabit Ethernet

- ▶ iSCSI over Gigabit Ethernet
- ▶ ISLs (attaching multiple switches or directors together)
- ▶ Interoperability (attachment to other vendors switches)

MDS 9000 switches and directors also support optional CWDM SFPs to provide aggregation of multiple links onto a single optical fiber through a passive optical mux.

The IP Services Modules and Multiprotocol Services modules described in 6.2.6, “Optional modules” on page 74, provide the Gigabit Ethernet interfaces, to enable iSCSI and FCIP capabilities for the 9200 and 9500 families.

FICON

IBM qualification for FICON covers the Cisco MDS 9216 Multilayer Fabric Switch, Cisco MDS 9506 Multilayer Director, and Cisco MDS 9509 Multilayer Director. These switches are fully compliant with FC-SB-2 and FC-SB-3 standards. The Cisco Mainframe Package optional license is required as discussed in 6.3.2, “Optional licensed feature packages” on page 84.

Cisco VSAN technology can help to provide greater security for a FICON traffic in an intermix environment. FICON can also be encapsulated over FCIP using the IP Services Module to provide an alternative to using channel extender devices.

The Cisco FICON environment can be managed from z/OS using the CUP function, from Cisco Fabric and Device Managers or by using the Cisco SAN-OS CLI.

6.4.2 Supported port types

The Fibre Channel ports on all models of the MDS 9000 family provide an auto-sensing 1 or 2-Gbps SFP that use LC connectors. The operating port modes supported are described in the following sections.

Auto mode

Interfaces configured in the default auto mode are allowed to operate in either the fabric port (F_Port), fabric loop port (FL_Port), expansion (E_Port), or trunking E port (TE_Port) mode. The port mode is determined during interface initialization. For example, if the interface is connected to a node, server, or disk, it operates in F_Port or FL_Port mode depending on the N_Port or NL_Port mode. If the interface is attached to a third-party switch, it operates in E_Port mode. If the interface is attached to another MDS 9000 switch, it may become operational in TE_Port mode.

TL_Ports, SD_ports and ST_ports are not automatically determined during initialization and must be administratively configured.

E_Port

In E_Port mode, an interface functions as a fabric expansion port. This port can be connected to another E_Port to create an ISL between two switches. E_Ports carry frames between switches for configuration and fabric management. They serve as a conduit between switches for frames destined for remote N_Ports and NL_Ports. E_Ports support class 2, class 3, and class F service.

An E_Port connected to another switch can also be configured to form a PortChannel.

F_Port

In F_Port mode, an interface functions as a fabric port. This port can be connected to a node (server, disk, or tape) operating as an N_Port. An F_Port can be attached to only one N_Port. F_Ports support class 2 and class 3 service.

FL_Port

In FL_Port mode, an interface functions as a fabric loop port. This port may be connected to one or more NL_Ports (including FL_Ports in other switches) to form a public arbitrated loop. If more than one FL_Port is detected on the arbitrated loop during initialization, only one FL_Port becomes operational. The other FL_Ports enter a non-participating mode. FL_Ports support class 2 and class 3 service.

Fx_Port

Interfaces configured as Fx_Ports automatically negotiate operation in either F_Port or FL_Port mode. The mode is determined during interface initialization, depending on the attached N_Port or NL_Port. This administrative configuration disallows interfaces to operate in other modes, such as preventing an interface to connect to another switch.

TL_Port

In translatable loop port (TL_Port) mode, an interface functions as a translatable loop port. It might be connected to one or more private loop devices (NL_Ports). The TL_Port mode is specific to Cisco MDS 9000 family switches and has similar properties as FL_Ports. TL_Ports enable communication between private loop devices and one of the following target devices:

- ▶ A device attached to any switch on the fabric
- ▶ A device on a public loop anywhere in the fabric

- ▶ A device on a different private loop anywhere in the fabric
- ▶ A device on the same private loop

TL_Ports support class 2 and class 3 services.

TE_Port

In trunking E_Port (TE_Port) mode, an interface functions as a trunking expansion port. It connects to another TE_Port to create an EISL between two switches. TE_Ports are specific to the Cisco MDS 9000 family. They expand the functionality of E_Ports to support these features:

- ▶ Multiple VSAN trunking
- ▶ Transport QoS parameters
- ▶ Fibre Channel trace (**fctrace**) feature

In TE_Port mode, all frames are transmitted in the EISL frame format, which contains VSAN information. Interconnected switches use the VSAN ID to multiplex traffic from one or more VSANs across the same physical link. This feature is referred to as *trunking* in the Cisco MDS 9000 Family.

TE_Ports support class 2, class 3, and class F service.

SD_Port

In switch port analyzer (SPAN) destination port (SD_Port) mode, an interface functions as a SPAN. The SPAN feature is specific to switches in the Cisco MDS 9000 family. It monitors network traffic passing through a Fibre Channel interface. This monitoring is done using a standard Fibre Channel analyzer, or similar switch probe, that is attached to an SD_Port. SD_Ports do not receive frames. They merely transmit a copy of the source traffic. The SPAN feature is nonintrusive and does not affect switching of network traffic for any SPAN source ports.

ST_Port

Interfaces configured as ST ports serve as an entry point port in the source switch for a Fibre Channel tunnel. ST ports are specific to remote SPAN (RSPAN) ports and cannot be used for normal Fibre Channel traffic.

Figure 6-16 shows an example of the port types that are available with the Cisco MDS 9000 family of products.

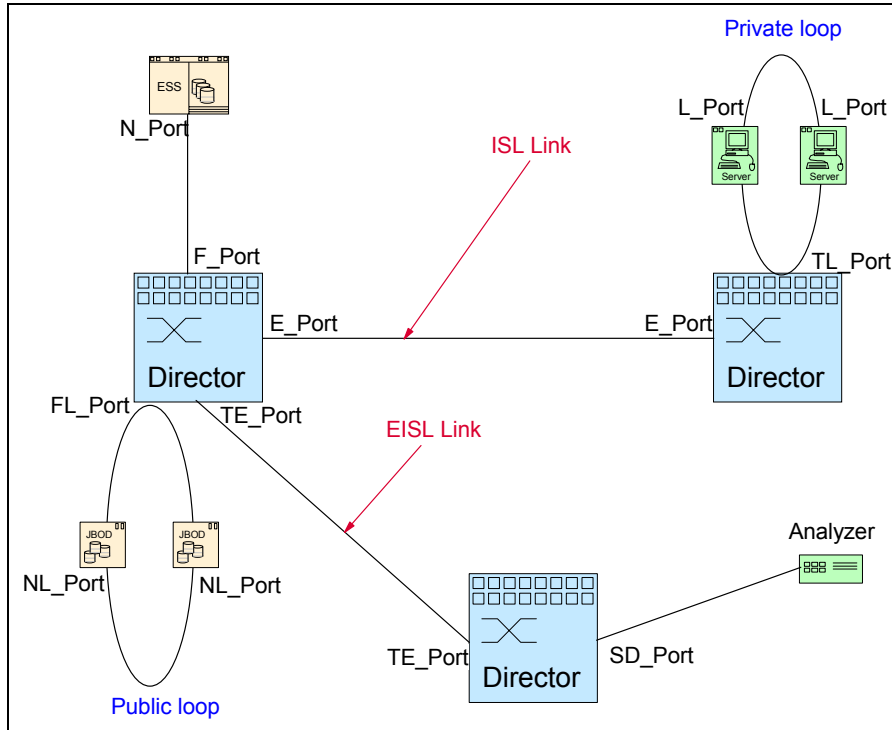


Figure 6-16 Cisco MDS 9000 family port types

6.4.3 VSAN

VSAN technology allows virtual fabrics (enabled by a mixture of ASIC functionality and software functionality) to be overlaid on a physical fabric. Cisco's approach is to position VSAN not as a single feature, but as an architectural approach that allows flexible delivery of many features.

A given device can belong only to one VSAN. Each VSAN contains its own zoning, fabric services, and management capabilities, as though the VSAN were configured as a separate physical fabric.

VSANs offer the following features:

- ▶ Ease of configuration is enhanced because devices can be added or removed from a VSAN fabric without making any physical changes to cabling.
- ▶ Traffic is isolated to within a VSAN, unless IVR is implemented. Separate companies or divisions of a company can be segregated from each other without needing separate physical fabrics.

- ▶ Fabric services are provided separately to each VSAN. Smaller fabrics are simpler and generate fewer Registered State Change Notifications (RSCNs) between switches. Each VSAN runs all required protocols such as Fabric Shortest Path First (FSPF), domain manager, and zoning.
- ▶ Redundancy can be configured, for example, on a dual HBA server by having each HBA in a separate VSAN. This is the same as you would typically have each HBA in a separate physical fabric if you did not have VSANs.
- ▶ Duplicate FCIDs can be accommodated on a network provided the devices are in separate VSANs. This allows for IVR connection of fabrics that were previously completely separate.

Figure 6-17 represents a typical SAN environment that has a number of servers, each with multiple paths to the SAN. In this case, the SAN consists of a Fibre Channel director attached to a disk and tape subsystem.

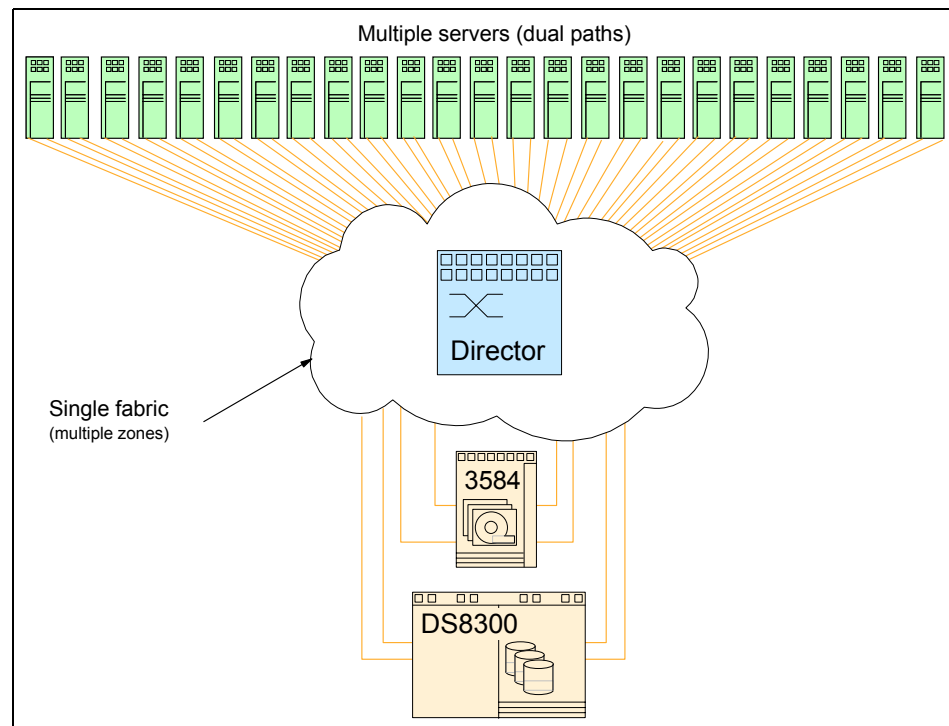


Figure 6-17 Traditional SAN

Figure 6-18 shows how the same scenario is implemented using Cisco's VSAN.

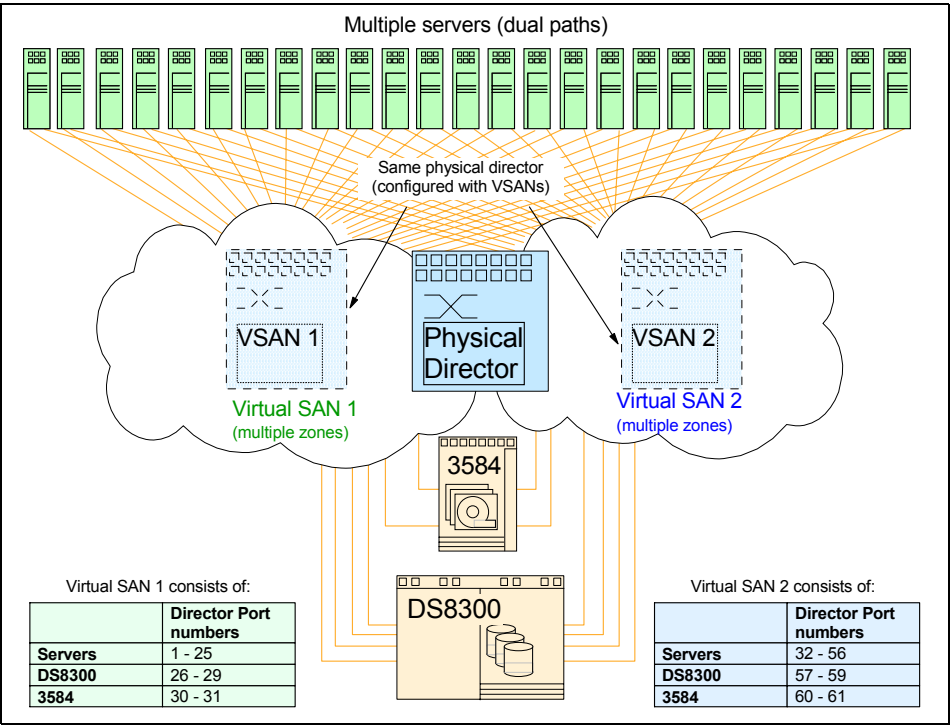


Figure 6-18 Virtual SAN

In this example, the servers are still connected to the SAN, but the SAN consists of a single 9509 attached to the same disk and tape subsystems. In this case, we configured the first 31 ports in the director into a VSAN called Virtual SAN 1, and the second 31 ports into another VSAN called Virtual SAN 2. The servers have a connection to each VSAN, thereby providing a solution that consists of multiple SAN fabrics.

The VSANs cannot communicate with each other, unless IVR is implemented. They appear to be totally separate fabrics. They have their own FSPF tables, domain manager, and zoning requirements. Any traffic disruption in one VSAN does not impact the other VSAN. A port cannot belong in multiple VSANs.

Note: A new feature in SAN-OS 2.x is support for WWN-based VSANs.

VSANs compared to zones

Table 6-1 shows the main differences between a zone and a VSAN.

Table 6-1 VSANs compared to zones

VSANs	Zones
A VSAN is a logical fabric with its own routing, naming, and zoning protocols.	A zone is a logical group of ports or WWNs which are allowed to talk to each other.
VSANs can contain multiple zones.	Zones are always contained within a VSAN. They cannot span a VSAN.
VSANs limit the reach of fabric services transmissions.	Zones limit the reach of I/O transmissions.
Membership is defined using a VSAN ID to Fx ports. As of SAN-OS 2.x, membership can be defined using WWN.	Membership is typically defined using WWN or port number (Fx).
HBAs may belong only to a single VSAN, the VSAN associated with the Fx port.	HBAs can belong in multiple zones.
VSANs enforce membership at each E_Port, source port, and destination port.	Zones enforce membership only at the source and destination ports.

Registered State Change Notifications

The Registered State Change Notification service propagates information about a change in state of one node to all other nodes in the fabric. In the event of a device shutting down, for example, the other devices on the SAN are informed and then know not to send data to the shutdown device, avoiding time-outs and retries.

There are two types of RSCNs. Switch RSCNs (SW_RSCN) are passed from one switch to another, for example when a new device comes online, and the local switch needs to inform the other switches. SW_RSCNs are sent at a VSAN level (within a VSAN). The second type of RSCN is issued by the switch to an end device that informs it of a change within a zone to which the end device belongs. This type of RSCN is sent to only those devices in the affected zone.

Default and isolated VSANs

Up to 1024 VSANs can be configured on a physical SAN. Of these, one is the default VSAN (VSAN 1) and another is an isolated VSAN (VSAN 4094). User-specified VSAN IDs range from 2 to 4093.

Default VSAN

The factory settings for switches in the Cisco MDS 9000 family have only default VSAN 1 enabled. If you do not need more than one VSAN for a switch, use this default VSAN as the implicit parameter during configuration. If no VSANs are configured, all devices in the fabric are considered part of the default VSAN. By default, all ports are assigned to the default VSAN.

Isolated VSANs

VSAN 4094 is an isolated VSAN. All non-trunking ports are transferred to this VSAN when the VSAN to which they belong is deleted. This avoids an implicit transfer of ports to the default VSAN or to another configured VSAN. All ports in the deleted VSAN are disabled.

VSAN membership

Port VSAN membership on the switch is assigned on a port-by-port basis. By default, each port belongs to the default VSAN. Trunking ports have an associated list of VSANs that are part of an allowed list.

VSAN attributes

VSANs have the following attributes:

- ▶ The *VSAN ID* identifies the VSAN as the default VSAN (VSAN 1), user-defined VSANs (VSAN 2 to 4093), and the isolated VSAN (VSAN 4094).
- ▶ The *administrative state* of a VSAN can be configured to an active (default) or suspended state. When VSANs are created, they can exist in various conditions or states.
 - The *active state* of a VSAN indicates that the VSAN is configured and enabled. By enabling a VSAN, you activate the services for that VSAN.
 - The *suspended state* of a VSAN indicates that the VSAN is configured but not enabled. If a port is configured in this VSAN, it is disabled. Use this state to deactivate a VSAN without losing the VSAN's configuration. All ports in a suspended VSAN are disabled. By suspending a VSAN, you can preconfigure all the VSAN parameters for the whole fabric and activate the VSAN immediately.
- ▶ The *VSAN name* text string identifies the VSAN for management purposes. The name can be from 1 to 32 characters long and it must be unique across all VSANs. By default, the VSAN name is a concatenation of VSAN and a four-digit string representing the VSAN ID. For example, the default name for VSAN 3 is VSAN0003.
- ▶ *Load balancing* attributes indicate the use of the source-destination ID (src-dst-id) or the originator exchange OXID (src-dst-ox-id, the default) for load balancing path selection.

Operational state of a VSAN

A VSAN is in the operational state if the VSAN is active and at least one port is up. This state indicates that traffic can pass through this VSAN. This state cannot be configured.

Deleted VSAN

When an active VSAN is deleted, all of its attributes are removed from the running configuration.

VSAN-related information is maintained by the system software:

- ▶ VSAN attributes and port membership details are maintained by VSAN manager. This feature is affected when you delete a VSAN from the configuration. When a VSAN is deleted, all the ports in that VSAN are made inactive and the ports are moved to the isolated VSAN. If the same VSAN is recreated, the ports are not automatically assigned to that VSAN. You must *explicitly* reconfigure the port VSAN membership.
- ▶ VSAN-based runtime (name server), zoning, and configuration (static route) information is removed when the VSAN is deleted.
- ▶ Configured VSAN interface information is removed when the VSAN is deleted.

6.4.4 Inter-VSAN Routing

IVR is available when the Enterprise Package license has been applied to a switch running v1.3 (2a) SAN-OS or later. IVR helps to allow data traffic to flow between VSANs while maintaining the VSAN segregation because no management data is passed. This proves useful, for example, when a host defined in one VSAN is required to have access to a tape drive defined in another VSAN. This feature reduces the amount of required hardware to meet the needs for multiple systems.

An IVR is defined in a similar manner to normal zoning within a VSAN. Instead of working within a VSAN and performing the zoning definitions, we work from the IVR group to create an IVR zone set, which can be activated or deactivated without affecting the VSANs.

Figure 6-19 shows how the same scenario is implemented using Cisco's IVR.

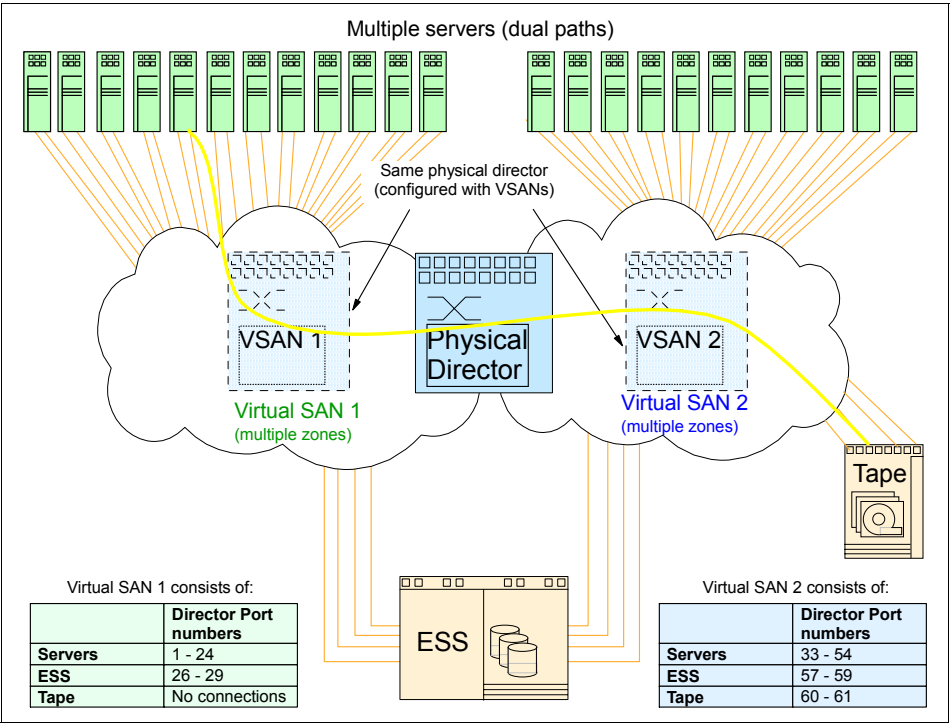


Figure 6-19 Inter-VSAN Routing

In this example, two groups of servers are connected to different VSANs within a single MDS 9509 Director. The disk has paths that are defined to both VSAN 1 and VSAN 2, although the requirement is for all servers to access all tape drives. In this case, we configured the first 29 ports in the director into VSAN 1 and the second 31 ports into VSAN 2.

The VSANs cannot communicate with each other, and they appear to be totally separate SANs. By defining an IVR zone, we can allow a data connection from a server in VSAN 1 through to a tape drive in VSAN 2. No management data is passed over this connection, and any disruptions in one VSAN do not have any impact on the other VSAN.

Tip: Use VSANs on an exception basis. For example, use them for multivendor switch interoperability, to isolate separate companies in a shared services environment, to manage QoS between test and production environments (using the new VSAN-based QoS feature), and to isolate less reliable FCIP links from disrupting the main fabric.

If you have a lot of IVRs in your design, you may have too many VSANs. If you consider a LAN analogy, you would not install several routers into the middle of your corporate LAN.

6.4.5 PortChanneling

PortChanneling is Cisco's term for exchange-based load balancing across multiple ISLs. An exchange is usually a single SCSI command and the response it evokes, so it is of fairly short duration (milliseconds or seconds). However, exchanges can be longer in a FICON environment since FICON improves efficiency by retaining the exchange-id for multiple commands.

PortChanneling can also be implemented based on source ID (that is, a server HBA port) and destination ID (for example, a disk system HBA port) pairs, which give less granular load balancing but provide some traffic isolation if that is preferred.

PortChanneling does not do load balancing at a frame level. Frame-based load balancing requires additional out-of-order frame management intelligence.

With PortChannels, users can aggregate up to 16 physical ISLs into a single load-balanced bundle. The group of Fibre Channel ISLs designated to act as a PortChannel can consist of any port on any 16-port switching module within the MDS 9000 chassis, allowing the overall PortChannel to remain active upon failure of one or more ports, or failure of one or more switching modules. These PortChannels support the following functions:

- ▶ Increase the aggregate bandwidth on an ISL or EISL by distributing traffic among all functional links in the channel
- ▶ Load balance across multiple links and maintains optimum bandwidth utilization

Load balancing is based on a source ID (SID), destination ID (DID), and, optionally, the originator exchange ID (OXID) that identify the flow of the frame.

- Provide high availability on an ISL

If one link fails, traffic previously carried on this link is switched to the remaining links. If a link goes down in a PortChannel, the upper protocol is not aware of it. To the upper protocol, the link is still there, although the bandwidth is diminished. The routing tables are not affected by a link failure.

Figure 6-20 shows ISLs and PortChanneling.

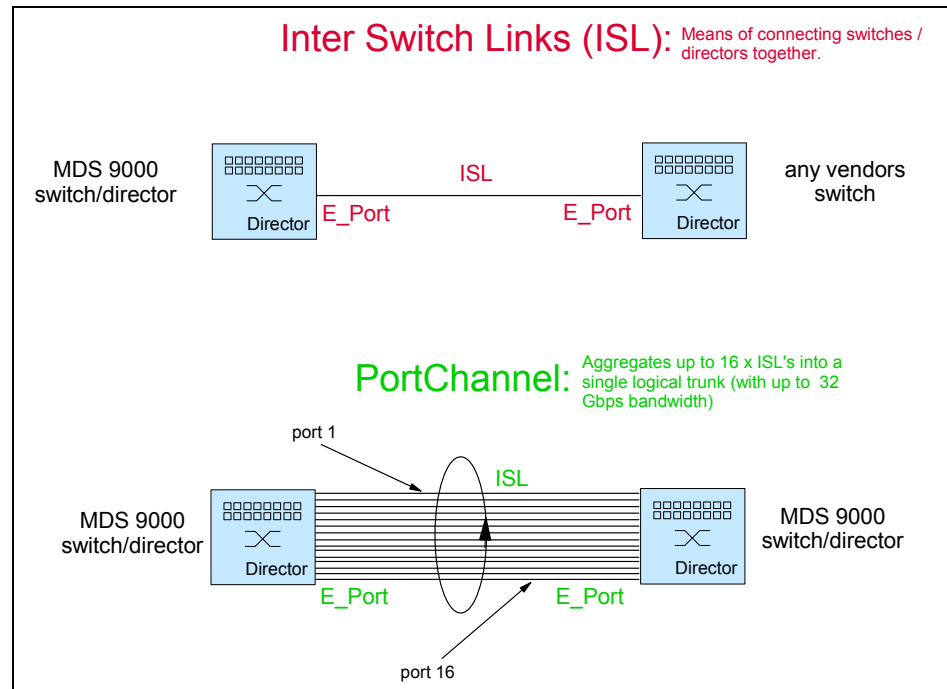


Figure 6-20 PortChannels and ISLs on the Cisco MDS 9000 switches

6.4.6 Trunking

The Cisco MDS 9000 family uses the term *trunking* to refer to an ISL that carries one or more VSANs. Trunking ports receive and transmit EISL frames. EISL frames carry an EISL header containing the VSAN information. When EISL is enabled on an E_Port, that port becomes a TE_Port.

Trunking is also referred to as *VSAN trunking*, because it applies only to a VSAN. If a trunking enabled E_Port is connected to another vendor's switch, the trunking protocol ensures that the port will operate as a standard E_Port.

Figure 6-21 shows a diagram of trunking. It also demonstrates how you can use a combination of PortChannels and trunking to create an aggregate bandwidth of up to 32 Gbps between switches.

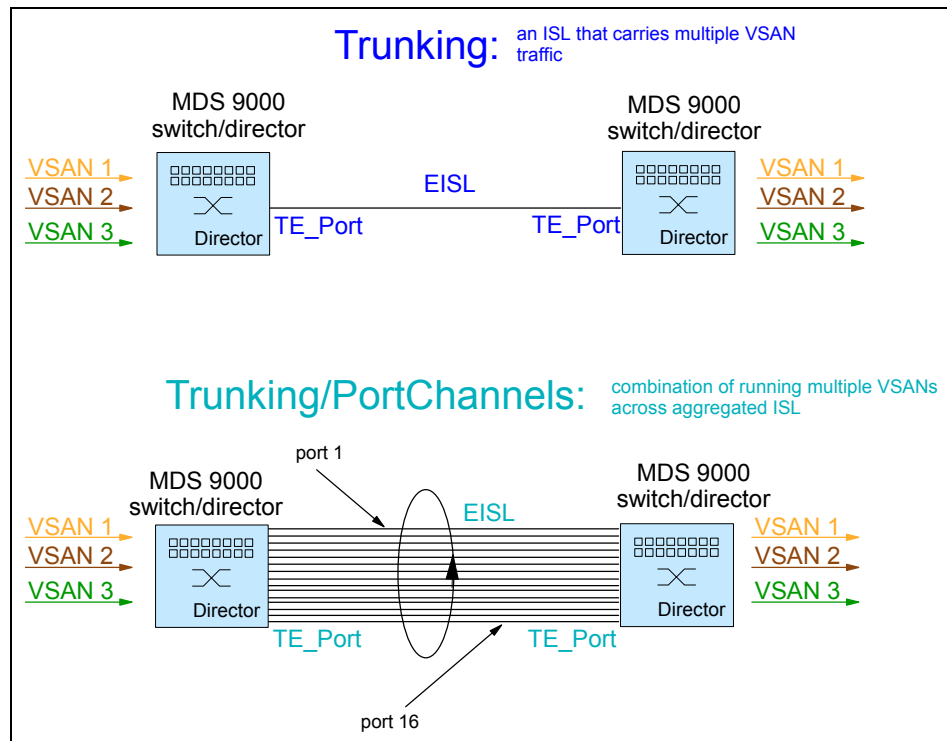


Figure 6-21 Trunking and PortChanneling

6.4.7 Quality of Service

QoS is generally managed in a Fibre Channel environment by reducing the number of receiver ready (R_RDY) buffer credits to a sender. Four distinct QoS priority levels are available: three for Fibre Channel data traffic and one for Fibre Channel control traffic. Fibre Channel data traffic for latency-sensitive applications can be configured to receive higher priority than throughput-intensive applications using data QoS priority levels. Control traffic is assigned the highest QoS priority automatically, to accelerate convergence of fabric-wide protocols such as FSPF, zone merges, and principal switch selection.

Data traffic can be classified for QoS by the VSAN identifier, zones, N-Port WWN, or FCID. Zone-based QoS helps simplify configuration and administration by using the familiar zoning concept.

Note: Zone-based QoS (introduced in SAN-OS 2.1) also requires the SAN-OS Enterprise Package.

QoS offers the following primary advantages:

- ▶ Provides latency to reduce frame loss in congested networks
- ▶ Prioritizes transactional traffic over bulk traffic

The Cisco MDS 9000 family supports QoS for internally and externally generated control traffic. Within a switch, the control traffic is sourced to the supervisor module and is treated as a high priority frame. A high-priority status provides absolute priority over all other traffic and is assigned in the following cases:

- ▶ Internally-generated, time-critical control traffic (generally Class F frames)
- ▶ Externally-generated, time-critical control traffic entering a switch in the MDS 9000 range from another vendor's switch

High priority frames originating from other vendor switches retain the priority as they enter a switch in the MDS 9000 family.

By default, the QoS feature for control traffic is enabled but can be disabled if required.

6.4.8 Fibre Channel Congestion Control

Fibre Channel Congestion Control (FCC) is a Cisco proprietary flow control mechanism that alleviates congestion on Fibre Channel networks. A switch experiencing congestion signals this condition to the upstream (source) switch, which throttles the traffic by reducing the buffer-to-buffer credits.

By default, the FCC protocol is disabled. You can enable the protocol globally for all the VSANs configured in the switch, or selectively enable or disable it for each VSAN.

Congestion control methods

With FCC enabled, there are different congestion control methods:

- ▶ *Path quench control* reduces severe congestion temporarily by slowing the source to the whole path in the fabric.
- ▶ *Edge quench control* provides feedback to the source about the rate at which frames should be entered into the network (frame intervals).

FCC process

When a node in the network detects congestion for an output port, it generates an edge or a path quench message. These frames are identified by the Fibre Channel destination ID and the source ID.

Any receiving switch in the Cisco MDS 9000 family handles frames in one of the following ways:

- ▶ It forwards the frame.
- ▶ It limits the rate of the frame flow in the congested port.

Behavior of the flow control mechanism differs, based on the Fibre Channel DID:

- ▶ If the Fibre Channel DID is directly connected to one of the switch ports, the input rate limit is applied to that port.
- ▶ If the destination of the edge quest frame is a Cisco domain or the next hop is a Cisco MDS 9000 family switch, the frame is forwarded.
- ▶ If neither of these conditions is true, then the frame is processed in the port going toward the Fibre Channel DID.

All switches, including the edge switch, along the congested path process path quench frames. However, only the edge switch processes edge quench frames. The FCC protocol is implemented for each VSAN and can be enabled or disabled on a specified VSAN or for all VSANs at the same time.

Note: Cisco's FCC differs from standard buffer-to-buffer flow control. FCC looks at the source of congestion and passes messages upstream to report it to the nearest switch to the source so it can apply selective buffer-to-buffer quenching as appropriate. Standard buffer-to-buffer flow control simply delivers point-to-point flow control.

If you enable FCC on one switch, be sure to enable it on all switches in the fabric.

6.4.9 Switch port analyzer

The Cisco MDS 9000 family provides a feature called the *switch port analyzer*. As mentioned in 6.4.2, "Supported port types" on page 88, the SPAN or SD_Ports allow you to monitor network traffic through the Fibre Channel interface.

Traffic through any Fibre Channel interface can be replicated to a special port called the *SPAN destination port*. Any Fibre Channel port in a switch can be configured as an SD_Port. When an interface is in SD_Port mode, it cannot be

used for normal data traffic. You can attach a Fibre Channel analyzer to the SD_Port to monitor SPAN traffic.

SD_Ports do not receive frames. They only transmit a copy of the SPAN source traffic. The SPAN feature is non-intrusive and does not affect switching of network traffic for any SPAN source port.

SPAN sources

A SPAN source is the interface from which traffic can be monitored. You can also specify a VSAN as a SPAN source, in which case, all supported interfaces in the specified VSAN are included as SPAN sources. You can choose the SPAN traffic in the ingress direction, the egress direction, or both directions, for any source interface.

► Ingress source (rx)

Traffic entering the switch fabric through this source is spanned or copied to the SD_Port, as shown in Figure 6-22.

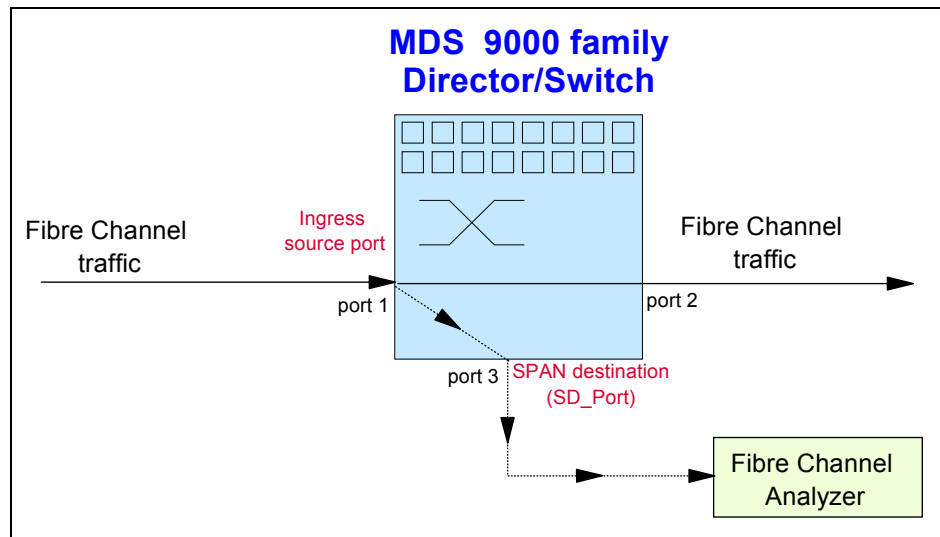


Figure 6-22 SD_Port for ingress (incoming) traffic

► Egress source (tx)

Traffic exiting the switch fabric through this source interface is spanned or copied to the SD_Port, as shown in Figure 6-23.

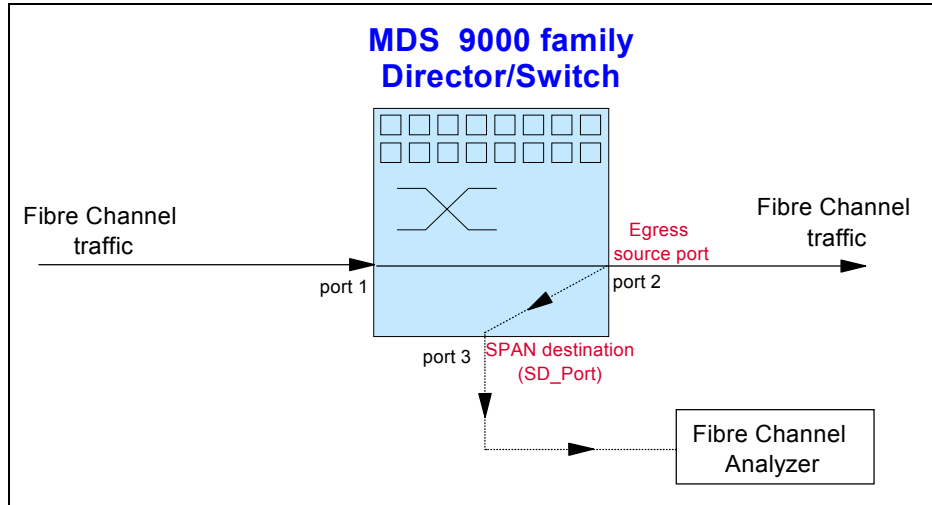


Figure 6-23 SD_Port for egress (outgoing) traffic

Allowed source interface types

The SPAN feature is available for the following interface types:

- ▶ Physical ports: F_Ports, FL_Ports, TE_Ports, E_Ports, and TL_Ports
- ▶ Interface sup-fc0 (traffic to and from the supervisor)
 - Fibre Channel traffic from the supervisor module to the switch fabric, through the sup-fc0 interface, is called *ingress traffic*. It is spanned when sup-fc0 is chosen as an ingress source port.
 - Fibre Channel traffic from the switch fabric to the supervisor module, through the sup-fc0 interface, is called *egress traffic*. It is spanned when sup-fc0 is chosen as an egress source port.
- ▶ PortChannels
 - All ports in the PortChannel are included and spanned as sources.
 - You cannot specify individual ports in a PortChannel as SPAN sources. Previously-configured SPAN-specific interface information is discarded.

VSAN as a SPAN source

When a VSAN as a source is specified, then all physical ports and PortChannels in that VSAN are included as SPAN sources. A TE_Port is included only when the port VSAN of the TE_Port matches the source VSAN. A TE_Port is excluded even if the configured allowed VSAN list might have the source VSAN, but the port VSAN is different.

Guidelines for configuring VSANs as a source

The following guidelines apply when configuring VSANs as a source:

- ▶ Traffic on all interfaces included in a source VSAN is spanned only in the ingress direction.
- ▶ When a VSAN is specified as a source, you will not be able to perform interface-level configuration on the interfaces that are included in the VSAN. Previously-configured SPAN-specific interface information is discarded.
- ▶ If an interface in a VSAN is configured as a SPAN source, you will not be able to configure that VSAN as a source. You must first remove the existing SPAN configurations on such interfaces before configuring VSAN as a source.
- ▶ Interfaces are only included as sources when the port VSAN matches the source VSAN.

6.5 Interoperability

Cisco's offering on interoperability includes support for Inter-VSAN routing to Brocade, McDATA, and QLogic switch products via an RPQ.

6.5.1 Switch interoperability modes

Switch interoperability modes enable other vendors' switches to connect to the MDS family. These modes are required because vendors often implement features on their switches that are not compatible with other manufacturers. Table 6-2 shows the modes.

Table 6-2 Cisco interoperability modes

Mode	Will interoperate with
Default	Cisco, QLogic
Interop 1	McDATA open mode (default for McDATA) Brocade interop mode
Interop 2	Brocade proprietary <17 port switches Core PID mode 0
Interop 3	Brocade proprietary >16 port switches Core PID mode 1, Core PID mode 3 (mode 0 emulation for switches that can't run mode 0 natively)

Note: VSANs are a valuable tool in managing interoperability between multivendor switches. Heterogeneous IVR was introduced in SAN-OS 2.1

The switch interoperability mode may disable a number of advanced or proprietary features, so it is worth understanding what these might affect before proceeding. Table 6-3 lists the changes required if interoperability mode is enabled on a Cisco MDS 9000 family switch or director.

Table 6-3 Interoperability mode changes

Switch feature	Changes if Interoperability Mode is enabled
Domain IDs	While Cisco implement the full standard specification 239 domain IDs (switch IDs), McDATA supports only 31 domain IDs within a fabric when using midrange switches. Therefore, Interop domain IDs are restricted to the range 97 to 127 to accommodate all vendors implementations.
Timers	All Fibre Channel timers must be set to the same value on all switches because these values are exchanged by E_Ports when establishing an ISL. The Time-Out Value timers are described in the following rows. (This is always required for connection of two switches)
F_S_TOV	Verify that the Fabric Stability Time-Out Value timers match exactly.
D_S_TOV	Verify that the Distributed Services Time-Out Value timers match exactly.
E_D_TOV	Verify that the Error Detect Time-Out Value timers match exactly.
R_A_TOV	Verify that the Resource Allocation Time-Out Value timers match exactly.
Trunking	Trunking is not supported between two different vendors' switches. This feature may be disabled on a per port basis.
Default Zone	The default zone behavior of permit (all nodes can see other nodes) or deny (all nodes are isolated when not explicitly placed in a zone) might change.
Zoning attributes	Zones can be limited to the WWPN, and other proprietary zoning methods (physical port number) can be eliminated.
Zone propagation	Some vendors do not pass the full zone configuration to other switches, only the active zoneset gets passed.
VSAN	This only affects the specified VSAN.

Switch feature	Changes if Interoperability Mode is enabled
TE_Ports and PortChannels	TE_Ports and PortChannels only apply when connecting from one MDS 9000 to another. Only E_Ports can be used to connect to non-MDS switches. TE_Ports and PortChannels can be used to connect to other MDS 9000 switches when interoperability mode is enabled.
Domain reconfiguration disruptive	This can require the entire switch to be restarted when changing the domain IDs.
Domain configuration nondisruptive	This only impacts the affected VSAN. Only the domain manager for the affected VSAN is restarted. Other VSANs are unaffected.
Name Server	Need to verify that all vendors have the correct values in their respective Name Server tables.

Interoperability mode in the Cisco MDS 9000 family can be enabled nondisruptively, but the default is to have this mode turned off.

It is still important to check with the OEM vendors involved in regard to the specific steps that must be taken.

6.5.2 Interoperability matrix

You can download the latest IBM interoperability matrixes for Cisco MDS switches from:

<ftp://service.boulder.ibm.com/storage/san/cisco/>

The interoperability matrixes include a list of servers, disk, and tape systems that have been tested and are supported with the MDS family of switches and directors. These lists also contain supported operating system versions and links to other Web sites that document the required HBA levels.

For combinations of technologies that are not explicitly supported, contact your local IBM office or IBM Business Partner to discuss submitting an RPQ.



Cisco solutions

Routers provide access to data which is located in a different fabric. The principal uses for storage routing are:

- ▶ Storage area network (SAN) extension over Internet Protocol (IP) networks, not the 9120 and 9140
- ▶ Lowering connection costs using Small Computer System Interface over IP (iSCSI), not the 9120 and 9140
- ▶ Achieving isolation and interoperability between different business units
- ▶ Managing scalability as your SAN environment grows
- ▶ Migrating from an older storage environment to a newer one

The latter three features are inherent to the entire Cisco MDS 9000 product family. The former two uses apply to all members of the family except for the MDS 9120 and the MDS 9140 because they do not have Ethernet ports.

7.1 SAN extension with FCIP

Fibre Channel (FC) distances have traditionally been limited to either local fiber runs using 9 micron longwave Fibre Channel, or high quality wide area networks (WANs), such as SONET and SDH, in combination with coarse wavelength division multiplexing (CWDM) or dense wavelength division multiplexing (DWDM) multiplexers.

The advent of Fibre Channel over IP (FCIP) has meant that applications that can tolerate the higher latencies of IP networks can now make Fibre Channel connections across standard corporate IP WANs. The advantages of this are that it uses a widely available and well understood infrastructure. This translates into lower cost.

We are still in a phase where people want FCIP over standard networks to be a panacea for all SAN extension applications. The inherent latencies involved are around five microseconds per kilometer (km) travelled in each direction and added latencies at every step (for example, up to 100 microseconds per router or firewall). This prevents FCIP from being used effectively for applications such as long distance synchronous replication or online transaction processing (OLTP). For example, a high quality network of 1000 km might have a latency of around 20 milliseconds. Given that a disk input/output (I/O) might only take 10 milliseconds, the problem with a 20 millisecond latency becomes obvious.

Because some corporate WANs provide uncertain quality of service, storage router vendors tend to be cautious about quoting distances for FCIP, and generally recommend that high quality WANs are necessary to provide services over anything more than 200 or 300 km.

In practice, the most common uses for FCIP are long-haul asynchronous replication, short-haul synchronous replication, and logical unit number (LUN) access over campus or metro distances. A client may choose to implement Fibre Channel tunneled into IP on a campus scale simply because the IP links are already in place. On a short IP network, the main problem becomes Quality of Service (QoS) since the latencies are not so large. The principles are the same whether running over 500 meters (m) or 5000 km. All that varies is the link latency and the service reliability and consistency.

7.1.1 Compression

FCIP compression in the Cisco MDS 9000 Family SAN-OS increases the effective WAN bandwidth. While Gigabit Ethernet ports for IP Storage Services can theoretically achieve up to a thirty to one (30:1) compression ratio, typical ratios in the field are less than two to one (2:1).

Hardware-based compression is also available on the 14+2 linecard in addition to the MDS 9216i integrated 14+2 controller.

When software compression is turned on with the IP Services (IPS) line card and set to modes 2 or 3, then the IPS runs small Fibre Channel frames together to use up space inside the IP packets rather than sending individual frames separately. This is only for use over WAN distances and is not recommended for cross-campus use.

Using jumbo packets can also improve throughput. Remember that jumbo packets need to be turned on through the entire data path.

Note: Cisco's SAN Extension Tuner helps to optimize FCIP performance. The tuner generates SCSI I/O commands that are directed to a specific virtual target. It reports I/Os per second and I/O latency results. SAN Extension Tuner is included with the FCIP enablement license package.

7.1.2 Using Inter-VSAN Routing with FCIP

The stability of WAN links varies by geography and provider. It is usually important to separate FCIP links running over WANs from your core Fibre Channel network. Because we want to protect both ends of the core fabric from FCIP link bounce (when the network may go up and down a few times in quick succession), we create transit virtual SANs (VSANs) between the two switches and then implement Inter-VSAN Routing (IVR). Transit VSANs consist solely of the FCIP ports on the switches.

Figure 7-1 shows an example of an asynchronous replication running over IP at a 500 km distance.

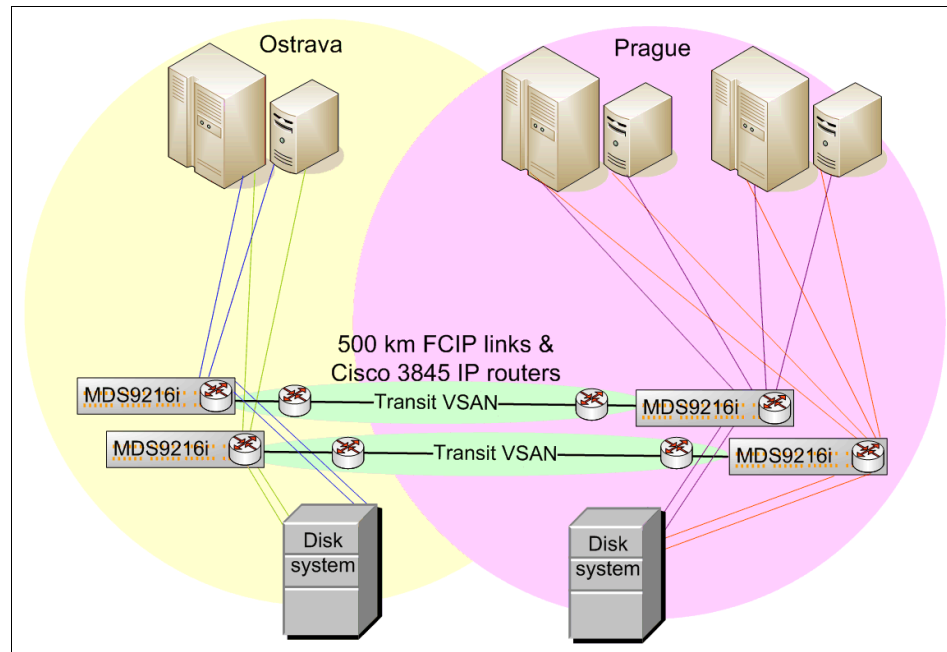


Figure 7-1 Tunneling FCIP using transit VSANs to mitigate link bounce

7.1.3 Using FCIP Write Acceleration

FCIP Write Acceleration (FCIP-WA) is an attempt to avoid the transport latency associated with long distance SCSI I/O operations. The IPS module allows the initiator to send the data to be written before the write command is processed by the remote target and a SCSI Transfer Ready message has time to travel back to start the data transfer in the traditional way. Essentially FCIP-WA spoofs the R_RDY.

FCIP-WA performance depends on the traffic profile of the SCSI operations being performed. The following I/O characteristics can benefit from FCIP-WA:

- ▶ Long distance high latency network
- ▶ Write intensive I/O
- ▶ High number of small SCSI writes (rather than low number of large writes)
- ▶ Disk system has low write latency

This suggests that the best use for FCIP-WA is in disk system replication, but results in practice are not always worth the added complexity. Across a 100 km link, replication using FCIP-WA can be expected to deliver around 10% improvement in throughput and a similar percentage reduction in latency on a given FCIP network.

Note: FCIP-WA is different from Fibre Channel Write Acceleration (FC-WA), which applies to FC-FC links without FCIP. FC-WA requires the Storage Services Module, but FCIP-WA does not.

7.1.4 Using Fibre Channel tape acceleration with FCIP

Fibre Channel tape acceleration is similar to FC-WA. However, where FC-WA assumes R_RDY but does not assume data is received until it receives acknowledgment from the target, tape acceleration goes one step further. Tape acceleration does not wait for R_RDY nor for acknowledgement that data has been received before sending in the next packet of data. Tape acceleration leaves the target tape in an uncertain state in the event of a link failure.

If you are using a sophisticated backup tool, such as Tivoli Storage Manager, then Tivoli Storage Manager can restart a migration from a disk storage pool to the tape storage pool from the point of failure. But if you use FC-TA, then Tivoli Storage Manager may think that a write has been completed when it has not. Then any restart would be fatally flawed.

Note: Tape acceleration spoofs both the R_RDY and the ACK, but it does not spoof the final tape mark.

7.2 Low-cost connection with iSCSI

You can create low-cost connections to disk storage using one of three ways.

- Fibre Channel Arbitrated Loop (FC-AL)

Using FC-AL does not require a switch port for each server, because up to 126 devices may share a single port. One Fibre Channel host bus adapter (HBA) is still required for each server.

- Network-attached storage (NAS) gateway

Using an NAS gateway, you need only to provision Fibre Channel ports for the gateway device, rather than for each server. Also no Fibre Channel HBAs are required for the servers, so the primary costs are for the gateway itself and for upgrading your Ethernet network to handle the increased traffic, as well as for establishing a VLAN for this new traffic.

Note: Some block I/O applications cannot be accessed effectively through an NAS gateway.

► iSCSI

iSCSI is like IP SAN. Using iSCSI, you do not need to provision Fibre Channel ports for each server. Also, no Fibre Channel HBAs are required, but iSCSI imposes processing overhead on each server. Therefore, in some cases, you may need a high performance Ethernet card with a TCP/IP offload engine (TOE) function. Again look at the costs associated with upgrading your Ethernet network, such as setting up a VLAN. Because iSCSI delivers block I/O, it is likely to be compatible with all applications.

All Cisco MDS multilayer switches have the capability for iSCSI, except for the MDS 9020, MDS 9120, and the MDS 9140. IP ports are available in the MDS 9216i, the 14+2 Multiprotocol Services (MPS) line card, and the 4 port and 8 port IPS line cards.

Note: The two Ethernet ports on the MDS 9216i and 14+2 MPS line card *cannot* be combined into a single EtherChannel. Two Ethernet ports on the IPS modules *can* be combined into a single EtherChannel, but only between ports that share the same application-specific integrated circuit (ASIC). PortChannel can be used.

Figure 7-2 shows how you can use iSCSI to provision disk storage to noncritical servers. You can also use iSCSI for critical servers. In general, you can expect lower performance and lower reliability on an Ethernet network than on a Fibre Channel network. When using iSCSI for critical servers, use iSCSI multipathing.

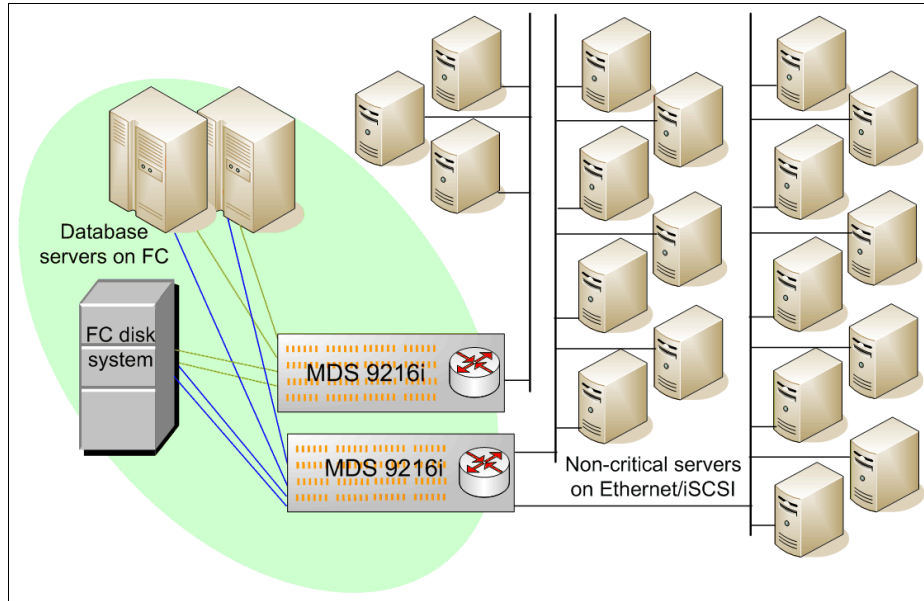


Figure 7-2 Using iSCSI routing to provision disk storage to non-critical servers

7.2.1 iSCSI Immediate Data

Cisco MDS 9216i and the IPS and MPS modules support iSCSI Immediate Data. iSCSI Immediate Data is similar to FC-WA. It spoofs the R-RDY and can send the initial payload with the initial write request.

7.3 Isolation and interoperability using IVR

Cisco's VSAN technology allows for the creation of separate logical fabrics on a shared physical infrastructure. There are cases where you need that isolation to be complete, but there are many other situations where you need to provide some access to that data. The following sections provide examples of ways in which you might use IVR.

7.3.1 Separating production from development

In addition to your main production environment, you may have a development or test environment which is subject to frequent reconfiguration and rebooting, or may be subject to a higher risk of failure due to less rigorous change controls. You need to isolate this from your production systems, but test systems also need occasional access to data that is stored on the production disk systems.

Figure 7-3 shows how you can achieve this fabric isolation by using a VSAN. Every Cisco MDS switch is also a router, so it can perform IVR.

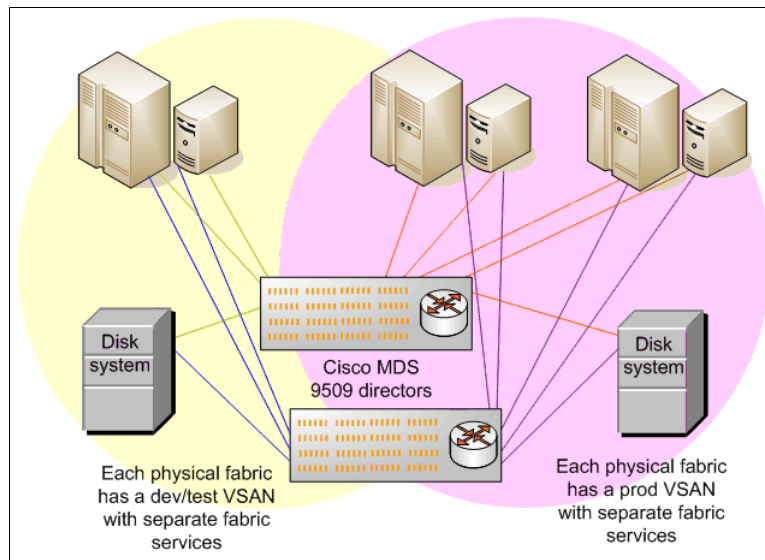


Figure 7-3 Every MDS switch is also a router: Dual director example

Figure 7-4 shows the same concept using separate switches for the two environments. Logically this is identical to Figure 7-3 but illustrates how you can route between two existing pairs of fabrics, so it also shows an example of scalability.

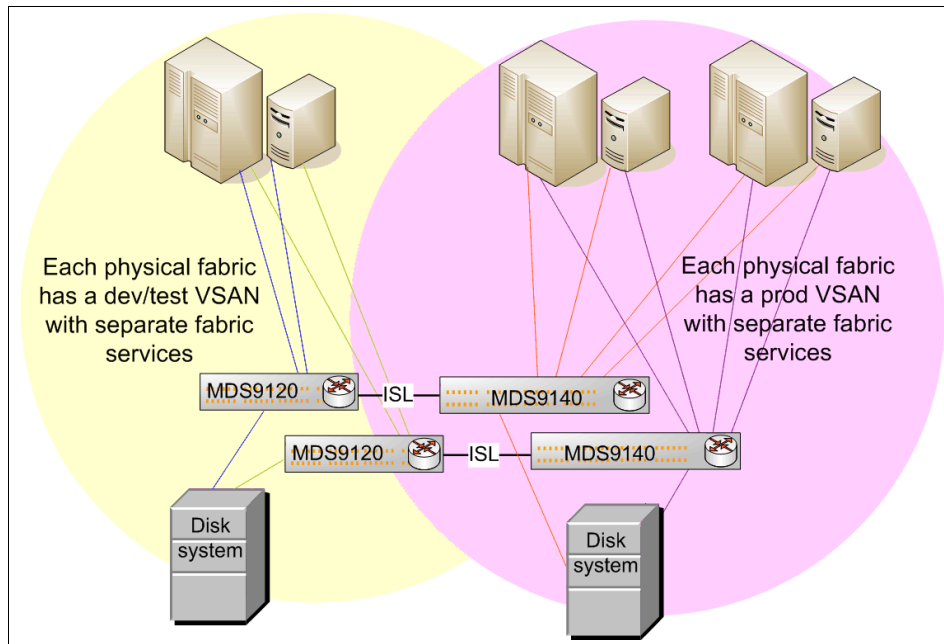


Figure 7-4 Every MDS switch is also a router: Four switch example

7.3.2 Separating corporate subsidiaries

A corporation can also choose to isolate subsidiary companies from each other while providing some shared services such as centralized backup. With Cisco MDS 9000, each VSAN can have a separate administrator with privileges granted only for that VSAN. This approach can also be used by a shared-services provider to host multiple customers on the same physical infrastructure.

Figure 7-5 shows an example where separate subsidiaries share a physical infrastructure, but live on isolated VSANs. IVR has been implemented to allow shared access to the backup infrastructure.

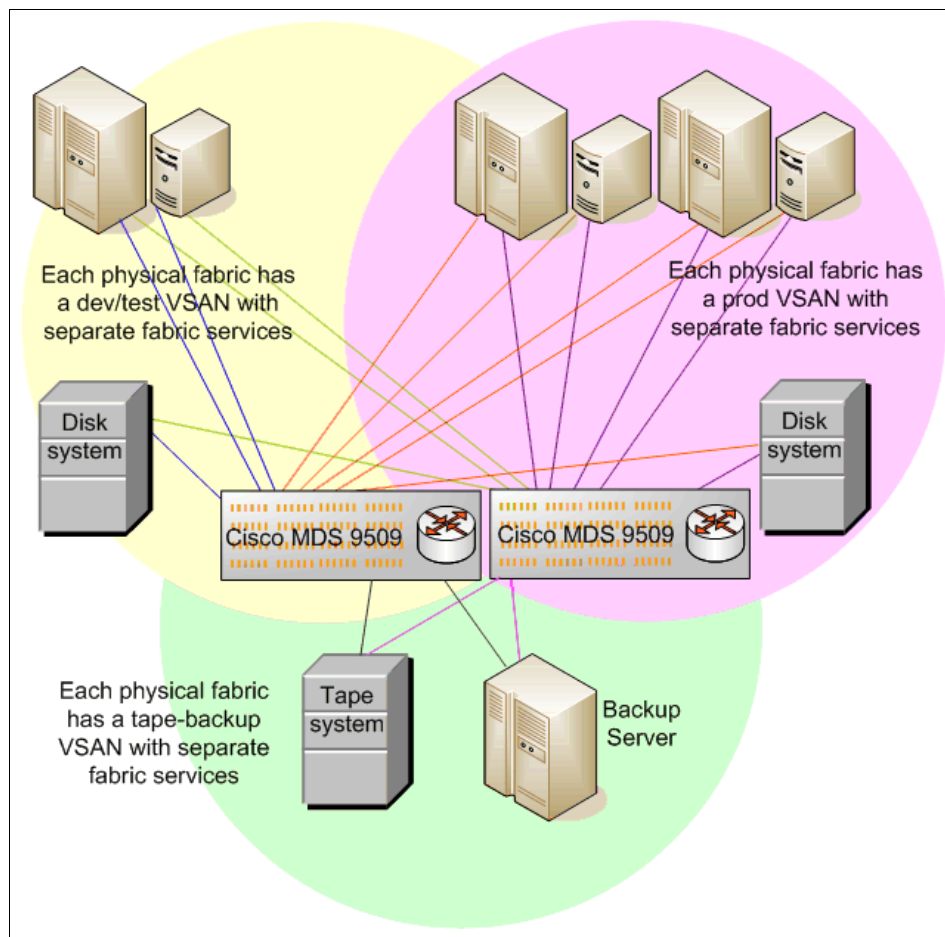


Figure 7-5 Using VSAN, IVR to isolate subsidiaries with access to shared tape

7.3.3 Isolation of multivendor switches and modes

You may have Fibre Channel switches from multiple vendors that each require different mode settings and behave slightly differently in the network. You might want to incorporate them into your network, but keep them isolated either for departmental reasons or simply keep the different modes of operation separate from each other. IVR gives the architect the confidence to combine switches from other vendors into the network, knowing that each VSAN has its own separate fabric services.

The Brocade VSAN in this case includes initiator devices attached to the Brocade switch and a single inter-switch link (ISL) port on the Cisco switch. The Cisco switch provides and manages the routing between the Brocade VSAN and the Cisco VSAN, which contains all of the other ports on the Cisco switch.

Figure 7-6 shows how switches from Brocade, McDATA, or QLogic can be incorporated into the network and yet be isolated into a separate VSAN, with IVR providing data sharing across the network.

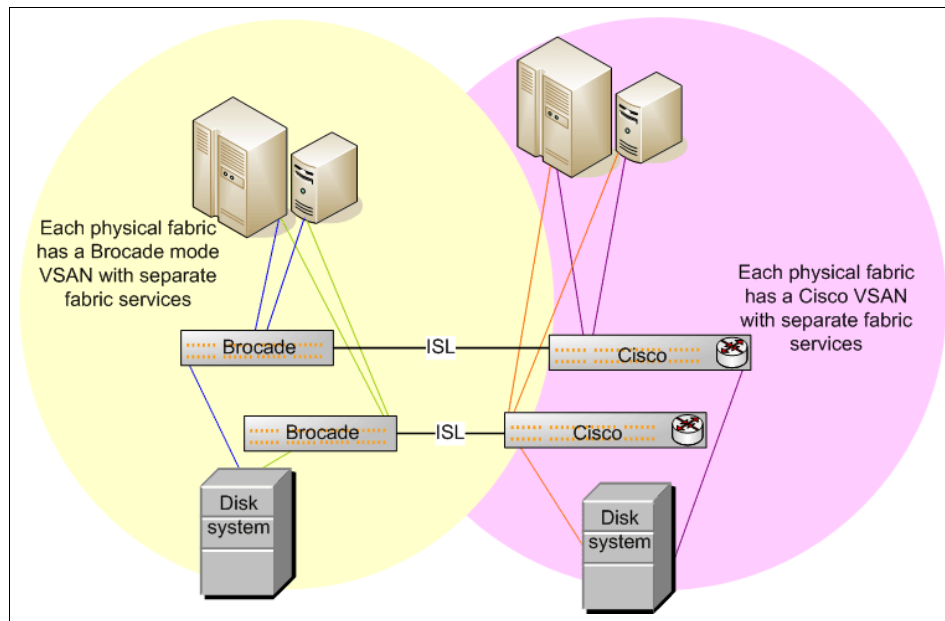


Figure 7-6 Using VSAN and IVR to provide multivendor isolation and integration

7.4 Managing scalability with IVR

It is also possible to use VSAN and IVR to prevent fabrics from growing too large because you add more ports on a modular basis. While modern switches typically include ways to limit Registered State Change Notifications (RSCNs), such as zone-limited RSCNs, some architects prefer to create smaller fabrics as a way to improve high availability. This is not necessarily a good idea because it introduces additional complexity. If you think about an Ethernet network, typically you do not want to introduce multiple routers into your core network. The same applies to VSAN and IVR. The rule of thumb is to *resist the urge to route*.

You can use VSAN and IVR to manage the growth of your network and to limit the impact of a misconfiguration in a network that is subject to regular change. As you add additional switch groups, you can isolate them as shown in Figure 7-7.

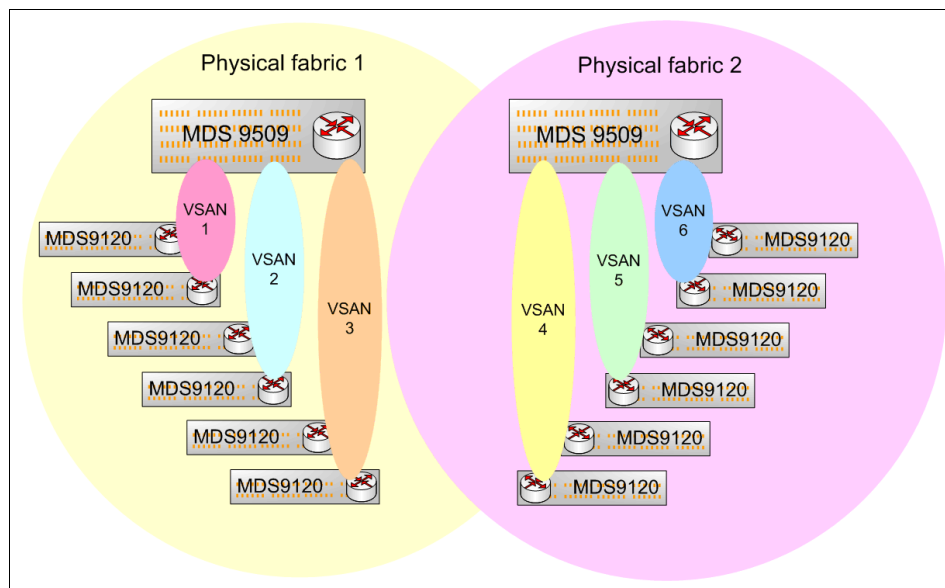


Figure 7-7 Using VSAN and IVR to manage scalability

7.5 Storage migration using IVR

Cisco's VSAN and IVR features can be valuable when considering migration from previous generation technologies. Figure 7-8 shows a site running an EMC CLARiiON FC4700 disk system and EMC Connectrix 1 Gbps switches, which are OEM from McDATA.

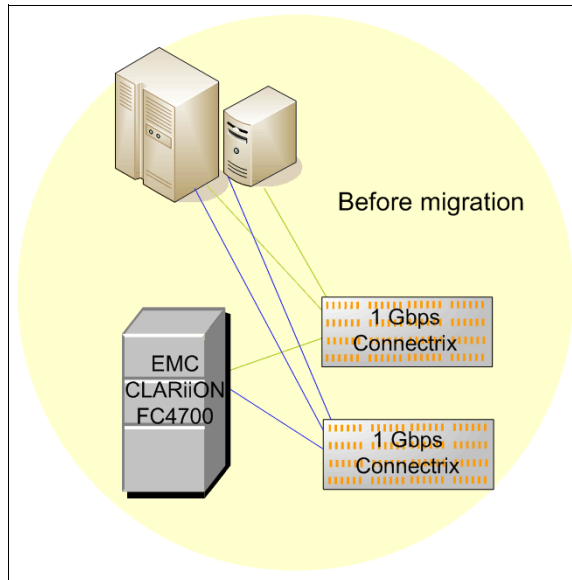


Figure 7-8 An existing SAN ready for upgrade

4. You can then remove the old fabric and disk system as shown in Figure 7-10.

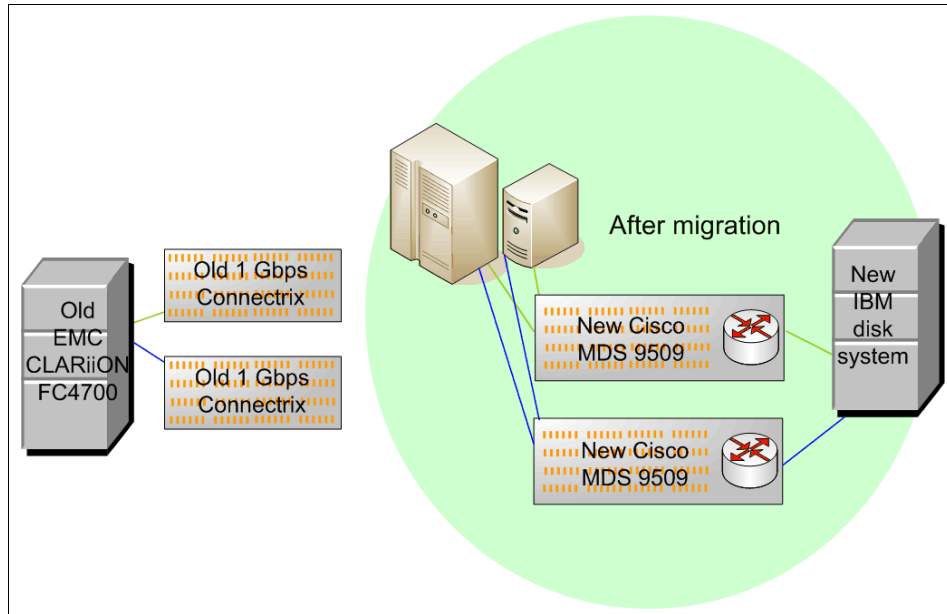


Figure 7-10 Connecting servers to the new fabric and disconnecting the old SAN



Cisco best practices

When you purchase a Cisco MDS 9000 switch, you also purchase a Fibre Channel (FC) router. The MDS 9000 is a feature rich family. The value you gain from the products depends on which features you choose to implement and how you go about it.

8.1 To route or not to route?

That is the question. If you think about an Ethernet network, typically you do not want to introduce multiple routers into your core network, and the same applies to your Fibre Channel network. Because every port on a Cisco MDS is also a router port, it can be easy to over-use virtual storage area network (VSAN) and Inter-VSAN Routing (IVR). The rule of thumb is to *resist the urge to route*.

Chapter 7, “Cisco solutions” on page 109, has examples of scenarios where you might want to route. These include the following situations:

- ▶ SAN extension over IP networks
- ▶ Lowering connection costs using Small Computer System Interface over IP (iSCSI)
- ▶ Achieving isolation and interoperability between different business units
 - Separating production from development
 - Separating subsidiary A from subsidiary B
 - Separating multivendor switches
- ▶ Managing scalability as your SAN environment grows
- ▶ Migrating from an older storage environment to a newer one

Keep in mind that a requirement for complete isolation does not imply a requirement for routing. Routing is required only when you want isolation alongside the capability of some access between isolated environments.

Before architects and implementers deploy IVR, read the white paper *Inter-VSAN Routing with the Cisco MDS 9000 Family of Switches & Cisco SAN-OS 2.1* from Cisco on the Web at:

http://www.cisco.com/en/US/netsol/ns514/networking_solutions_white_paper0900aecd80285738.shtml

Important: Be careful with introducing unnecessary VSANs into a SAN Volume Controller (SVC) environment. Refer to 8.5.1, “SAN Volume Controller interoperability” on page 129, for more details.

8.2 Piloting new technology

When you plan to use advanced features, such as FC over IP (FCIP), iSCSI, VSAN and IVR, compression, FCIP Write Acceleration (FCIP-WA), and Fibre Channel tape acceleration, it is important to implement the new technology initially as a pilot. You must do so with the understanding that the experience gained in your own environment will always be slightly unique.

When implementing leading-edge technologies, many clients prefer to avoid the uncertain outcomes that a pilot implies. Instead they secure implementation guarantees from vendors. In fact, the outcome can never really be guaranteed. Piloting allows the solution to be tailored based on lessons learned in your own environment.

8.3 iSCSI issues

Before architects and implementers deploy an iSCSI solution, read the white paper *iSCSI Design Using the MDS 9000 Family of Multilayer Switches* from Cisco on the Web at:

http://www.cisco.com/en/US/netsol/ns514/networking_solutions_white_paper09186a0080171d9e.shtml

There is a lot of enthusiasm and excitement about iSCSI. One thing that is certain is that it is slower than using Fibre Channel. In the white paper *Guide to iSCSI Performance Testing on the Cisco MDS 9000 Family*, testing was performed with an xSeries 345 (x345; dual 2.66 GHz with 1 GB RAM) running Microsoft Windows 2000, and an EMC CLARiiON CX600. This white paper offers advice on tuning and I/O block sizes, but remains inconclusive about the performance of iSCSI for specific real-world applications.

You can find this paper on the Web at:

http://www.cisco.com/en/US/products/hw/ps4159/ps4358/products_white_paper0900aecd801352e3.shtml

In Figure 8-1, the I/O response time in the bottom right graph is presumed to be in milliseconds.

The use of TCP/IP offload engine (TOE) Ethernet cards is one way to reduce the CPU overhead of iSCSI processing. However, using specialized hardware also detracts from the cost and ease-of-use arguments.

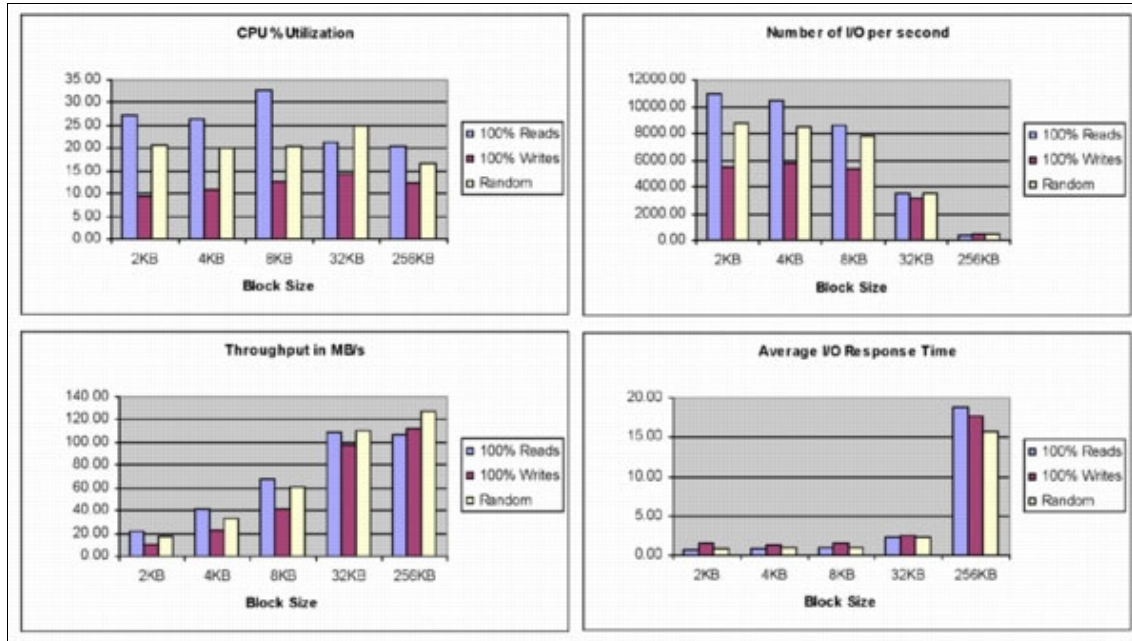


Figure 8-1 How block size affects iSCSI performance

8.4 IP network issues

Before architects and implementers deploy an FCIP solution, read the white paper *Designing FCIP SAN Extension for Cisco SAN Environments* from Cisco on the Web at:

http://www.cisco.com/application/pdf/en/us/guest/netso1/ns378/c649/cdccont_0900aecd800ed145.pdf

You can find additional white papers at the following Web site:

http://www.cisco.com/en/US/netso1/ns340/ns394/ns259/networking_solutions_packages_list.html

Also ask your telecommunications company about high Quality of Service (QoS) managed links, including information about offerings such as IP over SONET/SDH. Help your telecommunications company clearly understand your network quality expectations. And, make sure that you clearly understand the cost and management implications of any decisions that you make.

8.5 Interoperability

Cisco switches are explicitly supported for heterogeneous interconnection with:

- ▶ Brocade (including native mode)
- ▶ McDATA
- ▶ IBM *@server* BladeCenter®
 - Optical pass-through modules
 - QLogic modules
 - McDATA modules
 - Brocade modules

Heterogeneous IVR was announced to the field by IBM on 24 May 2005 in a document *IBM announces Cisco MDS 9000 Intelligent Fabric features*.

Note: At time of writing, the following interoperability documents did not include mention of heterogeneous IVR.

The Cisco MDS interoperability matrixes are available by clicking the individual product details at the following Web site:

http://www-1.ibm.com/servers/storage/san/c_type/

You can also find Cisco MDS interoperability matrixes by selecting the product information files from the following site:

<ftp://service.boulder.ibm.com/storage/san/cisco/>

The maximum hops supported is seven including one longwave ISL.

8.5.1 SAN Volume Controller interoperability

When deploying VSANs in an SVC environment, give careful consideration to how the different VSANs will access the SVC, given that in an SVC environment, typically all disk I/O travels via the SVC. Each SVC node has four Fibre Channel ports, so each can be part of a maximum of four fabrics or VSANs without routing via IVR.

8.6 Designing for availability

Redundancy can be used to offer some protection against hardware failure and against disruptive fabric events. The best practice with fabric design is usually to use dual redundant fabrics which have traditionally implied redundant hardware. With the advent of the VSAN, however, the equation is not so straightforward. The topics that follow briefly discuss some areas to focus on with respect to ensuring availability.

8.6.1 Fibre Channel router hardware

All members of the Cisco Fibre Channel router/switch family include redundant hot-swappable power supplies and fans, as well as the ability to restart a failed supervisor process.

As offered by IBM, the Cisco MDS 9509 also comes with two supervisor modules, each with its own control engine and crossbar fabric. The two crossbar fabrics operate in a load-shared active-active mode. Each crossbar fabric has a total switching capacity of 720 Gbps and serves 80 Gbps of bandwidth to each slot. The system does not experience any disruption or any loss of performance with the removal or failure of one supervisor module.

Even though a single Fibre Channel router/multilayer switch may be designed for high availability, two of them provide a more robust solution. Remember also that, although there are dual redundant clock modules in the Cisco MDS950n directors, if one clock module needs to be replaced, a director outage is required because these modules are not hot-pluggable.

Common sense and the principles of balanced system design must prevail. In some cases, you might assess that deploying two 99.999% available routers or directors, such as the MDS 9509, is unnecessary if the selected midrange back-end disk system is only 99.99% available.

8.6.2 Nondisruptive software upgrade

The Cisco MDS 9500 Series of Multilayer Directors supports the ability to upgrade the supervisor module and the switching module software on the fly without disrupting traffic flowing through the switch. During this upgrade process, the standby supervisor is upgraded. Then control automatically moves over to the upgraded supervisor (standby has become active, active has become standby), and the standby is automatically upgraded to the same level of software. Dual versions of software on the MDS are not supported.

This process allows maximum flexibility in upgrading the software while providing a path to revert back to a known level of stable software.

8.6.3 Inter-switch links

Relying on a single physical link between switches reduces the overall redundancy in the design. Redundant inter-switch links (ISLs) provide failover capacity should a link fail. Some architects are content with two separate fabrics, each with their own ISL. The issue with this is that, if an ISL fails, then you are in fabric failover, because one of the fabrics is effectively broken. By having at least two ISLs per fabric, you avoid fabric failover in the event that a link should fail.

Cisco's PortChanneling feature allows you to load balance across multiple ISLs. When deploying iSCSI or FCIP networks, check that the ports are capable of being PortChanneled.

If running FCIP over an externally provisioned network, architects should also understand whether multiple links take separate physical routes through the end-to-end network.

8.6.4 VSAN and IVR

VSAN and IVR can also be used to protect against disruptive fabric events. Fabric-level events have the potential to disrupt all devices on the fabric. Mistakes made when adding a switch or changing zoning configurations could ripple through the entire connected fabric.

In large or complex networks, designing with separate VSANs helps to isolate the scope of any such events. In smaller networks, however, configuring VSANs and IVR may simply add unnecessary complexity.

8.6.5 Backup

Some members of the Cisco MDS family offer a compact flash memory card. The settings can be backed up there, but it is generally better to back up to an external FTP server. Remember to apply sound change control procedures and retention of previous good configurations.

8.7 Designing for security

Security also impacts high availability since one of the leading causes of downtime is human error.

Role-based administration

The Cisco MDS 9000 Family of Multilayer Directors and Fabric Switches support a role-based security methodology to ensure that only authorized individuals have access to critical functions within the fabric. Each user is assigned to a role, better known as a group_ID, which is given a specific access level within each fabric. This access level dictates the commands, or more specifically, to which nodes of the command-line interface (CLI) command parser tree the particular role has access.

Centralized management

Roles can be defined and assigned locally within a switch by using CLI commands or can be centralized in a RADIUS server for easier management. Two default roles are provided: Network Administrator (full access) and Network Operator (read-only access). Up to 64 custom roles can be defined by the user. Only a user within the Network Administrator role may create new roles.

VSAN administration

VSANs contribute to the security of a network by maintaining isolation of devices onto different fabric services even though they may share a physical fabric. Because each VSAN is a separate virtual fabric, each VSAN has its own set of role-based administrators. This adds to the security of the SAN and makes it safer to administer a shared fabric.

Encryption

You must answer these key questions:

- ▶ How do I prevent someone from viewing or modifying confidential data?
- ▶ Does connecting a Fibre Channel network to an IP network impact the integrity of my data?

The principal security mechanism on a Fibre Channel fabric is zoning. Zoning works like an access control list (ACL), so that only devices that are on each others lists can talk to each other. VSANs provide an additional layer of security because they also define members. Non-members are not given access unless they are a member of another VSAN which has IVR access to the first VSAN.

Beyond ACLs, you can deploy encryption. The Cisco Multiprotocol Services (MPS) 14+2 line card and Cisco MDS 9216i switch both offer integrated hardware-based IP security protocol (IPSec) support, providing wire-rate encryption and decryption with Advanced Encryption Standard (AES) and Triple Data Encryption Standard (3DES).

Cisco provides a range of additional features which are designed to maintain the integrity of the fabric. For a discussion about some of these ideas, refer to the

article “Security - Beyond Zoning” in the Cisco user’s magazine *Packet* in the second quarter 2005 edition. You can find this edition on the Web at:

http://www.cisco.com/application/pdf/en/us/guest/netso1/ns513/c666/cdccont_0900aecd802c2b74.pdf

8.8 Designing for performance

The topics that follow briefly discuss some of the items that you need to consider when designing for performance.

8.8.1 Hardware selection

At the time of writing, Cisco supports maximum port speeds of 2 Gbps. Some disk systems, such as the DS4800, use 4 Gbps connections, but most are based on multiple 2 Gbps connections.

Place emphasis on the principles of balanced system design and promote lower throughput ports, such as iSCSI (shared 1 Gbps), and host-optimized line cards (with 3.2:1 oversubscription) as being suitable for most servers.

There are two common approaches to designing for performance. One approach is to try to understand the peak workloads, and then to project growth and allocate bandwidth accordingly. The other is to make every connection as high speed as possible, within certain cost restrictions.

When connecting back-end disk subsystems, ISLs, tape libraries, and high-throughput servers, use target-optimized ports:

- ▶ Four ports on the MDS 9120 20 port switch
- ▶ Eight ports on the MDS 9140 40 port switch
- ▶ Fourteen ports on the MDS 9216i and the MPS 14+2 line-card
- ▶ Sixteen ports on the MDS 9216A and the 16 port line card

When connecting low-throughput servers, you can use host-optimized ports:

- ▶ Sixteen ports on the MDS 9120 20 port switch
- ▶ Thirty-two ports on the MDS 9140 40 port switch
- ▶ Thirty-two ports on the 32 port line card

Or you can use iSCSI ports or Fibre Channel Arbitrated Loop (FC-AL). Vendors with an iSCSI solution are generally bullish about iSCSI, but there is still a lot of varying information in the market about performance. We recommend that you run a pilot test for iSCSI solutions before you place them into production.

While the bandwidth on most shortwave Fibre Channel networks installed today is underused, there are applications that perform high volume sequential I/Os, which can flood Fibre Channel ports. Remember that although a disk system may only have one, two, or four back-end 2 Gbps FC-AL loops, the system may be able to feed a lot of data onto the network from cache.

Note: You can use the Cisco SAN Extension Tuner to help understand and optimize FCIP performance. The tuner generates SCSI I/O commands that are directed to a specific virtual target. It reports I/Os per second and I/O latency results. SAN Extension Tuner is included with the FCIP enablement license package.

8.8.2 FCIP compression and FCIP-WA

Results can vary significantly from site to site when deploying FCIP compression and FCIP-WA.

FCIP compression

Cisco implements hardware compression in the MDS 9216i and on the MPS 14+2 line card. In IP Services, module compression is implemented in software.

Software compression has one advantage. In some modes, multiple small Fibre Channel frames can be stacked into a single IP packet to make more efficient use of the payload space.

Both approaches use an LZS compression algorithm which searches for repeat data strings in the input data stream and replaces these strings with data tokens shorter in length than the original data. A compression history table is built of these string matches, pointing to previous data in the input stream. The net result is that subsequent data in the stream is compressed based on earlier data. The compression ratio is higher with repeated data and lower with greater randomness.

Because compression varies as the data varies, we recommend that you perform a pilot for the different compression modes in your environment until you arrive at the one that delivers you the best throughput.

FCIP-WA

FCIP-WA performance depends heavily on the traffic profile of the SCSI operations being performed. The following I/O characteristics are well-suited to benefit from FCIP-WA:

- ▶ Long distance high latency network
- ▶ Write intensive I/O
- ▶ High number of small SCSI writes (rather than low number of large writes)
- ▶ Disk system has low write latency

This suggests that the best use for FCIP-WA is in disk system replication. Across a 100 km link, replication using FCIP-WA can be expected to deliver between 5% and 20% throughput improvement, with 10% being a realistic expectation.

Once again the benefits of FCIP-WA vary for each client. Piloting is the only appropriate way to be sure if it is a good fit for your network.



Cisco real-life solutions

The case studies in this chapter are intended to show real-life situations. They also show the pragmatic solutions that have been developed around them using the Cisco MDS multiprotocol technology.

Two case studies are presented here. The first one implements Inter-virtual storage area network (VSAN) Routing (IVR), Fibre Channel (FC) over IP (FCIP) tunneling over a campus IP backbone, and Small Computer System Interface over IP (iSCSI). The second one implements IVR and asynchronous disk mirroring via FCIP tunneling over a long-distance IP network.

Important: The solutions and sizing estimates that we discuss or make in this chapter are unique. Make no assumptions that they will be supported or apply to each environment. We recommend that you engage IBM to discuss any proposal.

9.1 University ZYX

University ZYX provides teaching and research services to 30 000 students across ten faculties: arts, business, architecture, fine arts, education, engineering, law, medicine, science, and theology. In addition, there are 20 interdisciplinary research clusters focused on emerging fields.

The example that follows departs from best practice in some ways. Based on a history of no Fibre Channel switch failures over four years, and to save cost, the customer decides that having dual physical fabrics is not important at their disaster recovery (DR) and engineering sites, especially given the capability to create VSANs each with their own fabric services. The customer also decides that a single director class router or switch will provide them with adequate reliability at the core of their network.

This approach is also echoed in their willingness to use singly-attached hosts and iSCSI for non-critical servers in many cases. The customer reasons that having less hardware is not so important, because they gain rich functionality with the MDS family, such as routing capability on every port of every switch.

9.1.1 Initial growth

University ZYX installed their first SAN in 2001 with an IBM TotalStorage Enterprise Storage Server® (ESS) model F20 and several 1 Gbps switches, primarily to support their PeopleSoft (now JD Edwards from Oracle) applications. In 2003, they added an IBM TotalStorage DS4500 disk system to hold Tivoli Storage Manager pools and other data that did not require tier one storage. That same year they also added a DS4500 disk system at their cold disaster recovery site across town in the IBM data center.

In early 2004, they added a pair of SAN Volume Controller (SVC) nodes with a 32 TB license to allow virtualization of their storage volumes for increased management flexibility. Figure 9-1 shows the layout of the SAN environment after four years.

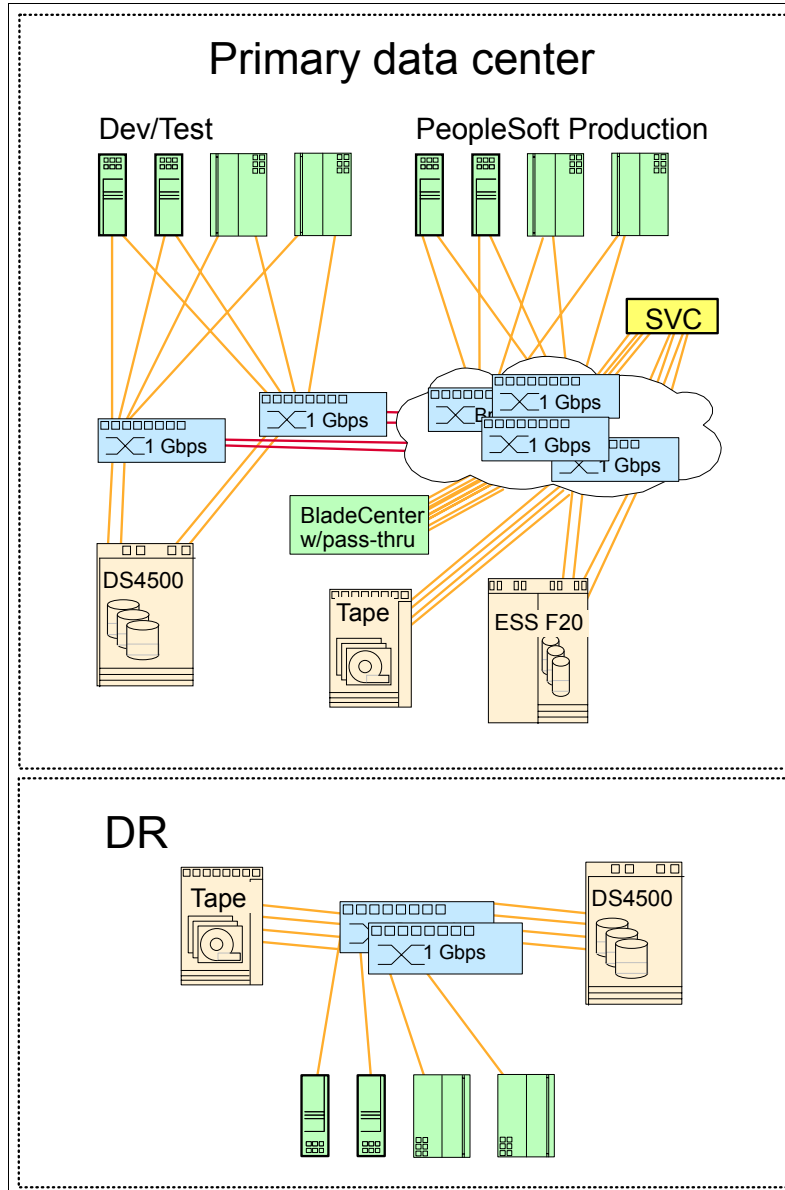


Figure 9-1 University ZYX SAN environment after four years

9.1.2 Lease expiration

At the end of 2004, the University realized that the four-year lease on the ESS and the 1 Gbps Fibre Channel switches was only six months away from full term. They began planning for replacements.

As part of this they wanted to leverage iSCSI to provide a flexible, low cost connection to a range of small servers. They also wanted to use their existing campus IP backbone to provide some sharing capability with the School of Engineering, which was approximately 2 km down the road. Engineering wanted to purchase their own disk system as well.

At the time, the Cisco MDS was the only Fibre Channel solution with FCIP and iSCSI being offered by IBM. Since the University was a content user of Cisco IP networking devices sourced from IBM Global Services already, they decided to proceed with Cisco MDS equipment. Based on three years of fault-free operation on the ESS, they also selected IBM TotalStorage DS8100.

9.1.3 Design and purchase of new systems

Engineering went ahead with the purchase of an IBM TotalStorage DS4300 and a Cisco MDS 9216i. They chose to run a single physical fabric and use the VSAN feature of the Cisco MDS to create logical subfabrics. They also knew that by deploying Cisco MDS, they had options for iSCSI and FCIP at their fingertips.

The IT services department decided to purchase a single MDS 9509 and initially decided on 3 x 32 port line cards and 2 x 14+2 Multiprotocol Services (MPS) card. After a review with Cisco, IBM, and the IBM Business Partner, it was decided that this configuration did not provide enough target-optimized ports. Two additional MPS 14+2 cards were substituted for one of the 32 port line cards. A decision was also made not to use CompactFlash to back up the MDS configuration, but to back it up to an external File Transfer Protocol (FTP) server instead.

Some of the 1 Gbps switches were owned and not leased. The customer decided to retain some of them for an additional year and to use VSAN and IVR to isolate and connect the 1 Gbps switches. The SVC adds complexity in that all data flows through it, so it either has to be a member of every SAN, or it must be accessed thorough IVR. When all of the 1 Gbps switches are retired, this aspect of the design could be simplified.

Figure 9-2 shows the network after the new multiprotocol switches/routers and the new IBM TotalStorage DS8100 and DS4300 were installed.

IVR provides the test/development system with access to the tape library and to the main DS8100 when needed. Each VSAN *must* have access to the SVC.

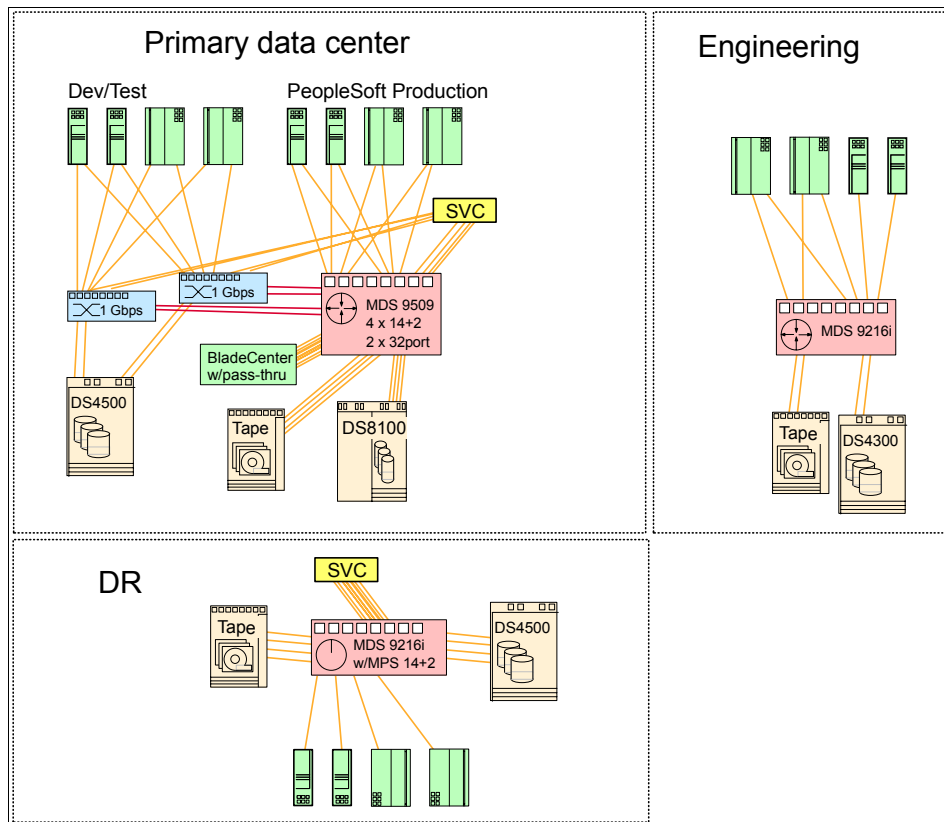


Figure 9-2 The network after installing the DS8100, DS4300, and Cisco MDS

9.1.4 Deployment of iSCSI and FCIP

iSCSI was deployed using the IP ports on the MPS 14+2 line cards. The main users are faculty file servers, which have a requirement for online or near-line storage. Additional IBM TotalStorage DS4000 EXP100s with SATA drives were purchased for the DS4500 to provision low-cost storage for the iSCSI-attached servers.

After consideration of a dedicated longwave Fibre Channel connection to engineering, the University opted instead to use the existing campus IP

backbone and use Cisco's FCIP functionality to connect the two sites, believing this to be a more flexible solution. The FCIP link is in a transit VSAN with IVR providing communication through the transit VSAN to the far site.

Figure 9-3 shows the physical network plan including FCIP and iSCSI.

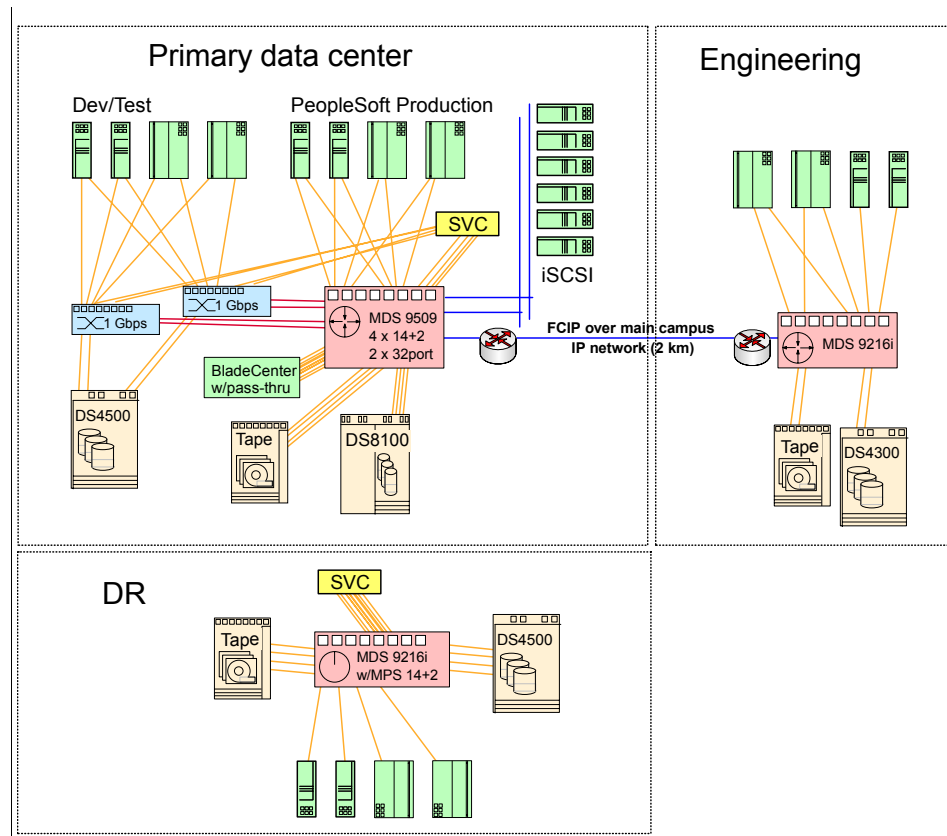


Figure 9-3 University ZYX planned network with FCIP and iSCSI

9.1.5 SVC synchronous replication for disaster recovery

The final stage in this technology refresh was the purchase of an SVC for disaster recovery and the establishment of a link to the IBM data center where the University's disaster recovery systems are housed. The advice from IBM is that a maximum 10 microsecond (ms) latency can be tolerated on SVC synchronous replication.

To estimate the latency of an IP link can be tricky. There is the 5 ms Fibre Channel cable latency per km in each direction, plus another 5 ms each way in

each Fibre Channel device, plus IP router latency which can be anything from 10 ms each way to much higher, especially if filters are applied. Then if traffic passes through other devices, such as firewalls, they also need to be factored in.

The real question is round-trip delay, rather than latency, since delay includes any congestion that may occur. Provided the round-trip delay never goes higher than 10 ms, then the replication will be successful using FCIP.

The University decided to proceed with an FCIP link. Figure 9-4 shows the physical network including SVC sync replication.

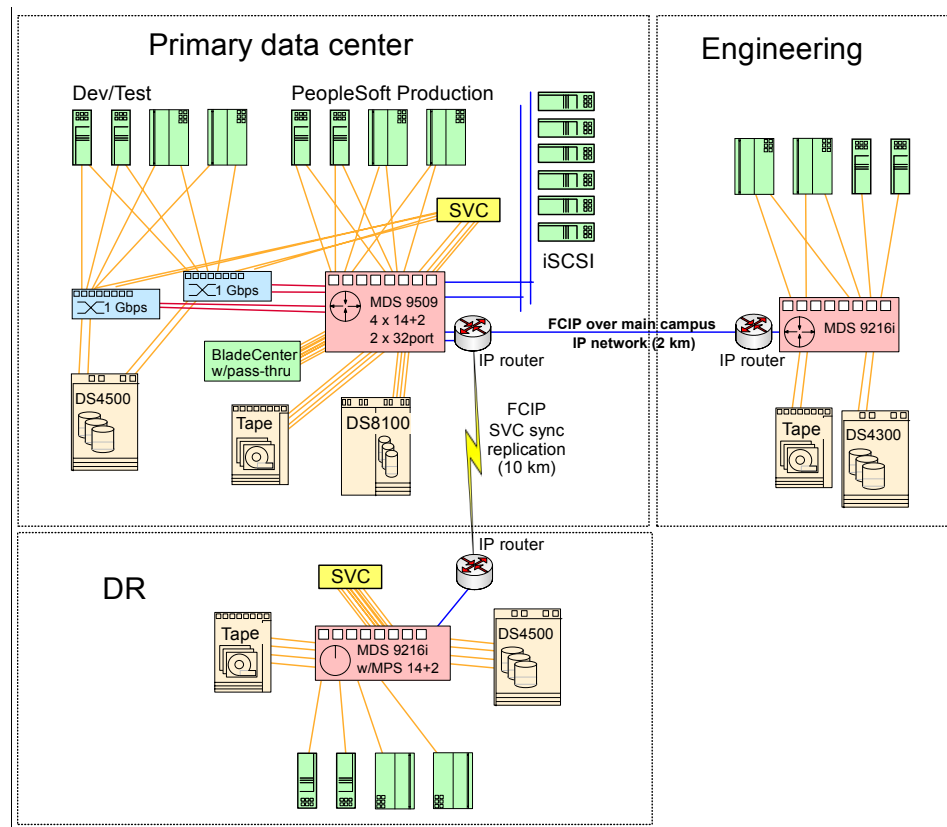


Figure 9-4 University ZYX network including SVC sync replication

9.2 Power Transmission Company ZYX

Power Transmission Company ZYX is a state-owned business that is set up to own and operate the high voltage electricity transmission grid that links generators to distribution companies and major industrial users. The company's head office is located in an area that is susceptible to earthquakes, but there is currently little provision for disaster recovery. There is a main engineering office about 600 km to the north, and a customer service call center about 400 km to the south.

9.2.1 Existing systems

The company uses an asset metering application, which monitors the condition of substations, meters, transmission lines, and other assets as events occur. The data is integrated with back-office systems and analytical tools for improved decision making.

Secure data communications are provided between remote devices and back-end applications. The system gathers, filters, and communicates data on usage and status. Communications gateways are based on Arcom Controls *Director series* equipment.

The production system runs on a pSeries server with AIX. Storage is shared between production and development/test on a CLARiiON CX600 disk system and an ADIC Scalar 100 tape library with two SCSI drives. Backups are done over the LAN using Tivoli Storage Manager. Current Fibre Channel switches are 1 Gbps.

Software includes IBM WebSphere® MQ Telemetry transport, IBM WebSphere Business Integration Adapters for utility industry processes and applications, and IBM WebSphere MQ Everyplace® software.

Figure 9-5 shows the existing SAN environment at Power Transmission Company ZYX.

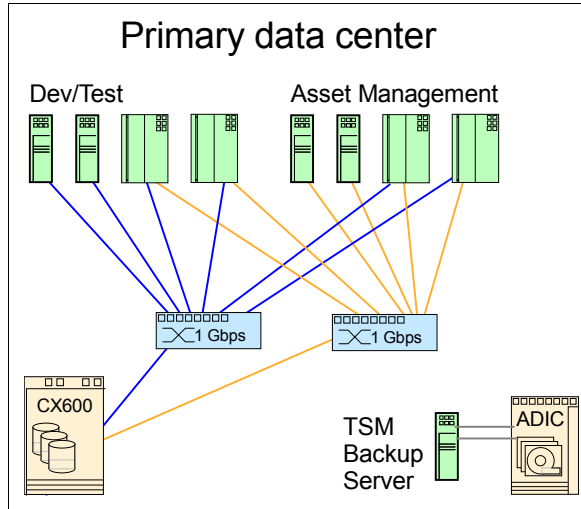


Figure 9-5 Existing SAN environment at Power Transmission Company ZYX

9.2.2 IT improvement objectives

The first main objective is to provide increased capacity and performance on the SAN in line with new application modules and increased usage of the system. The second main objective is to reduce business risk associated with IT infrastructure and site failure, including introducing:

- ▶ Dual-fabric attachment for all hosts
- ▶ Separation of development/test and production
- ▶ Improved backup/restore throughput
- ▶ Replication of data to a disaster recovery site

The company is an existing Cisco customer for Ethernet switches and IP routers and has been happy with their products. The new IT Infrastructure Manager also has a background as a network manager. The manager has a good relationship with the Cisco account team and has a some knowledge of the MDS family gained by reading the Cisco user's magazine *Packet*. It has become clear that FCIP tunneling will be the most cost-effective way to achieve remote asynchronous replication. Power Transmission Company ZYX has decided to use Cisco MDS multiprotocol switches/routers in rebuilding its SAN environment.

9.2.3 Deployment of new technology and establishment of the disaster recovery site

The company decided to set the 1 Gbps switches aside for use in the development/test environment. Because they plan to extend their Fibre Channel

network to other servers at a later time, they elect to deploy two Cisco MDS 9509 multiprotocol directors at the core of their new network. The company considered using a single MDS 9216i at the disaster recovery site since it could use the VSAN feature to provide separate fabric services. In the end, the company chose to deploy two MDS 9216i multiprotocol switches at the disaster recovery site.

As part of the technology refresh, the existing tape library and several of the existing servers will be redeployed to the disaster recovery site. The disaster recovery site will initially provide cold disaster recovery only.

Figure 9-6 shows the site with separation of development/test from production and of the tape VSAN. It also shows establishment of a disaster recovery site. Fibre Channel connections are color-coded on a per-VSAN basis at each site.

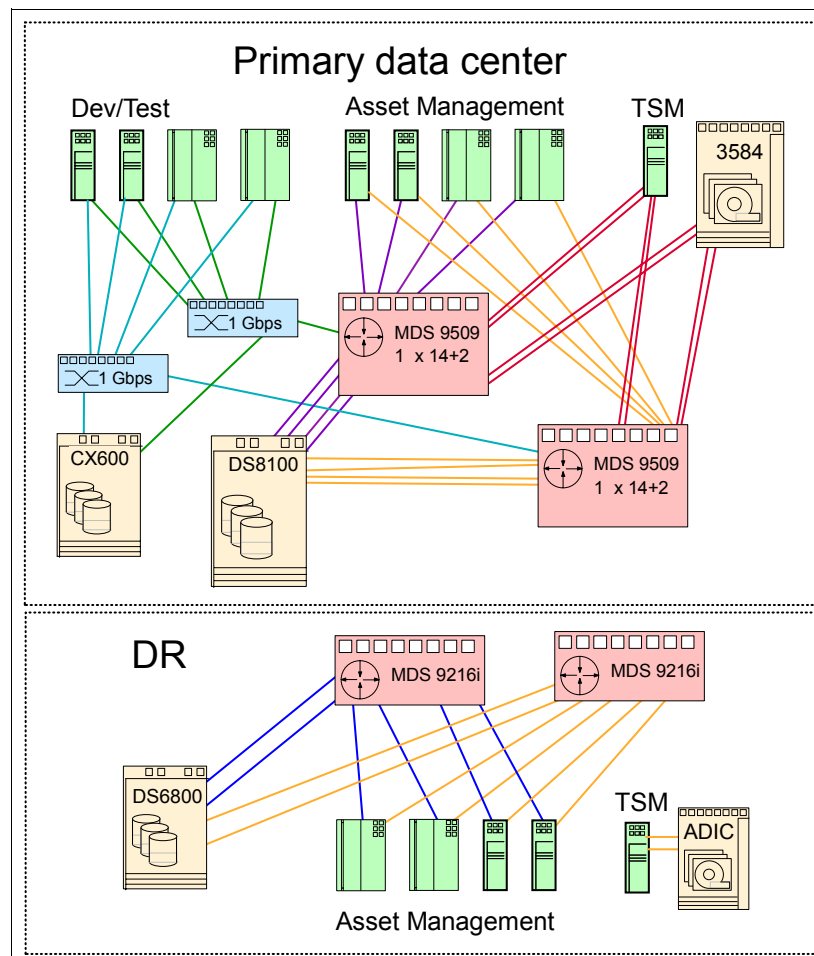


Figure 9-6 Development/test break from production; DR site established

Use of VSAN and IVR

VSANs have been deployed to provide isolation of the 1 Gbps switches. Each of these VSANs includes all of the ports connected to that 1 Gbps switch. IVR ensures that occasional access between the development/test and production areas can be accommodated.

A VSAN has also been used to create a separate fabric for the tape backup/restore solution. When other applications are added to the SAN, it is planned that these will be separated into their own VSAN to better manage change control across different business units.

9.2.4 Global Mirroring established to the disaster recovery site

Remember that when using Global Mirroring, this is essentially the same as using Global Copy plus periodic uses of FlashCopy® to provide a safe recovery point, so extra disk space must be allowed for the FlashCopy copies.

A sizing for the FCIP link was done using the Async PPRC Bandwidth Sizing Estimator available from IBM and IBM Business Partners. Figure 9-7 shows the values entered into the fields of the Async PPRC Bandwidth Sizing Estimator.

Input (in blue box below)					
Primary Site Configuration:			Workload (aggregate for primary site):		
Number of ESSs		1			
Configuration per ESS:				zOS	Open
Number of HAs to host(s)		6	IO rate (SIO/sec)	0	5000
Number of HAs to 2ndary		2	Read Hit Ratio	0.92	0.6
Distance to 2ndary (miles)		500	Write Hit Ratio	1	0.87
link data compression factor		2	Read/Write Ratio	3	2.33
			Destage rate	0.5	0.5
			Seq Prestage Rate	0.4	0.4
			Avg Block size (KB)	27	4
Desired Consistency Group Interval Time (sec)		0			
Desired max drain time (sec)		60			

Figure 9-7 Input to the Async PPRC Bandwidth Sizing Estimator

Figure 9-8 shows the output of the values entered in Figure 9-7.

Output (below)	
Min Aggregate link BW required (MB/sec)	5
Link type	Min number of active links required
OC-3	1
OC-12	1
OC-48	1
GigE	1

Figure 9-8 Output from the Async PPRC Bandwidth Sizing Estimator

Based on a 60 second drain time, we estimate that a bandwidth of 5 MB/s (or about 40 Mbps) is required. The **ping** round-trip time (RTT) on this network is observed to be approximately 20 ms.

Note: You can use the Cisco SAN Extension Tuner to help understand and optimize FCIP performance. The tuner generates SCSI I/O commands that are directed to a specific virtual target. It reports I/Os per second and I/O latency results. SAN Extension Tuner is included with the FCIP enablement license package.

The IBM TotalStorage DS6800 was configured with 12 ranks of 73 Gb 15 K RPM drives (48 drives in total) as one storage pool using Redundant Array of Independent Disks 5 (RAID5). During the design phase, using IBM Disk Magic, we checked to ensure that we had enough performance in the disaster recovery DS6800 to process 5000 input/output processors (IOPs), plus Global Mirror, and head-room for FlashCopy copies to be created. Figure 9-9 shows the utilization statistics from IBM Disk Magic on this configuration.

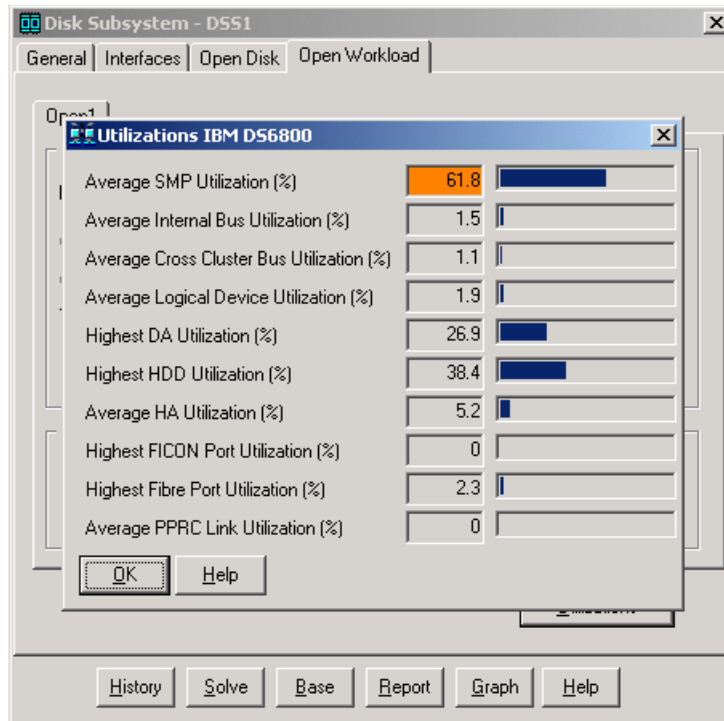


Figure 9-9 Utilization statistics for the disaster recovery DS6800 at 5000 IOPs

The company also implements Tivoli Storage Manager to stream copies of the backup both to local tape. In addition, to the remote Tivoli Storage Manager server over the IP network, the Tivoli Storage Manager traffic is transferred at night when the link is largely unused. The disaster recovery Planning module of Tivoli Storage Manager is also implemented to document the required workflow to complete a recovery.

Figure 9-10 shows that Global Mirroring has been established using FCIP tunneling. A transit VSAN was established that includes the FCIP ports on the MDS 9509 multiprotocol directors. This avoids fabric disruption if the wide area network (WAN) link is broken for any reason.

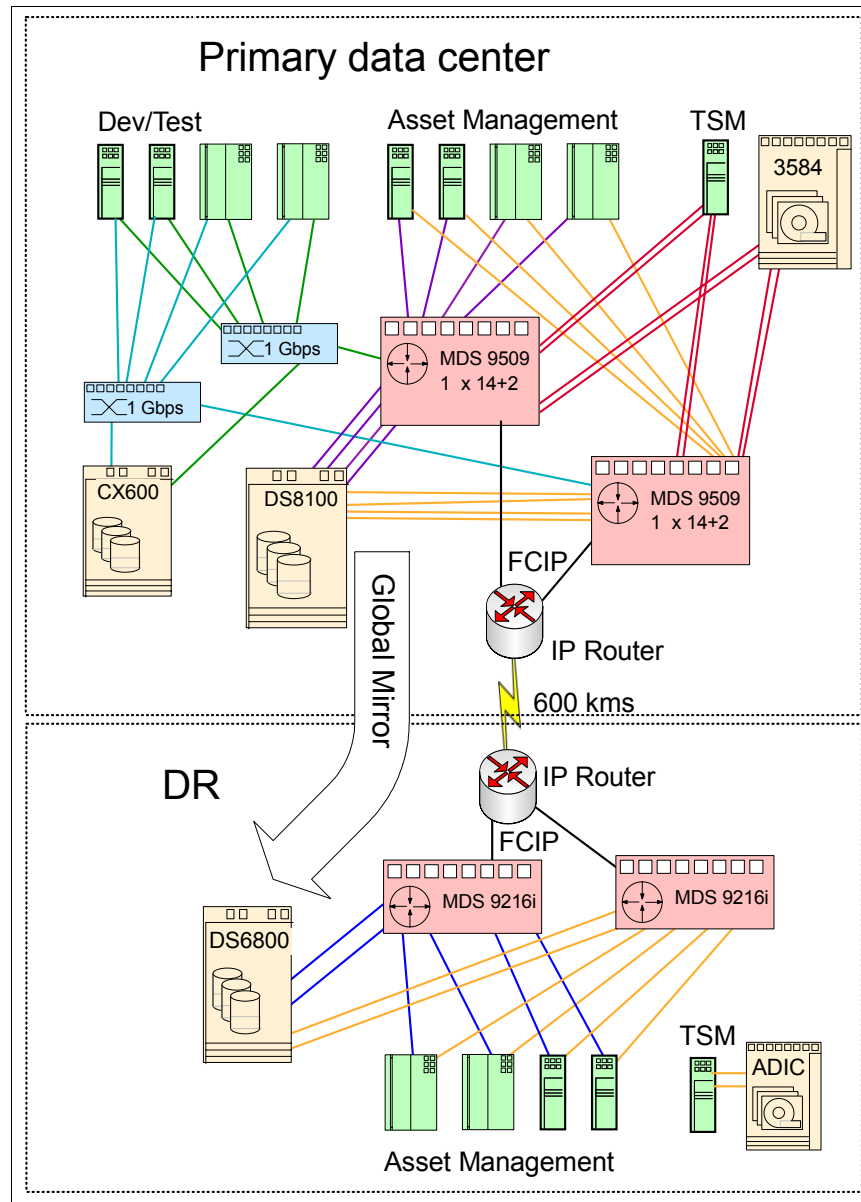


Figure 9-10 Global Mirroring established using FCIP tunneling and IVR



IBM TotalStorage m-type family routing products

This chapter describes the IBM TotalStorage m-type routers. IBM sells the following storage area network (SAN) routers:

- ▶ IBM TotalStorage SAN04M-R (2027-R04) with:
 - Entry level SAN router with four ports (two Fibre Channel (FC) ports at 1 Gbps speed
 - Two 1 Gigabit Ethernet ports in a 1U rack space that is designed for open systems IBM TotalStorage SAN solutions
- ▶ IBM TotalStorage SAN16M-R (2027-R16), which provides:
 - The 16 port base SAN router in a 1U rack space
 - The standard edition of SANvergence Management software
 - Rack-mount kit
 - Fully populated 2 Gbps shortwave small form-factor pluggables (SFPs) on all ports

10.1 Product description

The IBM TotalStorage SAN m-type routers support On Demand Business and other applications that require interconnection of SAN islands to provide any-to-any connectivity to fabric connected devices.

10.1.1 IBM TotalStorage SAN04M-R

The IBM TotalStorage SAN04M-R (2027-R04) is the McDATA Eclipse 1620 SAN Router. It provides four ports in a 1U rack space. Two ports are Fibre Channel 1 Gbps ports; the other two ports are intelligent ports for IP connectivity. Each of the two IP ports is provided with two connectors: one standard RJ45 and one SFP. Either one of those can be used, but not both at the same time. RJ45 is provided to connect Fast Ethernet and the SFP is for Gigabit Ethernet connections.

The IP ports support Internet Fibre Channel Protocol (iFCP) or Internet Small Computer Systems Interface (iSCSI) connectivity. The base functionality supports Fibre Channel, iFCP, Ethernet, and iSCSI with a maximum number of fifty iSCSI server connections. The optional firmware version (iFCP Enterprise) adds compression and Fast Write functionality. For enhanced management of the SAN router, clients may order Enterprise SANvergence Management Software.

SAN router features

The SAN router supports iSCSI, iFCP, and R_Port for trunking to both IP backbones and existing Fibre Channel fabrics. The SAN router connects to a wide range of end systems, including Fibre Channel, and Fibre Channel initiators and targets. The SAN routers support TCP/IP routing over extended distances at wire speed. The SAN routers offer:

- ▶ SAN internetworking for scalable and fault-tolerant SANs
- ▶ Support for full fabric, private, and public loop Fibre Channel devices
- ▶ Patent-pending Fast Write technology for maximizing throughput across long distances
- ▶ Compression for increased bandwidth

SAN router physical description

All the ports are located on the front of the SAN04M-R router. Only the Fibre Channel SFPs are field replaceable units (FRUs). The cooling fans are located on the rear. They are not hot-swappable.

When viewed from the front, standard power connections are located on each side of the device. The unit contains two independent power supplies for

redundancy and higher availability. Power supplies are not hot-swappable. The power supplies are input rated between 100 and 240 volts alternating current (VAC), at between 47 Hz and 63 Hz.

Fibre Channel and IP connectivity ports

There are two user-configurable Fibre Channel ports located on the front of the SAN router. The SFPs provide Fibre Channel connectivity at 1 Gbps. These ports can be configured as:

- ▶ FC_Auto (default)
- ▶ FL_Port
- ▶ F_Port
- ▶ L_Port
- ▶ R_Port

An LED to the left of each Fibre Channel port indicates the configuration and status of the associated port.

There are two intelligent ports for IP connectivity. Each IP port is provided with two connectors, one standard RJ45 and one SFP. Either connector can be used, but not both at the same time. RJ45 is provided to connect Fast Ethernet, and SFP is for Gigabit Ethernet connections. The IP ports support iFCP or iSCSI connectivity.

Management ports

Two management ports are located on the front of the SAN router. An RS-232 serial port can be connected to a VT100 terminal emulator for access to the command-line interface (CLI). An RJ45 port can be connected to the local area network (LAN) for out-of-band management using the SAN Router Element Manager or the SANvergence Manager. The RJ45 management port can be accessed by any PC on the LAN with a Web browser.

Operational features

Table 10-1 describes the SAN router features. Some features are optional and might not be present in some SAN router software versions.

Note that both SAN16M-R and SAN04M-R do not support Metro Fibre Channel Protocol (mFCP) links and mSAN routing since firmware release 4.7 and later. We discuss mFCP and mSAN routing in “mFCP and iFCP” on page 161.

Table 10-1 Features of the SAN router

Feature	Description
Intelligent ports	Four intelligent ports can be configured for iSCSI or iFCP.
iFCP standards track protocols	The SAN router supports the Internet Engineering Task Force (IETF) draft standard for iFCP, which provides connectivity and networking for existing Fibre Channel devices over a TCP/IP network.
iSCSI	The SAN router is capable of providing iSCSI connectivity.
R_Port	Support for FC-SW2 standard E_Port and Brocade interoperability mode allows you to fully integrate the SAN router into an existing Fibre Channel SAN that includes one or more Fibre Channel switches. In the McDATA Open mode, R_Port supports Cisco and Qlogic E_Ports as well.
Fast Write	The Fast Write software feature available on intelligent ports improves the performance of write operations between Fibre Channel initiators and targets in a wide area network (WAN). The improved speed depends on the WAN round-trip time (RTT), available buffer space on the target, number of concurrent I/Os supported by the application, and application I/O size.
Zoning	Using SANvergence Manager, network management software, or the CLI, you can create zones across networks. You can use zone sets for periodic reallocation of network resources. For example, you can have one set of zones for daytime data transactions and another set of zones for nighttime backups. You can create zones across networks.
Real-time and historical system logs	The Element Manager and LogViewer can be used to look at current system log messages from the connected SAN router.
Compression	Compression technology that is available on intelligent ports identifies repetitive patterns in a data stream and represents the same information in a more compact and efficient manner. By compressing the data stream, more data can be sent across the network even if slower link speeds are used.
Jumbo Frames	Since the maximum Fibre Channel payload size is 2112 bytes, two regular Ethernet frames are required. The Jumbo Frame option extends the Ethernet payload to 2112 bytes. With the support of Jumbo Frames, a Fibre Channel frame can be mapped to one Ethernet frame, providing more efficient transport. For iSCSI traffic, up to 4K size frames are supported.

Element Manager overview

The SAN Router Element Manager, a Web-based Java applet, is used to configure, monitor, and troubleshoot the router. Table 10-2 summarizes the configuration and monitoring functions of the Element Manager software.

Table 10-2 SAN Router Element Manager

Feature	Description
SAN Router Configuration	SAN Router Inband IP Address Date-Time System Properties Default Zoning Behavior Password Management SNMP Traps
Port Configuration	Fibre Channel and TCP Ports (supporting iSCSI and iFCP) Management Port Static Routing
iFCP Gateway Configuration	iFCP Setup Remote Connection Configuration Port Redundancy Configuration
iSCSI Configuration	Device Configuration RADIUS Server Configuration
SAN Router Operations	System Log Upgrade Firmware Reset the System Configuration Backup, and Restore
Monitoring	Device View LEDs and icons, system information icons Message Log Setting Polling Interval
Reports and Statistics	Address Resolution Protocol (ARP) Table Gigabit Ethernet Port Statistics Fibre Channel Port Statistics Fibre Channel Device Properties MAC Forwarding Table Storage Name Server (mSNS) Internet Protocol Forwarding Table Remote Gateway Statistics GraphicsPort Traffic Statistics Ping iFCP Compression Rate Statistics VLAN Configuration Statistics

10.1.2 IBM TotalStorage SAN16M-R

The IBM TotalStorage SAN16M-R (2027-R16) is the McDATA Eclipse 2640 SAN Router. It provides:

- ▶ The 16-port base SAN router in a 1U rack space
- ▶ The standard edition of SANvergence Management software
- ▶ Rack-mount kit
- ▶ Fully populated 2-Gbps shortwave SFPs on all ports

Twelve ports are user configurable as a 1 Gbps or 2 Gbps Fibre Channel or as a Gigabit Ethernet. The remaining four ports are intelligent Gigabit Ethernet ports, which support optional extended distance iFCP connectivity when activated.

Base functionality of the 12 user configurable ports provides SAN routing on up to two Fibre Channel ports, Fibre Channel fabric support, and iSCSI support on Gigabit Ethernet ports.

Clients may order three optional firmware versions: iFCP with Fast Write and compression on the four intelligent Gigabit Ethernet ports, SAN routing on any of the 12 user configurable ports, and comprehensive bundle with full iFCP and SAN routing capability. For enhanced management of the SAN router, clients may order the Enterprise SANvergence Management Software.

SAN router features

The SAN router supports iSCSI, iFCP, and R_Port for trunking to both IP backbones and existing Fibre Channel fabrics. The SAN router connects to a wide range of end systems including Fibre Channel, and Fibre Channel initiators and targets. SAN routers support TCP/IP routing over extended distances at wire speed. The SAN router offers:

- ▶ SAN internetworking for scalable and fault-tolerant SANs
- ▶ Support for full fabric, private, and public loop Fibre Channel devices
- ▶ Patent-pending Fast Write technology for maximizing throughput across long distances
- ▶ Compression for increased bandwidth

SAN router physical description

All ports and connectors are located on the front of the SAN router, except for the power connectors. The rear of the SAN router contains the power connectors and cooling fans. The FRUs are the optical transceivers and power supplies, which include internal fans.

Two standard power connections are located on the rear of the SAN router. Each power connection supplies ac power to a different power supply for power redundancy and backup. Either power supply can support the SAN router operation, but we recommend that you connect both, each to a different power

source. If one power supply fails, the SAN router continues to operate, but you must replace the failed power supply immediately.

Fibre Channel ports

There are 12 user-configurable Fibre Channel ports labeled 1 through 12 located on the front of the SAN router. These port connections are SFP connectors that provide 1 Gbps or 2 Gbps Fibre Channel or 1 Gbps Gigabit Ethernet connectivity. These ports can be configured as:

- ▶ FC_Auto (default)
- ▶ FL_Port
- ▶ F_Port
- ▶ L_Port
- ▶ R_Port

To the left of each Fibre Channel port is an LED that indicates the configuration and status of the associated port.

Ethernet ports for IP connectivity

The SAN router provides four intelligent ports for Gigabit Ethernet connectivity, labeled 13 through 16. The red labeled ports are intelligent ports that can be configured for iFCP. The white labeled ports can be configured for mFCP.

Any intelligent port (red labeled) can be configured for iSCSI. Each port has 256 Mb in buffers: 96 transmit, 96 receive, and 64 for overhead processing.

Management ports

Two management ports are located on the front of the SAN router. An RS-232 serial port can be connected to a VT100 terminal emulator for access to the CLI. An RJ45 port can be connected to the LAN for out-of-band management using the SAN Router Element Manager or the SANvergence Manager. The RJ45 management port can be accessed by any PC on the LAN with a Web browser.

Operational features

Table 10-3 summarizes the SAN router features. Some features are optional and might not be present in some SAN router software versions.

Table 10-3 Features of the SAN router

Feature	Description
Intelligent ports	Four intelligent ports can be configured for iSCSI or iFCP.
iFCP standards track protocols	The SAN router supports the IETF draft standard for iFCP, which provides connectivity and networking for existing Fibre Channel devices over a TCP/IP network.
iSCSI	The SAN router is capable of providing iSCSI connectivity.
R_Port	Support for FC-SW2 standard E_Port and Brocade interoperability mode allows you to fully integrate the SAN router into an existing Fibre Channel SAN that includes one or more Fibre Channel switches. In the McDATA Open mode, R_Port supports Cisco and Qlogic E_Ports as well.
Fast Write	The Fast Write software feature available on intelligent ports improves the performance of write operations between Fibre Channel initiators and targets in a WAN. The improved speed depends on the WAN RTT, available buffer space on the target, number of concurrent I/Os supported by the application and application I/O size.
Zoning	Using SANvergence Manager, network management software, or the CLI, you can create zones across networks. You can use zone sets for periodic reallocation of network resources. For example, you can have one set of zones for daytime data transactions and another set of zones for nighttime backups. You can create zones across networks.
Real-time and historical system logs	The Element Manager and LogViewer can be used to look at current system log messages from the connected SAN router.
Compression	Compression technology available on intelligent ports identifies repetitive patterns in a data stream and represents the same information in a more compact and efficient manner. By compressing the data stream, more data can be sent across the network even if slower link speeds are used.
Jumbo Frames	Since the maximum Fibre Channel payload size is 2112 bytes, two regular Ethernet frames are required. The Jumbo Frame option extends the Ethernet payload to 2112 bytes. With the support of Jumbo Frames, a Fibre Channel frame can be mapped to one Ethernet frame, providing more efficient transport. For iSCSI traffic, up to 4K size frames are supported.

Element Manager overview

The SAN Router Element Manager, a Web-based Java applet, is used to configure, monitor, and troubleshoot the router. Table 10-4 lists the Element Manager software configuration and monitoring functions.

Table 10-4 SAN Router Element Manager

Feature	Description
SAN Router Configuration	SAN Router Inband IP Address Date-Time System Properties Default Zoning Behavior Password Management SNMP Traps
Port Configuration	Fibre Channel and TCP Ports (supporting iSCSI and iFCP) Management Port Static Routing
iFCP Gateway Configuration	iFCP Setup Remote Connection Configuration Port Redundancy Configuration
iSCSI Configuration	Device Configuration RADIUS Server Configuration
SAN Router Operations	System Log Upgrade Firmware Reset the System Configuration Backup and Restore
Monitoring	Device View LEDs and icons, system information icons Message Log Setting Polling Interval
Reports and Statistics	Address Resolution Protocol (ARP) Table Gigabit Ethernet Port Statistics Fibre Channel Port Statistics Fibre Channel Device Properties MAC Forwarding Table Storage Name Server (mSNS) Internet Protocol Forwarding Table Remote Gateway Statistics GraphicsPort Traffic Statistics Ping iFCP Compression Rate Statistics VLAN Configuration Statistics

10.2 SAN router architecture

This section describes the basic functions, features, and internal architecture of the McDATA routers. It also explains basic terminology used in McDATA SAN routing.

10.2.1 SAN routing terminology

The introduction of SAN routing technology brought with it new jargon and terminology. Routing has caused us to define new terms so that we can describe routed SANs and their properties effectively. In addition to the standard SAN routing terms, each vendor uses different terms along with their products. Table 10-5 introduces some terms.

Table 10-5 *McDATA SAN routing terms*

Term	Definition
R_Port	The SAN routing port. It is used on McDATA SAN router side to identify a connection to a Fibre Channel switch. The opposite end to an R_Port on the Fibre Channel switch is the E_Port.
mSAN	Metro area SAN. This is the actual fabric, which is interconnected via SAN router to one or more other fabrics.
iSAN	Internetworked SAN. iSAN is a logical name which represents one or more mSANs (fabrics) interconnected via SAN routers across larger distance (outside the metro area).
IRL	Inter-router link. This is an IP-based connection between two SAN routers. It uses the iFCP.
iFCP	Internet Fibre Channel Protocol. This protocol is used to connect two or more mSANs together. Usually it uses the external, high-latency, lower-bandwidth networks outside the metro area.

Figure 10-1 shows an example of interconnected mSANs into an iSAN. From firmware release 4.7 and later, an mSAN can consist only of one SAN04M-R or SAN16M-R. These can be interconnected via one or more IRLs, using the iFCP. Inter-router iFCP connections provide path fail over capability. We recommend that you always use at least two connections for availability. All of the mSAN routers are interconnected together via an external WAN and come together to build an iSAN using the iFCP.

Figure 10-1 also shows an example of a typical routed SAN. It is a set of individual fabrics interconnected by SAN routers. A routed SAN functions as a single, large SAN, provides any-to-any connectivity, while keeping the particular

fabrics autonomous. Should any event occur within an individual fabric, such as component failure, fabric reconfiguration, Registered State Change Notification (RSCN) broadcast, such an event will not be propagated into the other fabrics.

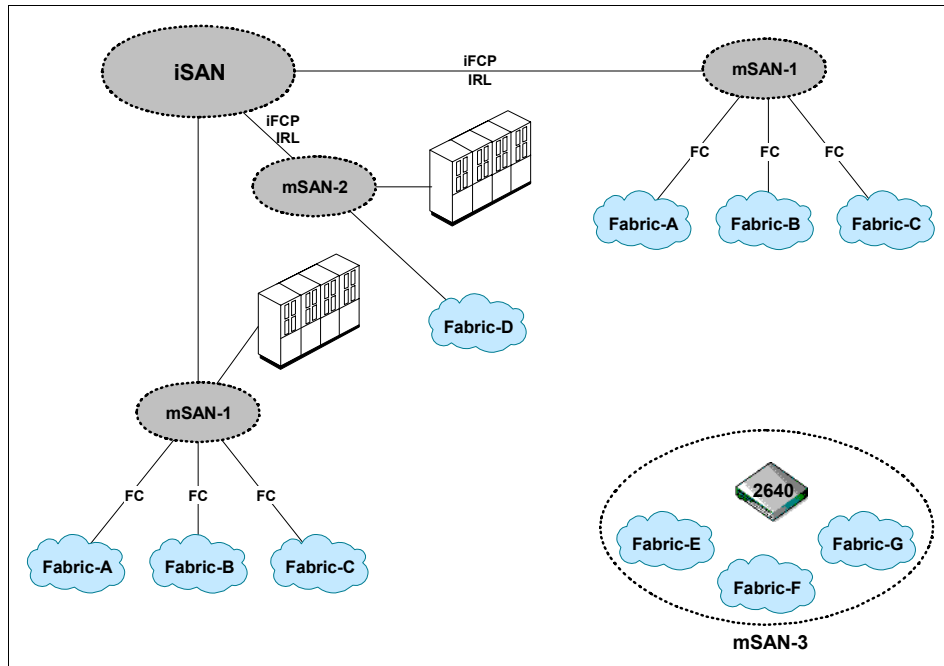


Figure 10-1 mSAN and iSAN interconnections

10.2.2 SAN routing features

This section describes the various features used in McDATA SAN routing.

mFCP and iFCP

mFCP and iFCP are used to interconnect McDATA routers. Even though these protocols are similar to each other from a high-level perspective, they are used in different environments and have different features.

The TCP stack of both iFCP and mFCP is governed by a Saturn processor. Each iFCP port has its own dedicated Saturn processor. Apart from the TCP stack, the Saturn processors accommodates Fast Write and Compression as well.

The mFCP was used to interconnect another SAN16M-R router in a metro area, by using a high-speed, low latency, usually dedicated LAN or VLAN. There is no mFCP support on the SAN04M-R router. Also, since firmware version 4.7, there is no support for mFCP connections *between* SAN16M-R.

The Gigabit Ethernet connection for mFCP must be full duplex, with symmetric flow 802.3X control. However, it does not use Fast Write, compression, and rate limiting. It is a McDATA proprietary, User Datagram Protocol (UDP) and currently requires a direct connection. mFCP uses Gigabit Ethernet ports only.

Both routers that use mFCP must be on the same subnet. Two SAN16M-R routers cannot be interconnected together with a Fibre Channel-based IRL. UDP by design does not support retransmission, packet reordering, or duplicate packets detection. Therefore, a fast and reliable connection is required. To prevent buffer overflow on each router's side, which would lead to the drop of packets, the SAN16M-R uses 802.3x Symmetric Flow Control.

In addition, up to four mFCP links can be aggregated into a single pipe by using 802.3ad Link Aggregation. Each packet uses Differentiated Service Code Point (DSCP) to ensure Quality of Services (QoS). DSCP is defined in RFC 2598 and defines Expedited Forwarding. In the McDATA mFCP implementation, Expedited Forwarding has set the "do not fragment" flag. The result is that packets are transmitted at the maximum possible maximum transmission unit (MTU) and do not fragment if a device with a lesser MTU set is in the path. Rather, the packet is dropped. Make sure you use the appropriate MTU size when planning to implement SAN16M-R routers with an mFCP link or links into your environment.

Tip: You can find RFC 2598 and related RFCs on the Web at:
<http://www.rfc-editor.org/>

iFCP is used in completely different environments, having lower bandwidth and high latency links over greater distances, usually across a WAN. The iFCP is a TCP-based protocol. Therefore a dedicated CPU is used to run the TCP stack within the SAN16M-R. Only the red labeled ports on the SAN16M-R have the dedicated CPU, and you cannot run iFCP on any other ports but these.

TCP itself provides a variety of services, such as packet retransmission, packet reordering, and duplicate packet detection. However, to transmit storage traffic effectively (to use the available bandwidth at a maximum possible sustainable rate) further optimization is implemented. These are the Fast Write algorithm, LZO compression, and rate limiting.

The 802.3ad Link Aggregation is not implemented in iFCP. One iFCP link can serve up to 64 TCP connections per single R_Port with the SAN04M-R or 256 TCP connections per single R_Port with the SAN16M-R. A TCP session is established as a result of PLOGI, PDISC, or ADISC in one of the iSAN-connected fabrics.

iFCP maps Fibre Channel frames to an IP datagram. One Fibre Channel frame is mapped per single IP datagram. Figure 10-2 shows the Fibre Channel to IP encapsulating mechanism.

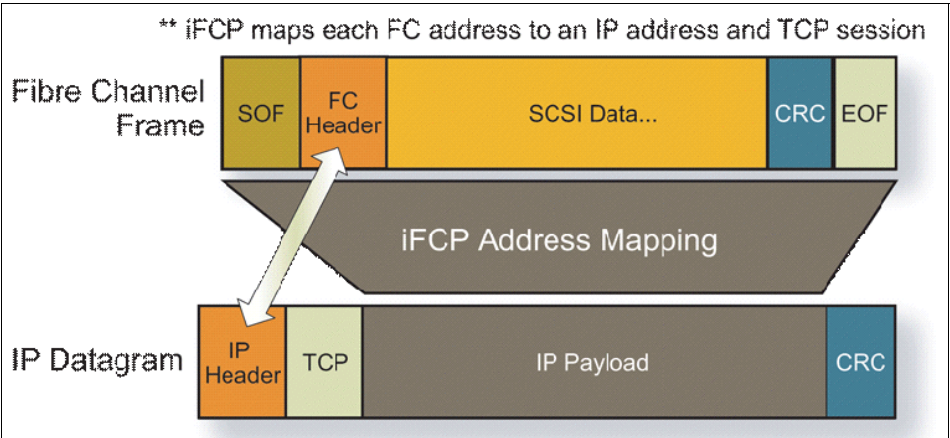


Figure 10-2 Fibre Channel Frame to IP Datagram encapsulation

Note: To compare different gateway protocols, such as iFCP and FCIP, refer to 1.2, “Gateway protocols” on page 4.

Table 10-6 summarizes the fundamental differences between mFCP and iFCP.

Table 10-6 mFCP and iFCP comparison

Technology	mFCP	iFCP
OSI Layer 4 Protocol	UDP	TCP
Intelligent ports with CPU only	No	Yes
Fast Write	No	Yes
Compression	No	Yes
Rate limiting	No	Yes
802.3ad Link Aggregation	Yes	No
802.3.x Flow Control	Yes	Yes
Must be on the same Inband IP subnet	Yes	No

The Inband IP address is the router’s internal SAN address, as shown in Figure 10-7 on page 169.

When a SCSI write operation is performed in a native Fibre Channel environment, multiple handshakes occur before a block of data is sent from a SCSI initiator to a SCSI target. First, the SCSI initiator sends a message with the total amount of data it wants to send. The SCSI target responds with a transfer ready message (FC_XFER_RDY) and indicates how much data it is prepared to receive. When the amount of data to send is settled between the SCSI initiator and target, the SCSI initiator sends the data. After a successful write, the SCSI target sends another FC_XFER_RDY and the handshake starts over again.

McDATA Fast Write helps mitigate the impact of the higher latency of IP networks over distance. When a write operation is started by an initiator, the local router forwards it to the remote router. That is the normal SCSI operation, which ensures that the commands are delivered to the target in the same order as initiated on the initiator.



The remote router, however, acts like a virtual target to the local initiator and sends back the FC_XFER_RDY message, prompting it to send the whole data segment for the write operation. The initiator sends the data and does not require any other handshake messages. From the remote target point of view, the remote router acts as an virtual initiator. FC_XFER_RDY messages are exchanged between this virtual initiator and remote target, so the handshake round-trip messages do not have to travel across the high-latency link. Figure 10-4 shows the flow of the SCSI write operation between remote sites using Fast Write.

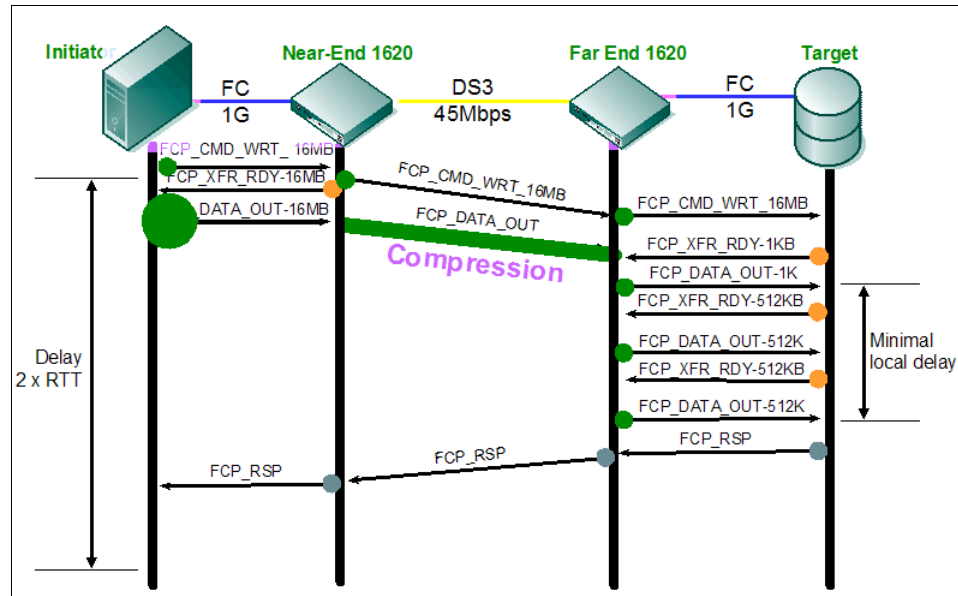


Figure 10-4 SCSI Write over high-latency environment with Fast Write

Fast Write tracks the status of all of its open Fibre Channel sessions at both the local and remote site. Any error condition on the target is detected by the initiator during the final SCSI completion message and leads to error recovery. Also, the Fibre Channel protocol has incorporated checksum mechanism in its design to ensure data integrity. The Fast Write does not interfere by the final SCSI completion message; this is sent at the end of the session to the real initiator. All of these aspects ensure data integrity, so there is no real danger of corrupt data when using Fast Write.

iSCSI gateway

The m-type routers enable communication between iSCSI initiators and Fibre Channel targets. Fibre Channel storage can be either directly attached to the router or can use standard R_Port connections from connected fabrics. iSCSI

initiators can be directly connected to the intelligent ports on m-type routers, or they can use an intermediate network.

The iSCSI protocol is designed to map the SCSI protocol over TCP/IP. Figure 10-5 shows SCSI to IP mapping.

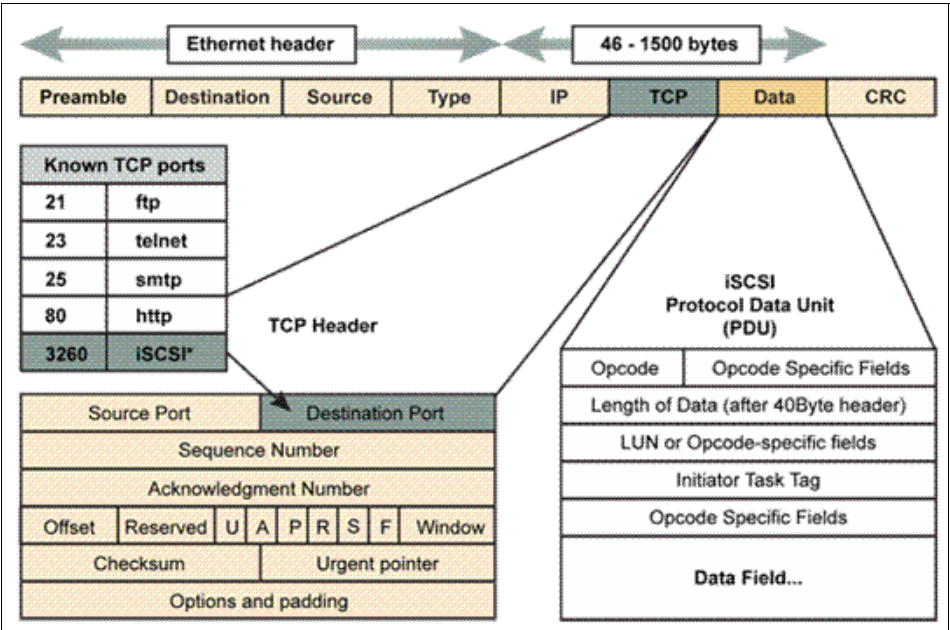


Figure 10-5 iSCSI protocol

In an iSCSI session, initiators establish iSCSI sessions with targets. Session IDs are generated to uniquely identify individual conversations between specific iSCSI nodes within the corresponding Network Entities. An initiator logging on to a target includes its iSCSI name and an initiator session ID (ISID). A target, responding to the login request, generates a unique target session ID (TSID), in combination with its iSCSI name. A single ISID/TSID session pair may have multiple TCP connections between them, as per the results of login negotiation. However, if multiple TCP connections for that session have been established, individual command and response pairs must flow over the same TCP connection. This is known as *connection allegiance*.

Figure 10-6 shows how you can use an m-type router to connect iSCSI hosts to your existing Fibre Channel environment.

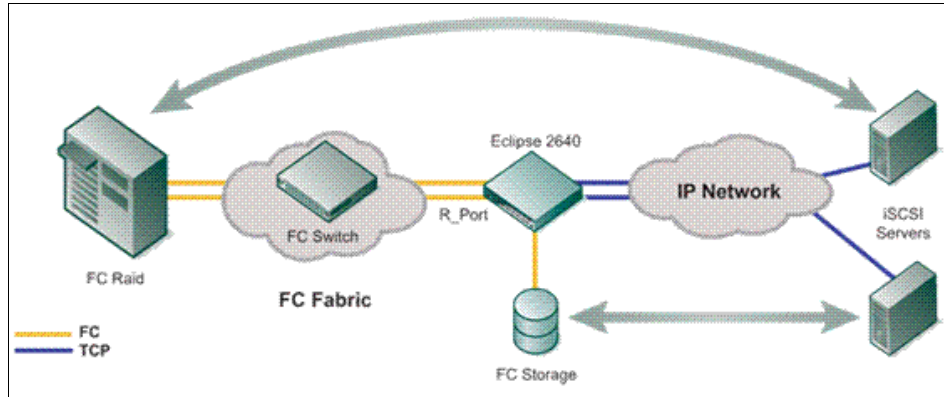


Figure 10-6 Connecting iSCSI servers to an existing fabric using m-type router

Selective ACKnowledgement

TCP Selective ACKnowledgment (SACK) is defined by RFC 2018. On a “lossy” network, when using SACK, the receiving host informs the sender that the data has been received. That means the receiving host can acknowledge packets out of order. The sender then retransmits only those packets that have been lost. For example, if receiving hosts acknowledges packets 1, 2, 3, 4, 6, and 8, then only packets 5 and 7 are retransmitted.

Without SACK, the sender has to retransmit all packets *after* the first missing packet, or in this example, packets 5, 6, 7, and 8.

Compression

Another optimization implemented in the McDATA SAN routers (both SAN04M-R and SAN16M-R) is compression. The LZO algorithm is implemented in the McDATA routers and runs on a dedicated Saturn processor on each iFCP port. It is not done on the shared CPU of the Eclipse SAN router itself. You can choose from four different compression modes.

► LZO

This is a frame-based algorithm. If you have many active initiator-target sessions opened on the iFCP link in your environment, this method works best.

► Fast LZO with history

This mode provides a 2 byte compression at a time with keeping the 8 byte history. This uses more memory, but offers a higher compression ratio. We recommend that you use this mode in an environment with a faster link (at least T3 and faster) and with a fewer active initiator-target sessions.

- ▶ LZO with history

This mode Provides 1 byte of compression at a time with keeping the 8 byte history. It offers a higher compression ratio at the expense of speed. Therefore we recommend that you use this mode in an environment with fewer initiator-target sessions and slower link (such as T3 and E3).

- ▶ Deflate

This mode is suited to provide the best compression ratio compared to other modes, but at the cost of speed. Therefore, we recommend that you use this mode on slower lines, such as 10 Mbps Ethernet and slower.

Rate limiting

Rate limiting prevents ingress traffic from entering faster than egress traffic, which results in buffer overflow and dropping packets. Dropped packets cause TCP to resort to flow-control which lead to a throughput decrease.

Interoperability mode

McDATA supports OEM interoperability through the use of McDATA Open Fabric (Interop-mode). All m-type family SAN routers have set the Open Fabric Mode as the default mode. With Open Fabric Mode, worldwide name (WWN) zoning is available, and port zoning is not. Features which are implemented differently by each vendor might be unavailable too. McDATA and Brocade interoperability is not supported by IBM except by an RPQ.

10.2.3 SAN routing architecture

We can differentiate between routing at an mSAN level, which can be done within a particular router, and routing at an iSAN level, which is done among routers interconnected via an iFCP link. Routers do not need to be on the same subnet.

To provide routing services, m-type family routers use the following internal network architecture:

- ▶ Router internal IP address
- ▶ SAN internal IP address for each port
- ▶ External IP address for each non-Fibre Channel port

Table 10-7 summarizes the default IP address settings for SAN04M-R and SAN16M-R routers.

Table 10-7 Default IP address settings for SAN04M-R and SAN16M-R routers

Description	SAN04M-R default	SAN16M-R default
Router internal IP address	192.168.111.100	0.0.0.0
SAN internal IP addresses	192.168.111.103 192.168.111.104	0.0.0.0
Subnet mask	255.255.255.0	0.0.0.0
External IP addresses	0.0.0.0	0.0.0.0
Subnet mask	0.0.0.0	0.0.0.0
Default gateway	0.0.0.0	0.0.0.0
Management IP address subnet mask	192.168.100.100 255.255.255.0	192.168.100.100 255.255.255.0

Figure 10-7 shows a diagram of the router internal network architecture.

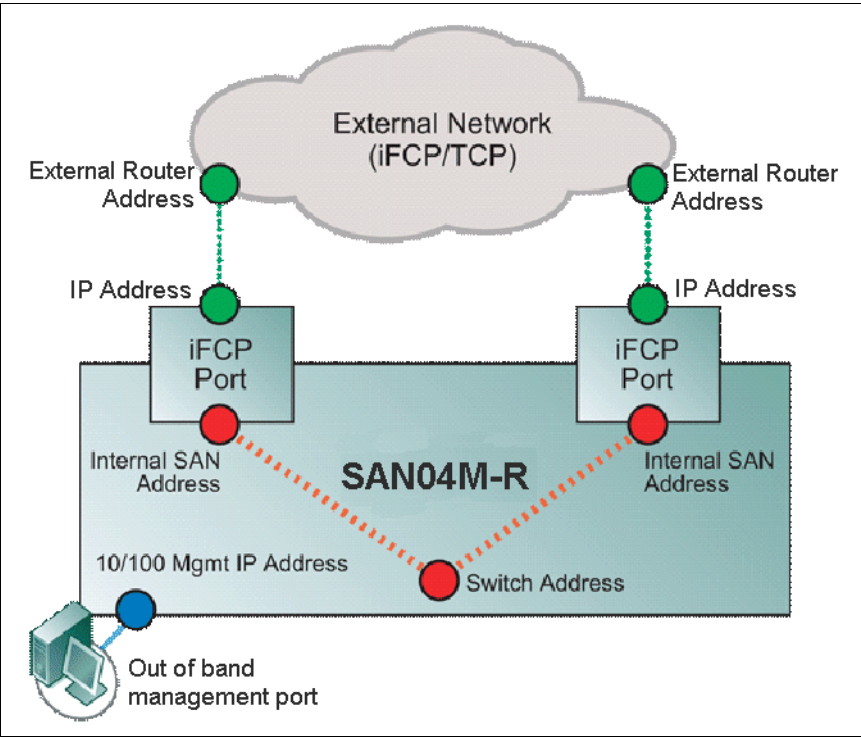


Figure 10-7 SAN router internal network architecture

Routing at the fabric level

Even though routing at the fabric level is not done by SAN routers, we cover it here briefly for the sake of completeness. At the fabric level, switches are interconnected via E_Ports. Path selection (routing) is governed by Fabric Shortest Path First (FSPF) based on calculating the cost of the particular path. The simple name server (SNS) service is used to register Fibre Channel nodes. All switches keep the SNS updates and are, therefore, all aware of the devices in the SNS.

Routing at the mSAN level

At the mSAN level, fabrics are interconnected together, but not merged. The interconnection is done using one or more SAN routers. There can be no more than two SAN16M-R routers in a single fabric and only one SAN04M-R in a single fabric. SAN16M-R routers can be connected together with up to four IRL mSAN links.

Fabric manager

Routing at the mSAN level uses each router's R_Ports. One of the R_Ports connected to the particular fabric is in charge of controlling the routing to and from that particular fabric. This port is called the *fabric manager*. The fabric manager is selected during the fabric build only, and it is always that port with the lowest node worldwide name (nWWN). The fabric manager is selected during fabric build when all nWWNs are sent across the fabric. Only the R_Port nWWNs are eligible to become a fabric manager. To distinguish the R_Ports from other ports in Fabric, especially E_Ports, the following method is used.

Each fabric component vendor has assigned to it a particular address range, the Organizational Unique Identifier (OUI). Within this pool of addresses, McDATA reserved a subset of addresses for routers only. When nWWNs are sent across the fabric build, the m-type family routers recognize the R_Ports and assign fabric manager R_Port to that port with the lowest nWWN. After the fabric manager is selected, it cannot be changed. It remains the fabric manager until another fabric build occurs. The SAN router cannot start a fabric build; it can only be initiated by a switch within the particular fabric.

Name server in mSANs

In an mSAN, two name servers are used: the primary Metro Simple Name Server (primary mSNS) and the secondary Metro Simple Name Server (secondary mSNS). The primary mSNS keeps a database of all the WWN nodes from that particular mSAN and other mSANs. It propagates this database to the entire iSAN (other mSANs' primary SNSs). The database of the primary mSNS is fed with data from each fabric's SNS. The primary mSNS uses the fabric manager for communication. The secondary mSNS serves as a client to the primary mSNS.

Its database contains only those WWN nodes entries, which are locally connected to the router.

Both mSNSs must be on the same inband IP address subnet. Otherwise they will not be able to communicate with each other. The inband IP address is the internal IP address of the router, as shown in Figure 10-7 on page 169.

The primary mSNS is selected automatically by fabric build, or it can be set manually. Primary and secondary mSNSs are synchronized by the time of primary mSNS selection only, which occurs during fabric build. After that, updates are sent via RSCNs. RSCNs are sent by using subnet broadcasts, for example to all devices on 10.0.0.255 and 255.255.255.0. Universal broadcasts, such as 255.255.255.255, are not used. If there is a change in the database in the secondary mSNS, a unicast packet is sent to the primary mSNS. The primary mSNS then sends a broadcast to all listening m-type routers.

In the mSNS database, the following information is stored:

- ▶ All storage entities in a local IP network, such as storage devices, Fibre Channel hosts, iSCSI initiators, and m-type routers
- ▶ WWN addresses for each Fibre Channel node
- ▶ Type of protocol and its properties for each registered entity
- ▶ iSCSI initiator name

The following services are provided by an mSNS:

- ▶ mSNS registration service
- ▶ mSNS State Change Notification service
- ▶ mSNS Keyed Query Service

The mSNS is responsible for partitioning the mSAN into zones.

Zoning

Zoning at the mSAN level is governed by mSNS. It provides a similar functionality to Fibre Channel zoning. Only devices which are members of a particular zone are allowed to communicate with each other. To share devices across an mSAN and between two SAN16M-R routers, you need to go through these steps:

1. Create mSAN zones with unique mSAN zone IDs (range from 1 through 512) on both routers. Zone names can be different unless they have the same mSAN zone ID.
2. If necessary, set a maximum bandwidth limit for your zones.
3. Add local devices into each zone on both routers.

Now zone members are mutually shared across the mSAN.

Routing domains and Fibre Channel Network Translation

The fabrics connected together by a router are kept separate. They do not merge together into one large fabric, and the Class F traffic is not allowed to pass through the router. Considering this, it is necessary to have a mechanism to enable traffic among those fabrics, the Fibre Channel Network Translation mechanism.

To allow cross-fabric communication, a device from one fabric must be presented with a unique fabric ID (domain ID) in the remote fabric. The problem is that usually the address space (fabric domain IDs) is reused in fabrics and, therefore, is not unique in mSANs. Two domain IDs, known as *routing domains*, are reserved for the purpose of unique addressing among fabrics. These reserved routing domains are:

- ▶ 0x7E for routing among fabrics within the mSAN boundaries
- ▶ 0x7F for routing among mSANs and for devices directly connected to one of the Fibre Channel ports on the router

If a remote nWWN node wants to communicate with another nWWN node within a local fabric, its fabric ID will always start with 0x7E.

Next, to locate the fabric that particular node comes from, four area IDs are available per single fabric and mapped to the routing domain. Four area IDs provide for addressing up to 1024 devices (4 x 256).

Finally, each fabric's R_Port has assigned to it the fabric ID of that particular fabric. This must be configured manually from the router's configuration interface.

Table 10-8 shows the mapping of the area IDs to fabric IDs for domain 0x7E.

Table 10-8 Area ID to fabric ID mapping for routing domain

Routing domain ID	Area ID	Fabric ID
0x7E	1 - 4	1
	5 - 8	2
	9 - 12	3
	13 - 16	4
	17 - 20	5
	21 - 24	6

Each egress traffic from the particular fabric undergoes network address translation (NAT) into 0x7E. Its fabric ID is mapped to a corresponding area ID range.

For example, if an initiator wants to communicate with a target in another fabric, the following actions occur.

1. The initiator queries the primary mSNS to see if the target is within the same zone and what its fabric ID is.
2. Since the target is in another fabric, the initiator sends a Fibre Channel frame to the router's routing domain 0x7E.
3. The router then performs the NAT back on the original domain of the target fabric and forwards the frame (received from the initiator) to the appropriate R_Port connected to the fabric where the target resides.

Note: The format as to how the routing domain IDs are propagated to each fabric depends on the mode in which the fabric operates. In McDATA's Open Fabric Mode, routing domains are 0x7E and 0x7F. However in both McDATA's and Brocade's native mode, the routing domains are presented as 30 and 31 respectively.

Path selection

Two aspects influence the selection of the R_Port when mSAN routing occurs.

- ▶ Path cost is considered using the standard FSPF mechanism.
- ▶ If more than one destination with the same FSPF cost exists, one path is selected using a round-robin algorithm.

However, traffic from one initiator is always sent through the same R_Port, unless the following actions occur.

- ▶ The target is disconnected from the fabric.
- ▶ R_Port reset occurs.
- ▶ Fabric is rebuilt.

Only the traffic from multiple devices is subjected to round-robin among equal cost R_Ports (if there are any), but not from one particular device. However, SAN routers do not have control over the interswitch link (ISL) selection between the switch and director and a SAN router.

Routing at the iSAN level

The iSAN consists of two or more mSANs interconnected with an existing WAN link, enabling the sharing of devices among mSANs over greater distances. The WAN infrastructure consists of existing IP switches and routers, as shown in Figure 10-8.

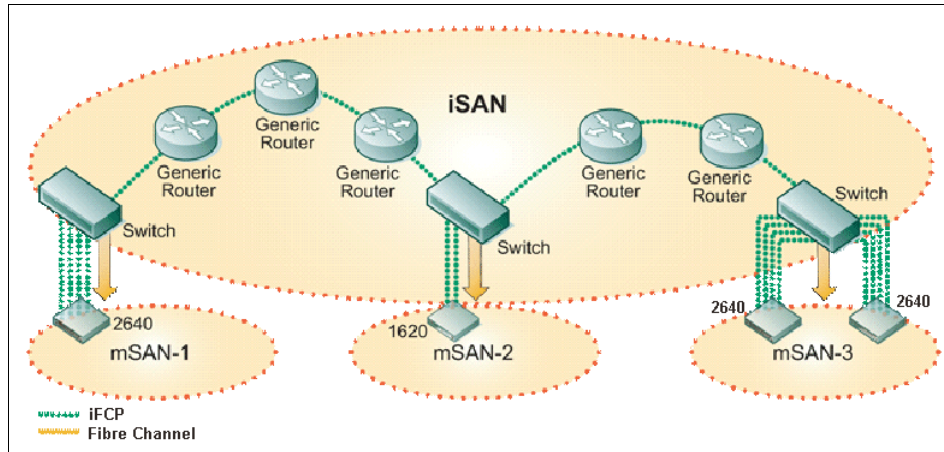


Figure 10-8 An example of iSAN architecture

From the iSAN perspective, we can observe that the following is taking place:

- ▶ Routing at the fabric level, fabric level zoning, and the primary mSNS in control
- ▶ Routing at the mSAN level, mSAN level zoning, both primary and secondary mSNS in control
- ▶ Routing at the iSAN level, iSAN zoning uses the same mechanism as mSAN zoning, primary mSNS in control

Path selection is done using FSPF. Note that the cost behind a router (everything presented as routing domain 0x7E) is not advertised at the mSAN or iSAN level. This means that equal cost may be seen from the FSPF perspective, not taking into consideration any other hops behind a router. As shown in Figure 10-9, servers from Fabric 1 see both iSAN and mSAN storage on the right side of the picture as directly attached to the router (presented as 0x7E), without considering FSPF within the fabric.

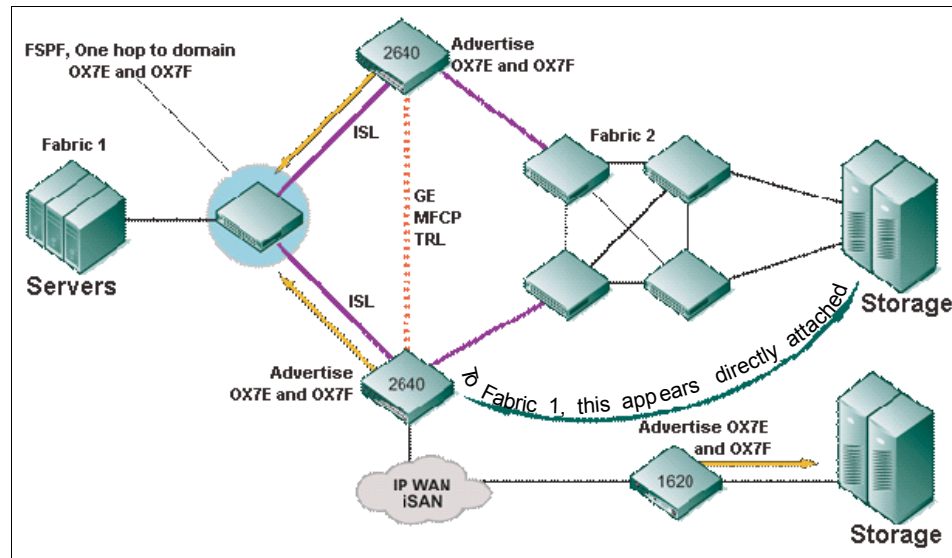


Figure 10-9 Additional FSPF costs behind the 0x7E domain



IBM TotalStorage m-type family solutions

Routers provide access to data which is located in a different fabric. The principal uses for storage routing are:

- ▶ Storage area network (SAN) extension over IP networks
- ▶ Lowering connection costs using Small Computer System Interface over IP (iSCSI)
- ▶ Achieving isolation and interoperability between different business units
- ▶ Managing scalability as your SAN environment grows
- ▶ Migrating from an older storage environment to a newer one

There are two IBM m-type OEM products to accommodate these features: the SAN04M-R and the SAN16M-R. The former one is especially suitable for iSCSI consolidation, simple SAN extension and migration. The latter one fits well for SAN extension and scalability and fabric isolation solutions.

11.1 SAN fabric local FC-FC routing

Figure 11-1 shows the local Fibre Channel (FC)-FC routing solution.

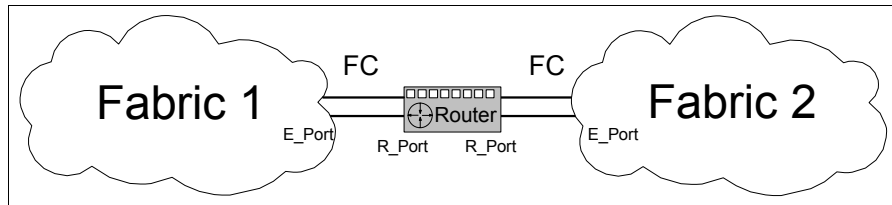


Figure 11-1 Local FC-FC routing between two SAN fabrics

In this case, one SAN16M-R is connected to different SAN fabrics using redundant Fibre Channel connections. From each fabric's switch point of view, the connection to the router appears as a standard E_Port, while on the router, the port is the R_Port. We can extend this configuration to span up to six SAN fabrics, each connected with two Fibre Channel connections.

If you have multiple switches in your fabric, we recommend that you distribute the connections to routers across them for maximum availability. If you are using a core-edge-fabric, we recommend that you connect routers to the core switches.

Some of the solutions that can be provided by local FC-FC routing include:

- Scalability

Fibre Channel addressing can theoretically support up to 16 million nodes in a single fabric. However, the practical limit for the amount of nodes is much lower. This is similar to TCP/IP networks, where the network address space is divided into smaller subnets of limited number of IP addresses, and traffic is routed between them. Usually the practical limit of a fabric from both technical and management standpoint is somewhere between 250 and 1000 nodes.

FC-FC routing allows you to divide the environment into several fabrics, while providing access to shared resources between fabrics.

- Multiple SAN administrators

In many cases, enterprises have several small SAN islands that are managed by different SAN administrators, often as a result of mergers or acquisitions. Using FC-FC routing between the fabrics allows each fabric to be managed separately from other fabrics. It also prevents the propagation of any management errors to other fabrics.

The mSAN configuration is the only part of the fabrics that needs to be coordinated among the SAN administrators. Since the mSAN zones need to

be defined on all fabrics before they can route traffic, the devices in each fabric are protected against unplanned access from the other fabrics.

- Interoperability between storage vendors

With FC-FC routing, you can separate fabrics built on different vendor's equipment to their own SAN fabrics, as shown in Figure 11-1. This helps to avoid problems with interoperability. This solution also allows each edge fabric to be supported and managed by the corresponding storage vendor, while enabling storage access between different SAN fabrics.

- Interoperability between old and new fabrics

In many cases when implementing a new SAN fabric, you already have an existing fabric. The existing fabric may have some parameter settings that you want or need to have set up differently in the new fabric. One good example is the Core PID setting.

By using FC-FC routing to connect the fabrics, you don't need to change the settings in the old fabric and are free to choose the settings you need for the new fabric. You can also use only a single Fabric OS level in any fabric, independent on the Fabric OS levels supported by the old hardware.

- Migration between old and new fabrics

The storage hardware is usually replaced with new hardware every three to five years. When refreshing the disk hardware, it may make sense to refresh the SAN hardware as well, especially if the new disk vendor is different than the former vendor.

FC-FC routing allows you to implement the new disk subsystems and SAN fabric in the final configuration. It also enables you to connect the complete new environment to the current SAN fabrics. This way you have simultaneous access from the servers to both old and new disk subsystems, and can use server-based tools, such as Logical Volume Manager (LVM), to migrate the data from the old disks to the new disks.

After you migrate any host to the new disks, you can move the Fibre Channel ports of the server to the new SAN fabric. Since you can do this one server at a time, the outage needed is minimized.

- Storage consolidation

Many enterprises implement a separate SAN fabric for tape backups. Without FC-FC routing, this requires a separate Fibre Channel adapter in each server that needs to be connected to the backup devices, as well as the additional fiber cabling to support these adapters. If you set up FC-FC routing between the normal SAN fabrics and the backup fabric, you can share the tape devices across any adapters in any fabric, as required.

Another example of storage consolidation is to implement a single IBM TotalStorage SAN Volume Controller (SVC) cluster across multiple SAN fabrics.

11.2 SAN extension with iFCP

Fibre Channel distances have been traditionally limited to either local fiber runs, using 9 micron long wave Fibre Channel, or high quality wide area networks (WANs) such as SONET and SDH in combination with coarse wavelength division multiplexing (CWDM) or dense wavelength division multiplexing (DWDM) multiplexers.

The advent of Internet Fibre Channel Protocol (iFCP) has meant that applications that can tolerate the high latencies of IP networks can now make Fibre Channel connections across standard corporate IP WANs. The advantages of this are that it uses a widely available and well understood infrastructure, which translates into lower cost.

We are still in a phase where people want iFCP over standard networks to be a panacea for all SAN extension applications. The inherent latencies involved are around 5 microseconds per km traveled in each direction with added latencies at every step (for example, up to 100 microseconds per router or firewall). This generally prevents iFCP from being used effectively for applications such as synchronous replication or online transaction processing (OLTP). For example, a high quality network of 1000 km might have a latency of around 20 milliseconds. Given that a disk I/O might only take 10 milliseconds itself, the problem with a 20 millisecond latency becomes obvious.

Because some corporate WANs provide uncertain Quality of Service (QoS), storage router vendors tend to be cautious about quoting distances for iFCP. They generally recommend that high quality WANs are necessary to provide services over anything more than 200 km or 300 km.

In practice, the most common uses for iFCP are asynchronous replication and non-critical access over campus or metro distances. A client may choose to implement Fibre Channel mapped into IP on a campus scale simply because the IP links are already in place. On a short IP network, the main problem becomes QoS since the latencies are not so large. The principles are the same whether running over 500 meters or 5000 km. All that varies is the link latency, the service reliability, and consistency.

Compression

iFCP compression in the m-type family routers increases the effective WAN bandwidth. While Gigabit Ethernet ports for IP Storage Services can theoretically achieve up to a thirty to one (30:1) compression ratio, typical ratios in the field are less than two to one (2:1).

The compression feature is implemented in both SAN04M-R and SAN16M-R routers and can be used with iFCP links. For more information regarding compression in m-type family routers, see “Compression” on page 167.

Using jumbo packets can also improve throughput. However, keep in mind that jumbo packets need to be turned on through the entire data path.

SAN extension over 700 km distance example

Figure 11-2 shows an example of an asynchronous replication running over IP at a 700 km distance.

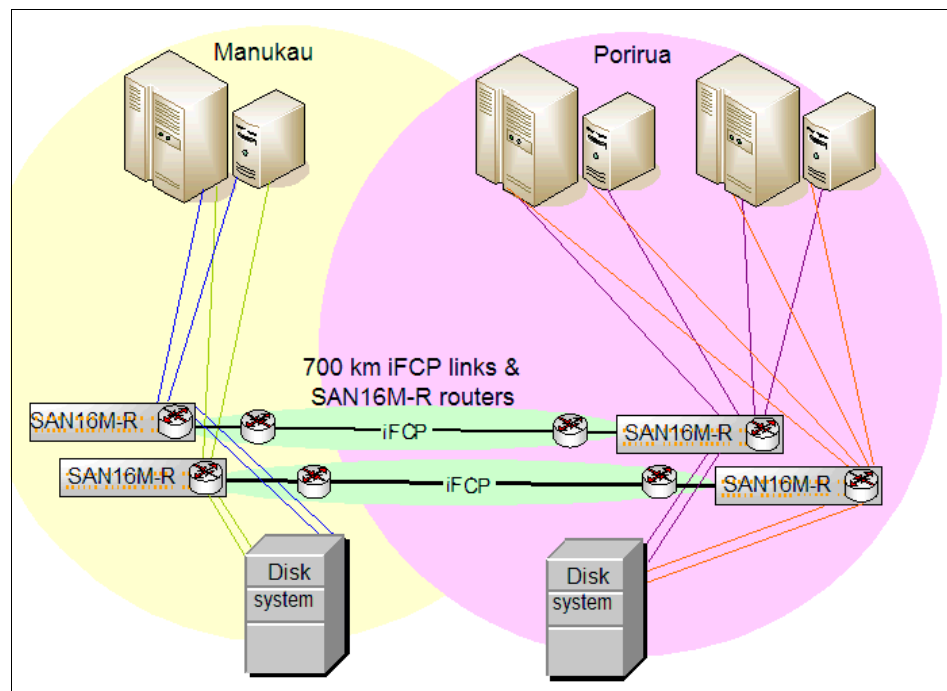


Figure 11-2 SAN extension over IP using iFCP over 700 km distance

Using iFCP Fast Write

The Fast Write mechanism is an attempt to mitigate the transport latency associated with long distance SCSI I/O operations. Fast Write performance depends heavily on the traffic profile of the SCSI operations being performed. The following I/O characteristics are well-suited to benefit from Fast Write:

- ▶ Long distance high latency network
- ▶ Write intensive I/O
- ▶ High number of small SCSI writes (rather than low number of large writes)
- ▶ Disk system has low write latency

These characteristics suggest that the best use for Fast Write is in disk system replication. However, some replication solutions, such as IBM Metro Mirror and IBM Global Mirror, already use mechanism similar to Fast Write.

Across a 100 km link, replication using Fast Write can be expected to deliver around 10% improvement in throughput and a similar percentage reduction in latency on a given FCIP network.

Learn the details about Fast Write in “Fast Write” on page 164.

11.3 Low-cost connection with iSCSI

There are three common ways to create low-cost connections to disk storage.

- ▶ Fibre Channel Arbitrated Loop (FC-AL)

Using FC-AL does not require a switch port for each server, because up to 126 devices may share a single port. However one Fibre Channel host bus adapter (HBA) is still required for each server.

- ▶ Network-attached storage (NAS) gateway

Using an NAS gateway, you need only provision Fibre Channel ports for the gateway device, rather than for each server. Also no Fibre Channel HBAs are required for the servers. Therefore, the primary costs are in the cost of the gateway itself, the cost of upgrading your Ethernet network to handle the increased traffic, and establishing a virtual LAN (VLAN) for this new traffic.

Note: Some block I/O applications cannot be accessed effectively through an NAS gateway.

► iSCSI

iSCSI can be thought of as an IP SAN. Using iSCSI, you do not need to provision Fibre Channel ports for each server. Also, no Fibre Channel HBAs are required, but iSCSI imposes a processing overhead on each server. In some cases, a high performance Ethernet card with a TCP/IP offload engine (TOE) function may be advisable. Again you need to look at the costs associated with upgrading your Ethernet network, such as setting up a VLAN. Because iSCSI delivers block I/O, it might be compatible with all applications.

M-type family routers have the capability for iSCSI. The SAN04M-R with standard firmware version can accommodate up to 12 iSCSI initiators. Optionally, clients may order optional firmware version, which enables 50 iSCSI initiators. The SAN16M-R can accommodate up to 50 iSCSI initiators per port and up to 200 per single router with the advanced firmware version.

Figure 11-3 shows how you can use iSCSI to provision disk storage to non-critical servers. iSCSI can also be used for critical servers. However in general, you can expect lower performance and lower reliability on an Ethernet network than on a Fibre Channel network. Using iSCSI for critical servers should be done using iSCSI multi-pathing.

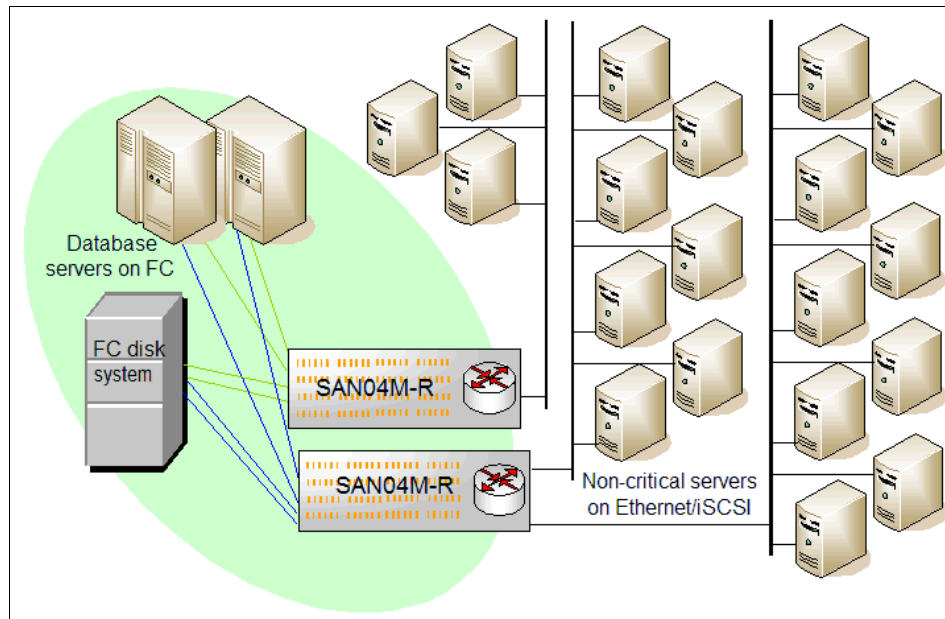


Figure 11-3 Using iSCSI routing to provision disk storage to non-critical servers

11.4 Isolation and interoperability using SAN routing

This section shows some examples of ways in which you might use SAN routing.

11.4.1 Separating production from development

As well as your main production environment, you may have a development or test environment which is subject to frequent reconfiguration and rebooting. Or you might have such an environment that is subject to a higher risk of failure due to less rigorous change controls. You need to isolate this from your production systems, but test systems also need occasional access to data that is stored on the production disk systems.

Figure 11-4 shows how to achieve this fabric isolation using two SAN routers.

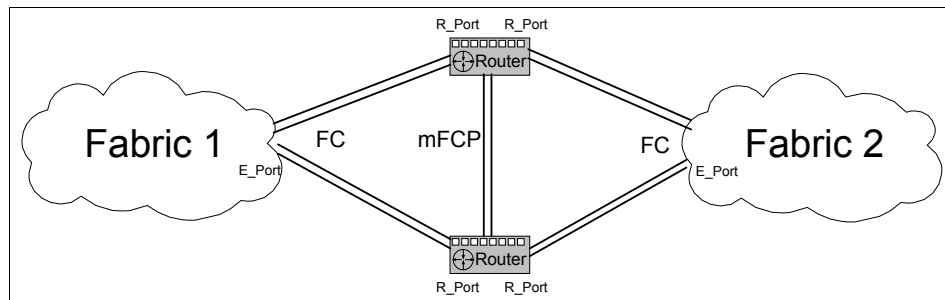


Figure 11-4 Separating fabrics using two routers

11.4.2 Separating corporate subsidiaries

A corporation can also choose to isolate subsidiary companies from each other while providing some shared services such as centralized backup. This approach can also be used by a shared-services provider to host multiple clients on the same physical infrastructure.

Figure 11-5 shows an example where separate subsidiaries share a physical infrastructure and allow shared access to the backup infrastructure.

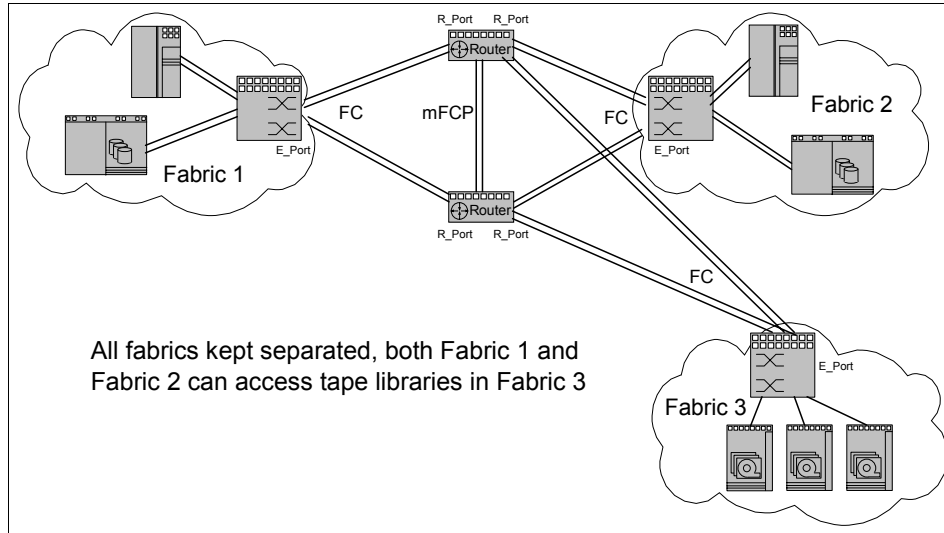


Figure 11-5 Using m-type SAN routers to isolate subsidiaries

11.4.3 Isolation of multivendor switches and modes

You may have Fibre Channel switches from multiple vendors that each require different mode settings and behave slightly differently in the network. You might want to incorporate them into your network, but keep them isolated either for departmental reasons or to keep the different modes of operation separate from each other. SAN routers gives the architect the confidence to combine switches from other vendors into the network, knowing that each SAN island has its own separate fabric services.

The Brocade fabric in this case would include initiator devices attached to the Brocade switch, the Brocade switch itself, and an inter-switch link (ISL) to the SAN router. The router provides and manages the routing between the Brocade and McDATA fabrics.

Figure 11-6 shows how switches from Brocade and McDATA can be incorporated into the network and yet be isolated when using SAN routers to interconnect them.

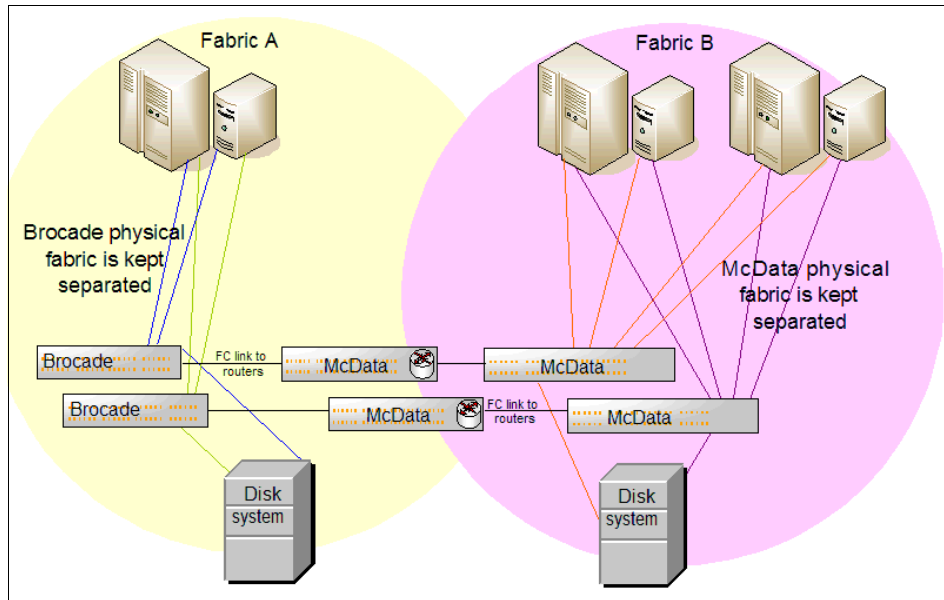


Figure 11-6 Using SAN routers to provide multivendor isolation and integration

11.5 Migration of existing storage to a new environment

You can use SAN routers to migrate data from your existing storage into a new environment. For example, assume that you have an HP XP512 storage system shared among AIX, HP-UX, and Windows servers. For historical reasons, each server platform has its own SAN fabrics and connections to the XP512. Each SAN fabric consists of a single 16-port, 1 Gbps switch.

The initial environment is shown in Figure 11-7. For the sake of clarity, we show only some of the servers.

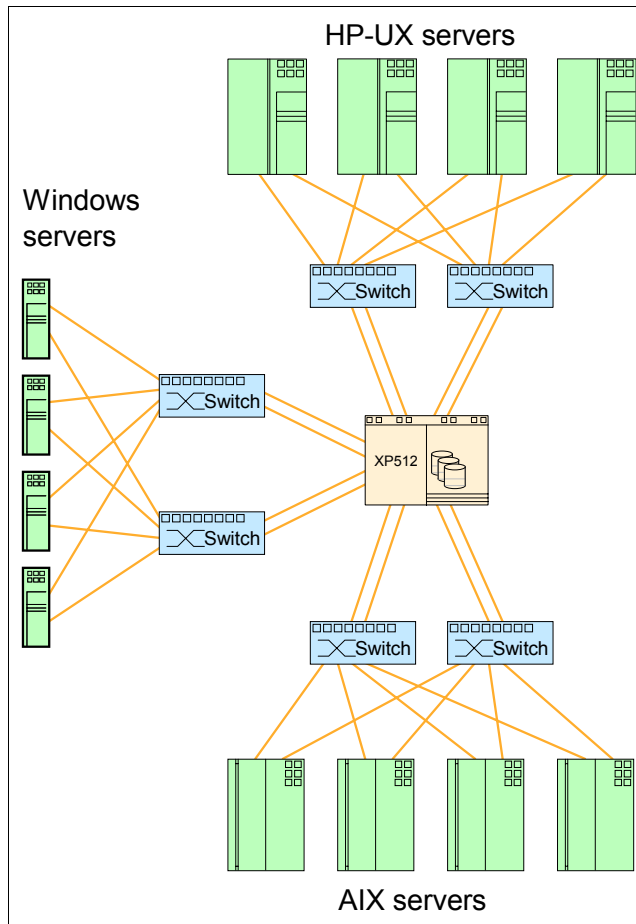


Figure 11-7 Initial storage environment

The newly implemented solution consists of following components:

- ▶ IBM TotalStorage DS8100 disk subsystem
- ▶ Two IBM TotalStorage SAN256M directors with 64 ports each
- ▶ Two IBM TotalStorage SAN16M-R routers

Figure 11-8 shows the interim solution.

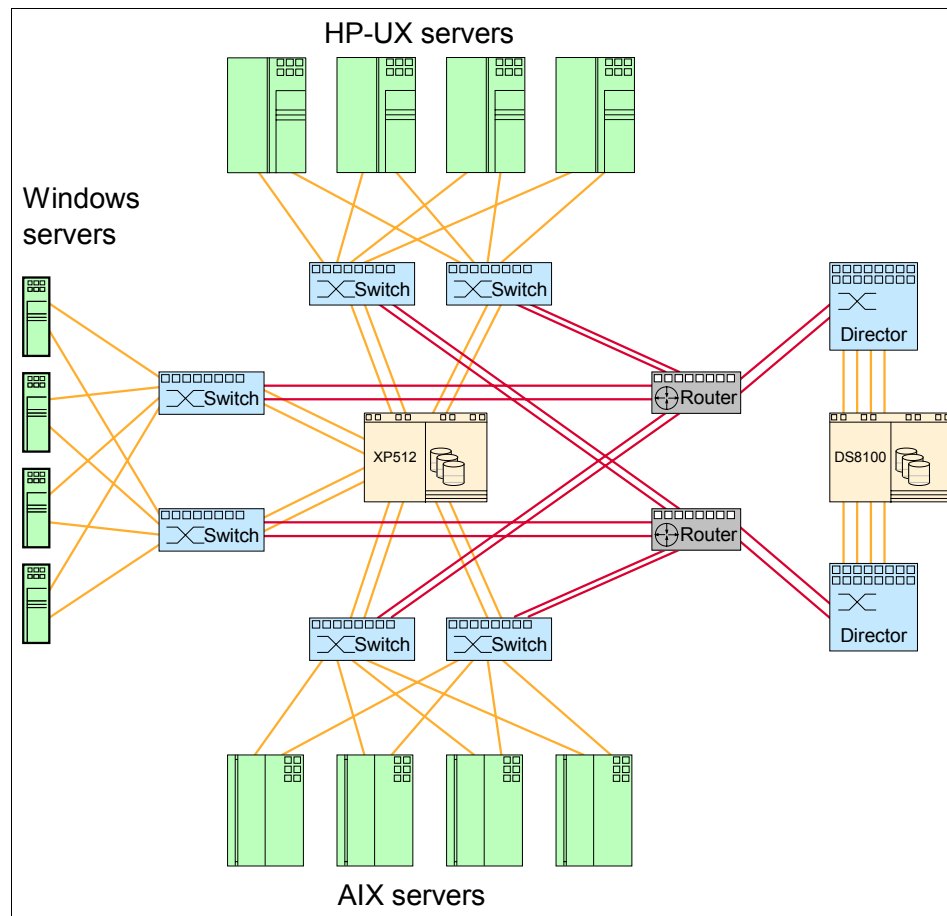


Figure 11-8 Migration interim storage environment

When the new environment is installed, we are ready to perform the migration. The migration typically includes the following steps:

1. Create mSANs for the server to access the DS8100.
2. Install the IBM Subsystem Device Driver (SDD) package and any other DS8100 specific software on the server.
3. Allocate new storage in the DS8100 to the servers.

4. Migrate all server data from the old storage to the new storage using the operating system based tools:
 - Native LVM for AIX
 - PVLinks for HP-UX
 - Veritas Volume Manager for Windows
5. Create zones to allow the server to access the storage from the new SAN fabrics.
6. Disconnect the server from the old switches and move it to the new directors.
7. Delete the mSAN zones created in step 1.
8. Disconnect the routers and old equipment.

After the migration is completed, we do not have any servers connected to the old switches and the XP512 is idle. At this time, we can remove the old storage hardware from the environment. The IBM TotalStorage SAN16M-R routers are also freed and can be used for other purposes, such as a SAN extension over iFCP.

Figure 11-9 shows the final solution after the migration.

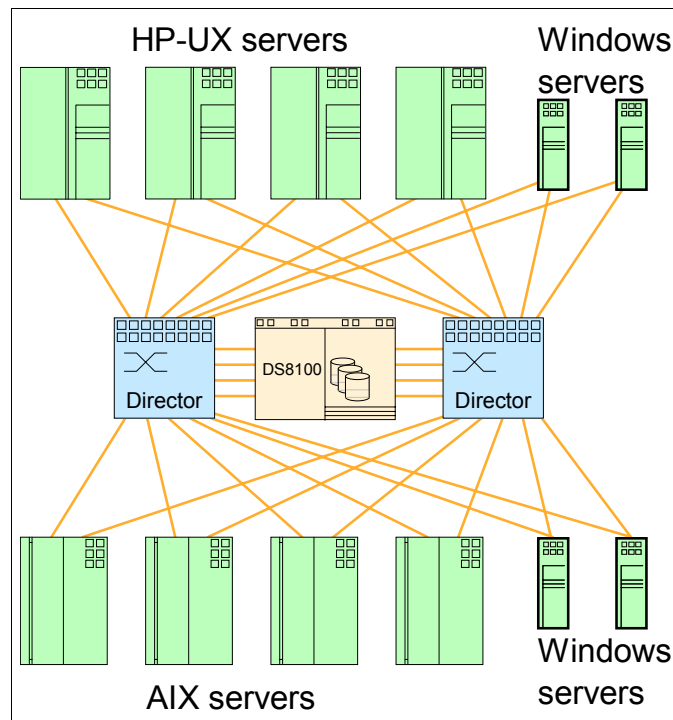


Figure 11-9 How the environment looks after the migration



IBM TotalStorage m-type family best practices

This chapter discusses various items for consideration that you need to plan before you introduce a storage area network (SAN) router into your SAN environment. This chapter does not present a comprehensive list, we think it offers a good starting point. Because each environment is different, there is no one list that provides answers to every question.

12.1 The planning checklist

When planning to introduce a SAN router to your environment, consider the following areas:

- ▶ Determine the amount of data you want to transfer among your interconnected fabrics. How much data of the total amount will change over a specific period of time? You need this information to size your link properly.
- ▶ Determine the wide area network (WAN) link type, its quality, and the number of independent paths of that link. Is it a shared or dedicated link? If it is shared, what other applications use this link? The best practice is to use dedicated links where possible; a shared link with a Quality of Service (QoS) mechanism is considered to be second best practice. Best-effort shared links are not recommended for storage traffic.

What is the maximum transmission unit (MTU) for your WAN? Do you use IP Security (IPSec)? Does your network support MTU auto-discovery? IPSec usually changes the MTU to a lower value, which can result in packet loss when your router's MTU setting is higher than that.

What is the average number of dropped packets? If this value is too large, the performance of the TCP-based communication is usually poor. The rule of thumb is that the network is considered to be well-performing if there is an average packet drop of one packet per million.

- ▶ If multiple WAN links are available, what are the differences among them? These include the bandwidth, latency, reliability, MTU size, and number of hops.
- ▶ Determine the type of security being used on your links. You need to open ports on your firewall, for example. We list all the ports an m-type router uses in 12.5.1, "Ports used by m-type SAN routers" on page 201.
- ▶ Do your network components support 802.3x flow control? The best practice is to use flow control to avoid packet drop due to buffer overflow.
- ▶ How many remote locations do you have? How many of them are within the mSAN area and how many are internetworked SANs (iSANs)? Do you have enough ports in your fabrics and IP infrastructure to accommodate routers and guarantee the aggregate bandwidth?
- ▶ Do you need to implement bandwidth management at the zone level? You can set up a zone with a minimum and maximum bandwidth that zone will use. This might be essential, especially on shared lines.
- ▶ What level of high availability is required? Will you need a backup WAN link? What will be the impact on the production systems if your primary link fails for a certain amount of time?

Decide how zones will be implemented. You can have your router append to your fabric's switches, or the zones may reside on routers only. If the latter is implemented, then every time you change a zone on the router level, you also need to change your zones at the switch level.

The best practice is to have your router append zones to fabric switches. All router zones have the SolP_ prefix, so none of your zones in particular fabrics will be modified.

- ▶ Create and maintain a matrix of all firmware and driver levels, operating system versions and maintenance levels. Consult with the vendors if there is a need for any upgrades prior installation.

12.1.1 The installation checklist

Consider the following items prior to a SAN router installation.

- ▶ Documentation

Have your list of IP addresses, port numbers, virtual LANs (VLANs), link types, fiber ports, protocols, equipment type and model, firmware, and driver version list ready and up to date.

- ▶ Cabling

For Gigabit Ethernet ports, you need cables with an LC connector on one end, the SAN router side, and an SC connector on the other, Ethernet switch side.

- ▶ Router management

Prepare a management workstation and connect it to the network that can reach the management ports of SAN routers. Reserve ports and IP addresses for connecting management ports. Set the router's management ports to Auto Negotiation (best practice). Make sure your firewalls (if there are any between SAN routers and the management application) allow Simple Network Management Protocol (SNMP) and Hypertext Transfer Protocol (HTTP) communication.

- ▶ Prepare a roll-back plan to return back to your starting point.

12.1.2 Running a pilot solution

Keep in mind that the primary task of your SAN is to serve your production servers. Therefore it is not a good idea to introduce new components directly into the live environment without prior testing.

Always follow the safe path and run a pilot solution. Your pilot solution should consist of all types of components that you have installed in your production environment. Develop the test cases and run them in your pilot environment. Try

any possible “what if” scenarios. The costs of downtime of a small test environment are incomparable to the costs of a downtime of your production SAN environment.

12.2 Fabric considerations

Before introducing any new component into your fabrics, it is a good practice to check the firmware levels of all the components and check with the product vendors to learn if there are any compatibility issues at given firmware levels. The m-type SAN routers can be interconnected together via mSAN or iSAN links only if they are on the same Enterprise Operating System (E/OS) level. Running different versions of E/OS is not supported, because they may not work together very well and cause problems.

Before upgrading the software on one of the interconnected routers, it is good practice to disable the mSAN, iSAN, or both ports. This will help to avoid undesired communication among the routers while they are not running the same level of code.

Do not use domain IDs Fast Ethernet and FF because the McDATA routers use these addresses for internal routing. Attaching a router to a Fibre Channel switch with these addresses can cause a fabric conflict.

Always make a backup of your router configuration before any router’s software upgrade. By doing so, you can avoid having to recreate your lost zoning configuration and setting vital parameters, such as interoperability mode, domain IDs, and the Core PID manually.

A good practice is not to fix anything that is not broken. You don’t need to upgrade your router’s Fabric OS just because there’s a new release available. Plan your upgrades carefully and perform them only when you have a good reason to do so.

12.3 Bandwidth and capacity planning

One of the main impacts on the final bandwidth requirements and capacity of the link is the type of application you will run across your interconnected fabrics.

Generally speaking, the highest and the most constant bandwidth requirements usually are ones that use synchronous data replication. Even higher bandwidth requirements can include a remote fabric tape library access. However in this case, the need for bandwidth usually occurs during the backup windows and tape management and vaulting tasks of your backup system. In the case of

asynchronous replication, the bandwidth requirements might be substantially lower.

12.3.1 Aspects that influence communication performance

This section discusses various aspects that have a significant influence on the link performance.

Link bandwidth

The link bandwidth is the most obvious factor that affects performance. It is also one of the key metrics used when provisioning the link. In storage environments, the link bandwidth used should always be the guaranteed bandwidth from the service provider.

If the guaranteed bandwidth is anything less than a full Gigabit Ethernet, it needs to be configured into the routers at both ends of the link to avoid overrunning the link. We recommend that you set the maximum allowed speed of the inter-router link (IRL) ports between 77% and 96% of the guaranteed bandwidth of the link at both ends of the link, depending on your link quality.

Latency and round trip time

Link latency is a metric of the round-trip time (RTT) it takes for a packet to cross the link. RTT is the time it takes for a datagram to be received and returned to the sender over a network. The key factors contributing to the link latency include:

- ▶ Distance
- ▶ Router and firewall latencies
- ▶ Time of frame in transit

Distance

The speed of light in optical fiber is approximately 208 000 km/s. Therefore the delay caused by a fiber connection is approximately 4.8 microns per km ($\mu\text{s}/\text{km}$). To calculate the round trip latency, we have to count this delay both ways.

For example, for a 100 km link, the round trip latency is approximately:

$$100 \text{ km} \times 4.8 \text{ } \mu\text{s}/\text{km} \times 2 = 960 \text{ } \mu\text{s}$$

Similarly, for a 1000 km link, the round trip latency is 9 600 μs , or 9.6 ms.

Router and firewall latencies

Any delay caused by routers and firewalls along the network connection needs to be added to the total latency. The latency varies a lot depending on the routers or firewalls and the traffic load. It can range from a few microseconds to several milliseconds.

You also need to remember that the traffic generally passes through the same routers both ways. Therefore, for round-trip latency, you need to count the one-way latency twice.

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

Time of frame in transit

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

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

Jumbo frames: Jumbo frames are a technique for maximizing the throughput of Ethernet networks by increasing the frame size from the default 1518 bytes up to 9000 bytes. To gain the maximum benefit, all devices in the network have to support the frame size. Non-jumbo routers break the frames down to 1518-byte frames.

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

For smaller frames, such as the Fibre Channel acknowledgement frames, the time in transit is much shorter. The minimum possible Fibre Channel frame is one with no payload. With Internet Fibre Channel Protocol (iFCP), the minimum size of a packet with only the headers is 96 bytes or 768 bits.

Table 12-1 details the transmission times of an iFCP packet over some common WAN link speeds.

Table 12-1 FCIP packet transmission times over different WAN links

Link type	Link speed	Large packet	Small packet
Gigabit Ethernet	1250 Mbps	14 μ s	0.6 μ s
OC-12	622.08 Mbps	28 μ s	1.2 μ s
OC-3	155.52 Mbps	113 μ s	4.7 μ s
T3	44.736 Mbps	394 μ s	16.5 μ s
E1	2.048 Mbps	8600 μ s	359 μ s
T1	1.544 Mbps	11 400 μ s	477 μ s

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

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

Congestion and dropped packets

An example of how congestion may occur is when more ingress ports are communicating to a single egress port. The buffers fill faster than they actually can drain out. The result is packet drop, which leads to retransmission.

12.3.2 Throughput and efficiency

Table 12-2 shows link speeds, approximate overhead, efficiency, and throughputs (with no overhead) for most common network link types. We use these parameters in our sizing examples in 12.3.3, “The amount of data and link sizing” on page 198.

Table 12-2 Example of throughput and efficiency of different network links

Link type	Link speed	Overhead	Efficiency	Throughput
100baseT Ethernet	125 Mbps	6.30%	93.70%	11.71 MBps
Gigabit Ethernet	1250 Mbps	6.30%	93.70%	117.13 MBps
OC-12 SONET	622.08 Mbps	8.92%	91.08%	70.82 MBps
OC-3 SONET	155.52 Mbps	8.92%	91.08%	141.64 MBps
T3	44.736 Mbps	5.05%	94.95%	5.31 MBps
T1	1.544 Mbps	4.42%	95.58%	0.184 MBps

12.3.3 The amount of data and link sizing

You can use different sizing methods and estimations depending on your environment and the application.

Link sizing for synchronous data replication

For synchronous data replication, you need to find the peak in write operations over a period of time. The theoretical maximum for a peak in write operations is influenced by many factors, such as central processing unit (CPU), application efficiency, networks, and storage devices.

For example, if we assume that the set of logical unit numbers (LUNs) that you need to replicate has a peak of 110 MBps during the system's busy hours. What speed of the WAN link is required for data replication?

The link speed of Gigabit Ethernet is 1000000000 bits per second. The link efficiency, under ideal conditions, is 93.7%. We can achieve a maximum throughput of 937 Mbps of storage data only, which is 117.13 MBps ($937/8=117.125$). We need only 110 MBps during peak periods for the data replication, so Gigabit Ethernet is the answer in our example.

However, you should monitor the utilization of your link and any possible increase in the write peaks as the environment grows over time.

Our calculation is simplified and does not take into consideration further link optimizations, such as compression or Fast Write. Compression usually leads to less overhead, so greater efficiency is achieved. However, using compression for synchronous replication cannot be considered as a best practice, because it introduces additional data processing, which introduces more latency. Latency then causes longer response times.

Here are other assumptions.

- ▶ There's a dedicated network (no shared bandwidth).
- ▶ We do not use Fast Write.
- ▶ Latency does not lead to congestion and dropped packets.
- ▶ We omit retransmits.

Link sizing for asynchronous data replication

When planning the link size for asynchronous data replication, we do not need bandwidth to accommodate data replication during peak hours. Instead, we need to find the total amount of data to be replicated and the amount of data to be changed in a period of time. While the former quantity is usually easy to determine, the latter is usually estimated as a best guess based on experience.

A good start is to assume that 20% of the total amount of data changes during the day. Your application's specific tools or operating system utilities may help you estimate this quantity.

For example, assume that we have 8 TB of data, with 20% changed over a 24-hour period. That is in average 69 GB change in one hour (20% from 8 TB = 1639; $1639 / 24 = 68.3$ GB per hour). We can calculate it with a conservative two to one (2:1) compression ratio, which is near 35 MB per hour ($68.3 / 2 = 34.15$).

The link speed of OC-12c is 622080000 bits per second. The link efficiency, under ideal conditions, is 91,08%. We can achieve maximum throughput of 567 Mbps of storage data only (without overhead), that is 70 MBps ($567/8=70.785$). We only need 35 MBps on average, so OC-12c is the right choice for our example. The amount of data, or the 24-hour data change ratio, can almost double and still provide enough bandwidth for asynchronous replication.

Link sizing for tape data backup

If planning for a link for remote tape communication, consider the total amount of data and the time period in which the data should be transferred from your hosts or backup servers to tapes. Another sizing factor might be the number of drives in the library. However, many of the new generation tape drives are capable of operating at data transfer rates able to consume the bandwidth of 1 Gbps Ethernet. For example, the IBM 3592 is capable of achieving a theoretical throughput of 120 MBps with compression, 40 MBps without compression. The conclusion is that you cannot feed, for example, two IBM 3592 drives with a 1 Gbps Ethernet link because this link gives you only a 117 MBps rate maximum. You need 240 MBps.

Let's consider an example where we have 10 TB of data to be transferred every weekend from primary data pools of our backup server to the remote tape library. The time for the data to be transferred cannot exceed 30 hours. We can calculate that:

$$10485760 \text{ MB} / 108000 \text{ seconds (30 hours)} = 97.1 \text{ MBps}$$

We need a Gigabit Ethernet link to accommodate this amount of data within the given period of time.

12.3.4 Fast Write and IBM products

Do not use Fast Write with IBM data replication products, such as Metro Mirror, Global Mirror, and Global Copy. These already include write optimization in the

product itself, so they would not benefit from McDATA's Fast Write on the router level.

IBM data replication products do not send round-trip messages (write command and transfer ready) to start sending data. Instead, they start to send the data immediately to the remote subsystem. This is similar to what the McDATA's Fast Write does. Therefore, the best practice is not to use Fast Write with IBM Metro Mirror, Global Mirror, and Global Copy.

12.4 Planning for availability

This section discusses some basic considerations to achieve a higher level of availability in your interconnected SANs using m-type routers. When planning your interconnected SAN for availability, always determine the level of availability that you really need. Calculate how much risk you can tolerate and what would be the impact on your production systems if you lost the fabric interconnectivity for any given period of time.

12.4.1 Hardware limitations

Routers, in general, are designed as a switch-class product, not a director-class product. Therefore a situation may occur when the router must be rebooted or replaced.

The m-type routers are not chassis-based, and the main electronic components are not redundant, except the power supplies and fans. In case of a power supply or fan failure, the router will remain operational, but a service window must be scheduled to replace it. In the case of any other component failure, the entire router must be replaced.

12.4.2 Multiple paths and path failover on a router level

To increase availability, we recommend that you use multiple paths to your fabrics and Metro Fibre Channel Protocol (mFCP) paths to a single router. One router can have multiple iFCP paths. However only one of them can be active at a time.

The backup iFCP path can be activated automatically by the router itself if the primary path fails. This is done by the heartbeat between two routers connected by an iFCP link. If the heartbeat is lost for a particular period of time (10 seconds), the backup path is activated. You can bring the primary path back online either manually or let the router do this automatically.

With the SAN16M-R router, you can design your mSAN for path failover. That means you can configure two routers in your mSAN, interconnect them via an

mFCP link, and connect each fabric's switch with two alternate paths, each to a different SAN16M-R router.

Figure 12-1 shows the path failover scenario at a router level. If one of the routers fails, all the traffic is transferred through the remaining router.

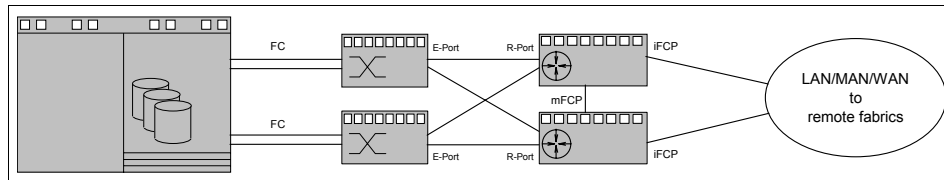


Figure 12-1 Path failover on router level.

It is a good practice to keep local storage traffic and traffic for replication separated. or rather to dedicate one pair of storage ports for data replication and the other pair for local traffic.

12.4.3 Fault isolation

If the WAN link in Figure 12-1 fails, fabric reconfiguration does not occur, because fabrics are not merged. Each fabric has its own Storage Name Server (SNS) database, domain space, and principal switch. No F-Class traffic is sent beyond an R_Port, and no Registered State Change Notifications (RSCNs) are propagated to remote fabrics.

12.5 Planning for security

With the introduction of SAN routers, the storage traffic leaves the known boundaries of Fibre Channel and traverses via IP networks to remote locations. This brings new security challenges and items to plan and consider before you can interconnect SANs.

12.5.1 Ports used by m-type SAN routers

For routers to communicate, many TCP and User Datagram Protocol (UDP) ports must be enabled as appropriate. In addition to this, some other ports must be enabled to manage routers. Table 12-3 lists all the ports that are being used by m-type router products.

To use virtual private network (VPN) with IPSec, we highly recommend that you secure communication among your m-type routers.

Table 12-3 Ports used by m-tpe routers

Description	Type	Port	Protocol
Inter-switch control	TCP	37121	iFCP
Redundancy control	TCP	37122	iFCP
Data	TCP	3420	iFCP
Data	TCP	3260	iSCSI
MTU Discovery	UDP	7	UDP Echo
SNMP Traps	UDP	162	SNMPv1
Telnet	TCP	23	Telnet
FTP	TCP	20	FTP control
FTP	TCP	21	FTP data
TFTP	UDP	-	TFTP
Java Applet	TCP	80	HTTP
Applet switch communication	UDP	161	SNMPv1
Switch message log communication	UDP	37009	
iFCP/mFCP ping results communication	UDP	37010	

12.5.2 Zoning

The Fibre Channel router ports need to be secured using zoning. You can zone either at pWWN or Node World Wide Name (nWWN) level. However pWWN zoning is not supported between mSANs that are interconnected by an iFCP connection.

One of the oldest best practices from the early SAN days is to have one initiator and one target in a single zone and not to have more than two ports in a single zone. This was a necessity in the past, but it is too conservative today. The host bus adapters (HBAs), their firmware, and OS drivers are much more mature than they were couple of years ago.

There is no single, best recommendation on how to use zoning in your environment. In general, we can organize the zones as follows:

- One initiator and one target per zone

This is not suitable for midrange and large environments. Imagine having to administer your zones in an environment of 50 servers, each with four HBAs

and two storage subsystems, each with six HBAs and two tape libraries with eight drives each. It would not be manageable to keep with this zoning approach.

- ▶ One initiator and multiple targets per zone

If you still want to stay conservative, but don't want to run into zone administration overhead, consider this rule for your zone configuration.

- ▶ Multiple initiators and multiple targets per zone

If your level of conservatism is not too high, and if you can take a calculated risk of possible "chatter" among your initiator HBAs within the zone, you may consider this approach.

Zoning at the mSAN or iSAN level should only be done for devices that you really need to have access to in remote fabrics. If there are only certain periods of time when you need to access a particular device in another mSAN, for example once a month, we recommend that you export (zone) this device only for that particular period of time that you need to have it exported.

12.6 Scalability and limitations

This section discusses some best practices and limitations that you need to consider when planning for interconnecting fabrics using SAN routers. Many of the numerical figures that we mention here are usually the technical limits of particular devices or practical recommendations. However, most of the SAN environments in the real world, at the time of writing, are within these limits.

There can be a:

- ▶ Maximum number of two SAN16M-R routers in an mSAN

There can be only one SAN04M-R in an mSAN, since it does not support an mFCP link. You cannot go beyond this limit.

- ▶ Minimum of two mFCP links

The mFCP links are configured in pairs. If you only activate one, you will receive an error message stating that mFCP will not be usable until its pair is also active. The pairs are 1-2, 3-4, 5-6, 7-8, 9-10, and 11-12.

- ▶ Maximum number of four mFCP connections between the SAN16M-R in an mSAN

We recommend that you use at least two mFCP connections (as per previous the point) for availability, regardless of performance requirements. You cannot go beyond this limit.

- ▶ Maximum number of four inter-switch links (ISLs) to each fabric from all mSAN routers

Since you can have a maximum of two SAN16M-R routers, we recommend that you have two ISLs from each router to the fabric. In a case where you only have one SAN16M-R router, it can have up to four ISLs to the fabric. A good practice is to always have at least two ISLs from one or both routers in an mSAN to each fabric.

- ▶ Maximum number of six fabrics can be connected in an mSAN

This applies whether you have one or two routers in the mSAN. The number of connected routers in an mSAN does not affect this limit.

- ▶ Maximum number of 48 switches in an mSAN

The number of connected routers in an mSAN does not affect this limit. Since you can have up to six fabrics, the number of switches in each fabric may vary, but it cannot exceed the total number of 48.

- ▶ Recommended average number of 12 switches per single fabric

You can have more than 12 switches in one fabric, but fewer than 12 switches in another fabric to keep the average around 12. For example, you can have 16 switches in one fabric and 8 in another fabric ($16 + 8 = 24$, $24 / 2 = 12$). You also need to keep in mind that there is a maximum number of 48 switches in an mSAN. You cannot exceed this limit.

- ▶ Maximum number of 1024 connected devices in a single fabric

Usually, this is the limit of the SNS database size in the fabric itself.

- ▶ Maximum number of 504 devices imported from one single fabric

The total number of imported devices from all fabrics per single SAN16M-R is 512. This is the limit of the mSNS database.

In addition to these limits, Table 12-4 shows the tested limits at the time of writing.

Table 12-4 Tested scalability limits at the time of writing

Metric	SAN04M-R	SAN16M-R
Maximum number of Fibre Channel Fabrics	2	6
Maximum number of switches (domains) per fabric	12	16
Total number of Fibre Channel switches in all interconnected fabrics in an mSAN	24	24
Combined maximum imported Fibre Channel devices from all fabrics with an mSAN	64	256

Metric	SAN04M-R	SAN16M-R
Maximum Fibre Channel devices in a connected fabric	1024	1024
Maximum number of R_Ports connected to a single fabric	2	4
Maximum number of SAN16M-R zones (recommended and tested, possible 512 and 1024 respectively)	128	256
Maximum number of loop devices off a single router FL_Port	8	32
Maximum number of loop devices attached to a single router	16	384
Maximum iFCP plus iSCSI sessions on a single Gigabit Ethernet-TCP port (initiator-target pairs)	64	64
Maximum iFCP plus iSCSI sessions (initiator-target pairs) per router	64	256
Maximum number of iSCSI initiators per router port	50	50
Maximum number of iSCSI initiators per Eclipse SAN router	50	200
Maximum number of iSCSI initiators in one mSAN (two routers, mFCP)	N/A	200
Maximum number of iSCSI sessions in one mSAN (two routers, mFCP)	N/A	256
Maximum iFCP point to multi-point connections per router port (one "site" to many "sites")	8	8
Maximum iFCP point-to-multipoint connections (one "site" to many "sites") per SAN router	16	32
Maximum number of Eclipse SAN routers in an mSAN	1	2
Maximum mFCP connections between two Eclipse SAN routers	N/A	4



IBM TotalStorage m-type family real-life routing solutions

This chapter discusses the details of some real-life solutions that were implemented with the IBM TotalStorage m-type family routing products. It covers the following solutions:

- ▶ Backup consolidation
- ▶ Migration to new storage environment
- ▶ Long distance disaster recovery over IP

Important: The solutions and sizing estimates that we discuss or make in this chapter are unique. Make no assumptions that they will be supported or apply to each environment. We recommend that you engage IBM to discuss any proposal.

13.1 Backup consolidation

In this scenario, we present a solution to consolidate the local area network (LAN)-free tape backups from two storage area network (SAN) fabrics.

13.1.1 Customer environment and requirements

The customer has two existing SAN fabrics and is currently using ArcServe software to back up the Windows servers in the SAN fabrics to a tape. The customer also has several application servers that do not have SAN attachment. Figure 13-1 shows the customer's environment.

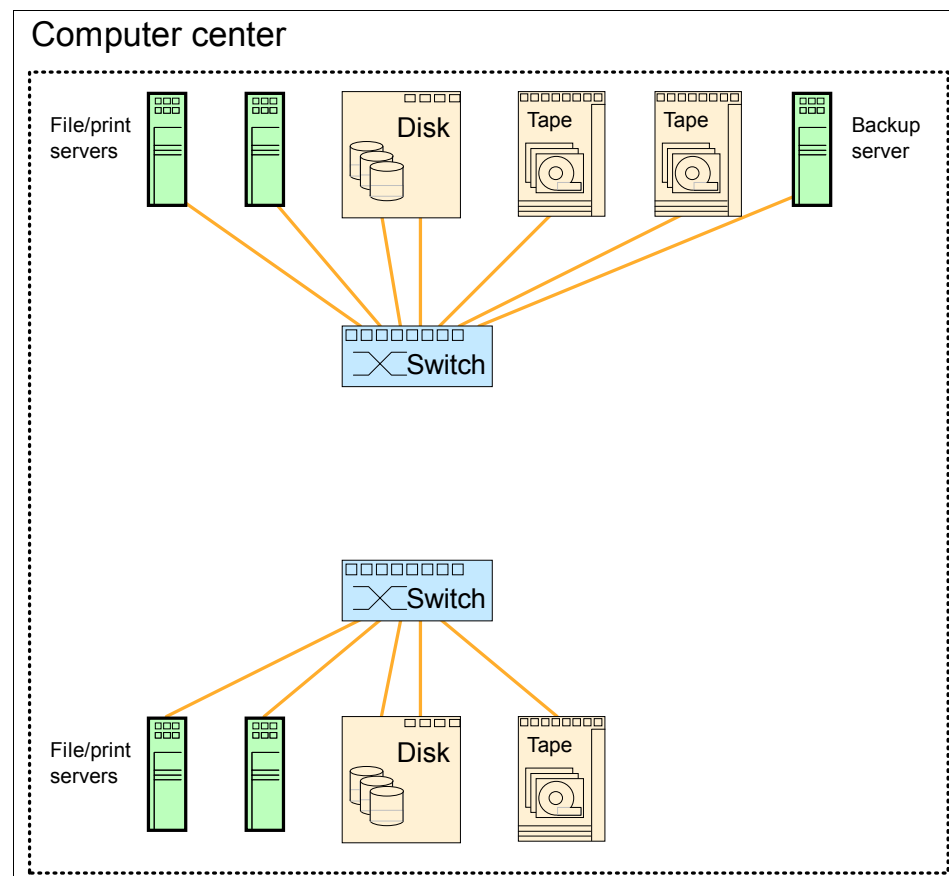


Figure 13-1 Current backup environment

The customer has the following requirements for the new solution:

- ▶ Consolidate the tape backups to a single Tivoli Storage Manager environment
- ▶ Provide for LAN-free backups from both current SAN fabrics
- ▶ Implement the new backup system to a location separate from the computer center
- ▶ Leverage the existing investment to SAN hardware

In the first SAN fabric, the customer currently has 160 GB of disk space, which is projected to grow to 630 GB in the near future. In the second SAN fabric, the customer has 100 GB of disk space.

13.1.2 The solution

Our solution has the following new components:

- ▶ IBM @server pSeries server for Tivoli Storage Manager
- ▶ IBM 3583-L72 tape library with four Fibre Channel drives
- ▶ IBM TotalStorage SAN32M-2 switch for the backup environment
- ▶ IBM TotalStorage SAN16M-R router

Figure 13-2 shows the new backup environment.

We locate the router in the computer center to minimize the need of fiber connections between the computer center and the backup site. All other components are located in a single rack at the backup site.

We connect each of the current fabrics and the new backup switch to the router with two inter-fabric links (IFLs) for redundancy. The customer provides the two long wave fiber connections required between the computer center and the backup site.

The Tivoli Storage Manager server will use its internal disks for both Tivoli Storage Manager databases and disk storage pools. Therefore it does not need any access to the existing customer SAN fabrics. The tape drives are divided evenly to the two Fibre Channel adapters in the Tivoli Storage Manager server.

We create a separate mSAN zone for each server in any SAN fabric that needs to have access to the tape drives. The mSAN zone will contain the worldwide name (WWN) of the host bus adapter (HBA) of the server and the WWNs of all the tape drives.

Since our new environment is only used for daily backups, it does not have as high availability requirements as SAN fabrics used for disk access. Therefore, it is adequate to have a single backup switch and a single router in the solution.

The application servers that are not connected to any SAN fabric are backed up to the Tivoli Storage Manager server over a LAN connection.

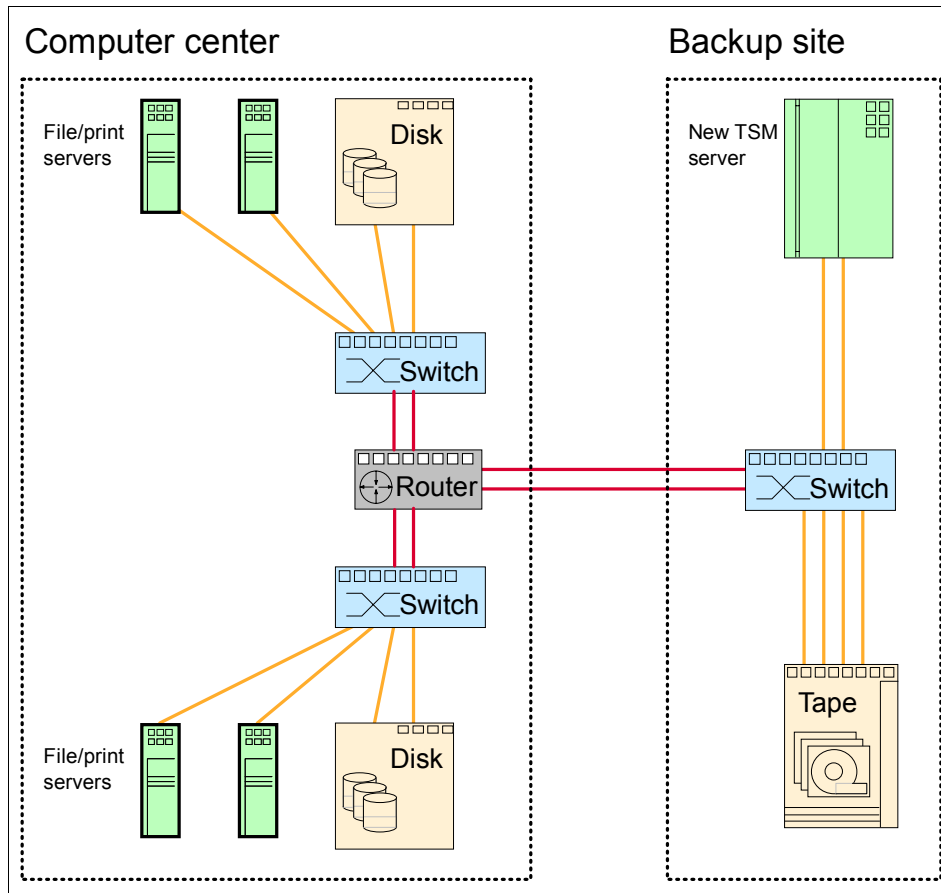


Figure 13-2 New backup environment

13.1.3 Failure scenarios

This section describes how the failure of different components affects the operation of our solution.

- Power failure

The Tivoli Storage Manager server, the tape library, and all of the SAN fabric components in the environment have dual redundant power supplies connected to different power circuits. Therefore, a power failure in one circuit does not have any effect on the operation.

- ▶ IFL failure

If an IFL fails, the system remains operational, but the maximum bandwidth available is reduced by 50%.

- ▶ Router failure

If the SAN router fails, it is impossible to run LAN-free backups. In this situation, the Tivoli Storage Manager client automatically uses a LAN-based method for any backups and restores. The Tivoli Storage Manager server and the servers not using LAN-free backups are not affected.

- ▶ Backup switch or Tivoli Storage Manager server failure

The failure of either the backup switch or the Tivoli Storage Manager server prevents any backup and restore activity.

13.2 Migration to a new storage environment

The following scenario presents a solution to migrate the customer's current storage environment to a new environment.

13.2.1 Customer environment and requirements

The customer has a Hewlett-Packard (HP) XP512 storage system that is shared between AIX, HP-UX, and Windows servers. Due to historical reasons, each server platform has its own SAN fabrics and connections to the XP512.

Each SAN fabric consists of a single 16-port, 1 Gbps switch. Since the lease period of the environment expires within a few months, the customer needs a new solution to replace the current environment.

Figure 13-3 shows the initial environment. For clarity you see only some of the servers.

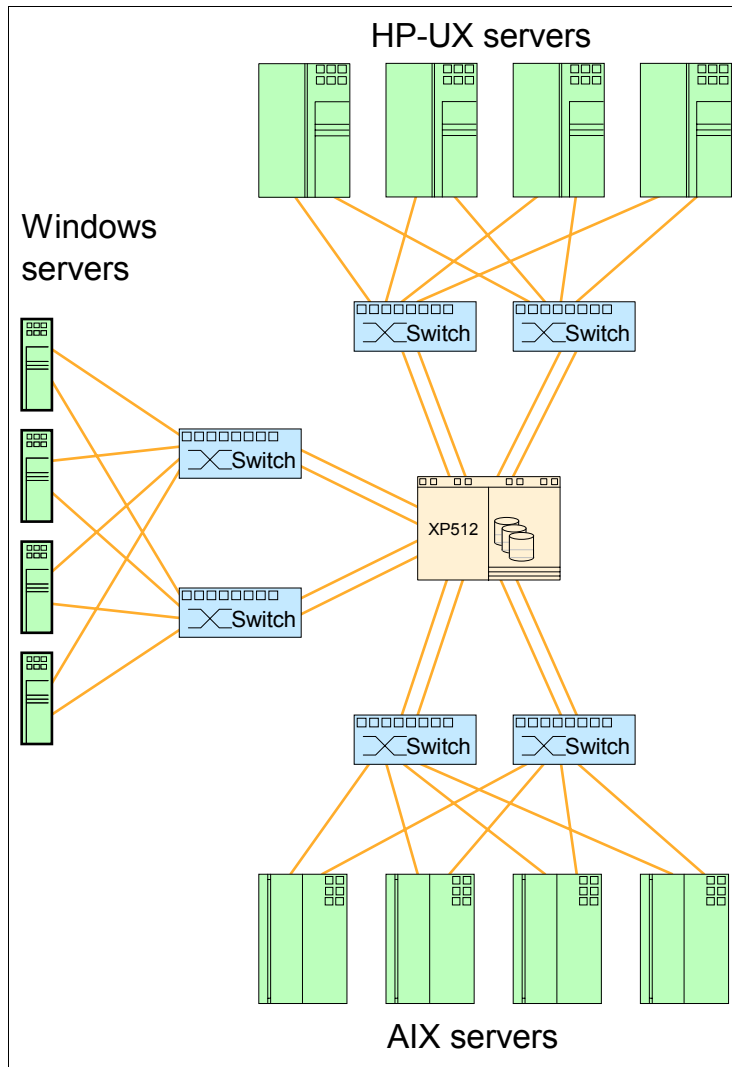


Figure 13-3 Initial storage environment

The customer has the following requirements for the new solution:

- ▶ New hardware to replace the current disk system and SAN fabric
- ▶ Flexibility in allocating ports between different platforms
- ▶ Scalability to support future applications
- ▶ Minimized amount of downtime of servers due to migration

13.2.2 The solution

Our solution has the following new components:

- ▶ IBM TotalStorage DS8100 disk subsystem
- ▶ Two IBM TotalStorage SAN140M directors with 64 ports each
- ▶ Two IBM TotalStorage SAN16M-R routers

We install the components of the new storage environment and connect the environment to the old environment with IFLs, as shown in Figure 13-4.

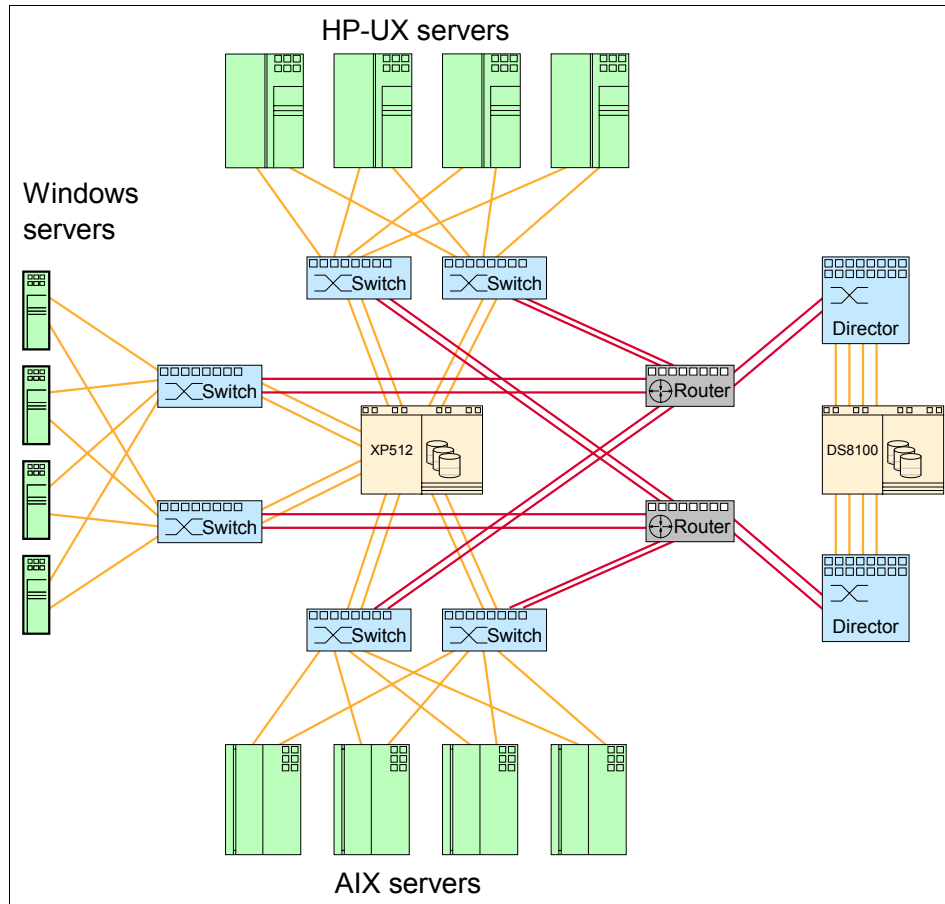


Figure 13-4 Interim environment for migration

In the new environment, all the DS8100 ports are shared among all servers. Since we only migrate a few servers at the same time, using a limited number of IFLs does not cause any performance degradation to the servers.

When the new storage environment is completely installed, we start migrating the servers one server or a group of servers at a time, using the following procedure:

1. Create mSAN zones to allow the server to access the DS8100.
2. Install the IBM Subsystem Device Driver (SDD) package and any other DS8100 specific software on the server.
3. Allocate new storage in the DS8100 to the servers.
4. Migrate all server data from old storage to new storage using the operating system-based tools.
 - Native LVM for AIX
 - PVLinks for HP-UX
 - Veritas Volume Manager for Windows
5. Create fabric zones to allow the server to access the storage from the new SAN fabrics.
6. Disconnect the server from the old switches and move it to the new directors.
7. Delete the mSAN zones created in step 1.

The only step that requires server downtime in the procedure is step 6. If the new cabling is prepared beforehand, this step should take little time.

After the migration of all servers is complete, we should have no servers connected to the old switches and the XP512 should be idle. At this time we can remove the old storage hardware from the environment. The IBM TotalStorage SAN16M-R routers are also freed and can be used for other purposes, such as SAN extension over Internet Fibre Channel Protocol (iFCP).

The final storage environment is shown in Figure 13-5.

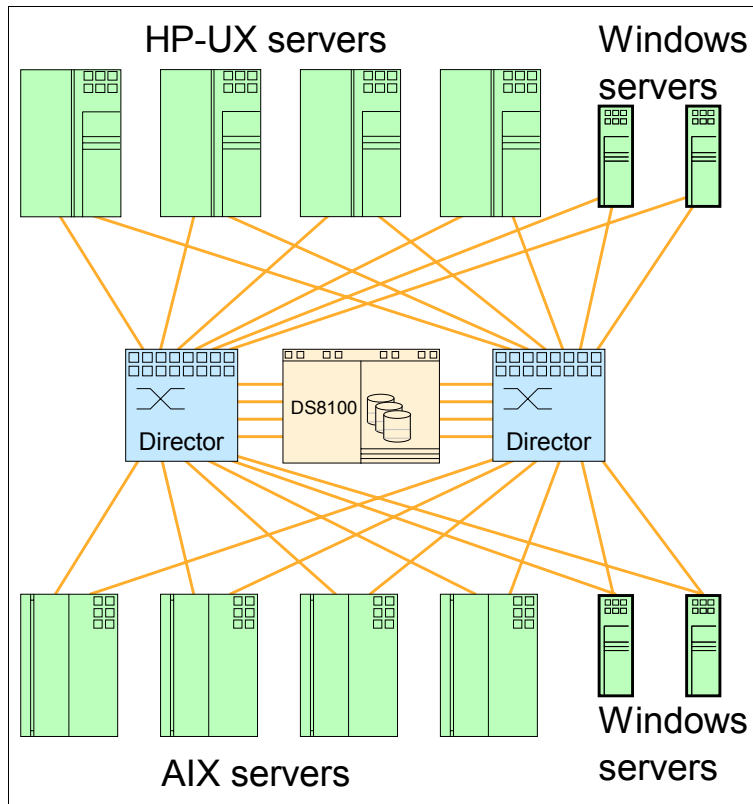


Figure 13-5 Final storage environment

13.3 Long distance disaster recovery over IP

In this scenario, we present a solution that allows for long distance disaster recovery (DR) over an IP connection.

13.3.1 Customer environment and requirements

The customer has three different SAN islands that need to be connected.

- ▶ Development SAN at the primary site
- ▶ Production SAN at the primary site
- ▶ DR SAN at the DR site

The distance between the primary site and the disaster recovery site is 600 km. The amount of data in the production environments is expected to grow to 5 TB within two years, and we expect 3% of the data to change during the peak hour.

The customer has the following requirements for the solution:

- ▶ Provide asynchronous replication for production data from the primary site to the DR site, with a 5 minute recovery point objective (RPO) and a 5 minute recovery time objective (RTO).
- ▶ Keep the dual fabrics of each SAN both physically and logically separate.
- ▶ Provide access to a point-in-time copy of productive data from the test environment at the development SAN.
- ▶ Provide for LAN-free backup from the development network to the tape library in the production network.

The detailed list of the current environment is:

- ▶ Production environment at the primary site
 - Dual SAN fabrics, based on IBM TotalStorage SAN140M directors
 - IBM TotalStorage DS8100 disk subsystem with eight Fibre Channel ports
 - IBM TotalStorage 3584 tape library with six IBM 3592 tape drives
 - Eight pSeries servers, with dual Fibre Channel adapters
 - Sixteen IBM @server xSeries servers, with dual Fibre Channel adapters
- ▶ Development environment at the primary site
 - Dual SAN fabrics, based on IBM TotalStorage SAN 32M-2 switches
 - IBM TotalStorage DS6800 disk subsystem with four Fibre Channel ports
 - Eight pSeries servers, with dual Fibre Channel adapters
 - Sixteen xSeries servers, with dual Fibre Channel adapters
- ▶ Disaster recovery environment at the disaster recovery site
 - Dual SAN fabrics, based on IBM TotalStorage SAN140M directors
 - IBM TotalStorage DS8100 disk subsystem with eight Fibre Channel ports
 - IBM TotalStorage 3584 tape library with six IBM 3592 tape drives
 - Eight pSeries servers, with dual Fibre Channel adapters
 - Sixteen xSeries servers, with dual Fibre Channel adapters

Figure 13-6 shows the environment. For clarity you see only some of the servers and connections.

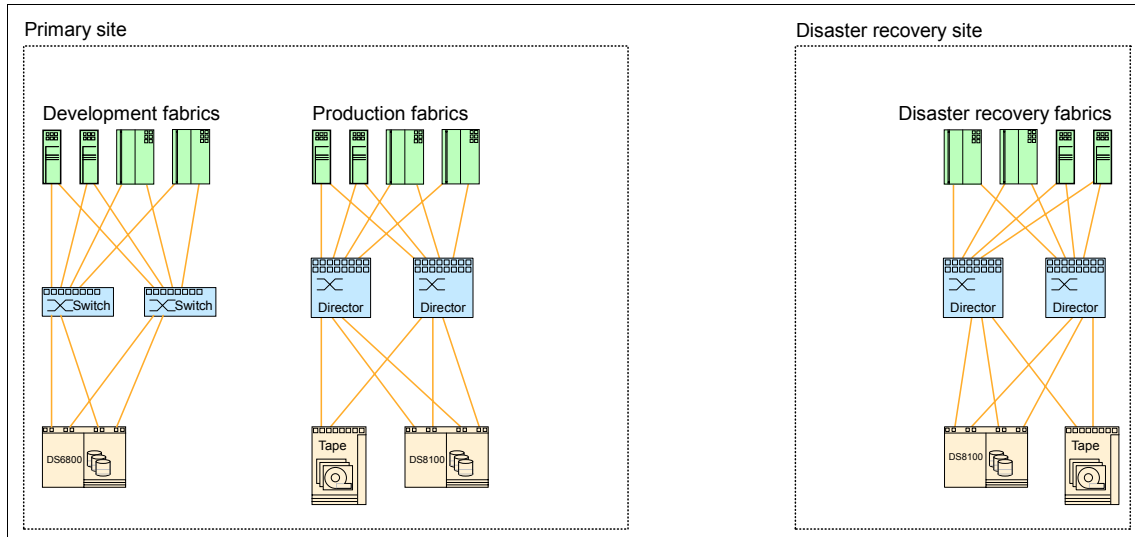


Figure 13-6 Customer environment

13.3.2 The solution

Our solution has the following components:

- ▶ DS8100 Global Mirroring feature for asynchronous replication
- ▶ Four IBM TotalStorage SAN16M-R routers (2027-R16)
- ▶ Four IP links between the 2027-R16 routers from the primary site to the disaster recovery site
- ▶ IBM eRCMF software to provide automatic failover of both the pSeries and xSeries servers

Figure 13-7 shows the complete solution.

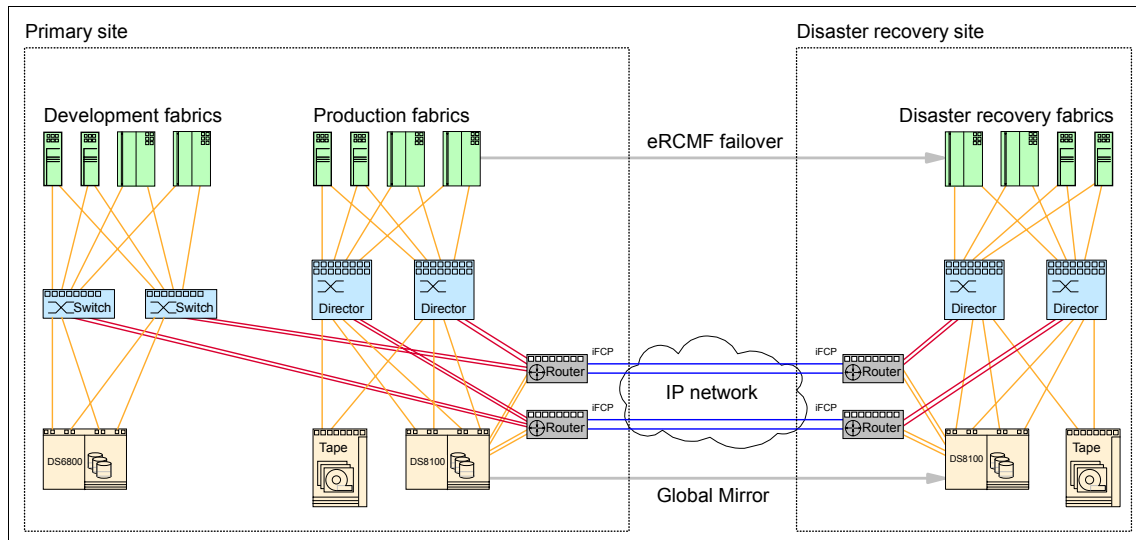


Figure 13-7 Disaster recovery solution

iFCP link sizing

Since we are using the iFCP links for Global Mirror between the DS8100 systems, we need to take into account any changes to the data when sizing the links. Based on the customer's requirements, the amount of data changing during the peak hour is 3% of 5 TB, or approximately 150 GB. If we assume that the changes are evenly divided over the hour, the changes are 2.5 GB per minute, or approximately 42 MBps or 336 Mbps. We use this number as a basis for our link sizing.

If we divide the amount evenly across four links, we get traffic of 84 MBps over each link. However, to allow for the loss of one link or any peaks in the traffic, we divide the traffic only across three links, giving us 33% extra bandwidth and 112 MBps traffic over each link. We also plan to have a maximum of 90% utilization on the link, so the minimum link speed we need is 125 MBps.

Each link can be implemented over an OC-3 line that has the capacity of 155 MBps. An alternative is to use an Multiprotocol Label Switching (MPLS)-based, shared connection, but due to possible router latency issues, we prefer the private OC-3 -based connection.

The most significant part of the OC-3 link latency is the propagation time of the light within the fiber. For a 600 km connection, with 1200 km round trip, it is:

$$1200 \times 4.8 \mu\text{s} = 5.8 \text{ ms}$$

We round this up to 6 ms to account for the packet transmission time over the 155 Mbps OC-3 link.

13.3.3 Normal operation

In normal operation, the production servers use only the DS8100 disks in the primary site. The DR servers are connected to the DS8100 in the DR site, but do not have the disk mounted or any applications running. The development servers are using the DS6800 disks in the primary site and some capacity from the DS8100 in the primary site.

For the DS8100 disk subsystems, four of the eight ports are used for host attachments; the remaining four are used for the Global Mirroring. The ports used for Global Mirroring are directly connected to the routers.

In addition to the normal zoning, we define the following mSAN zones to our environment:

- ▶ Separate mSAN zones for the HBAs of any server in the development fabric that needs access to the DS8100, containing:
 - The HBA of the server
 - Both Fibre Channel ports of the DS8100 used for host attachment in the fabric
- ▶ A separate mSAN zone for the HBAs of any server in the development fabric that needs access to LAN-free backup, containing:
 - The HBA of the server
 - All Fibre Channel ports of the tape drives in the primary site connected to the fabric

In addition, we define zones for each Global Mirror connection in the backbone fabric.

13.3.4 Failure scenarios

This section describes how the failure of different components affects the operation of our solution.

- ▶ Power failure

All of the SAN fabric components in the environment have dual redundant power supplies connected to different power circuits. Therefore a power failure in one circuit does not have any effect on the operation.

► iFCP link failure

The failure of a single iFCP link reduces the available bandwidth between the sites by 25%. However, since we assumed three available links in our sizing, the performance of the system will still remain adequate.

► Development fabric switch failure

The failure of a switch in the development fabric reduces the Fibre Channel bandwidth available for development and test servers by 50%. The traffic is automatically routed via the remaining paths by the SDD. The production environment is not affected.

► Primary site router failure

If the router at the primary site fails, the capacity of the Global Mirror connection will be reduced by 50%. However, since we rounded up our link speed, we still have about 300 Mbps or about 90% of the peak hour capacity available.

In addition, it reduces the Fibre Channel bandwidth available between development and test servers, and the storage in the production fabrics, by 50%.

► Primary site director failure

The director failure at the primary site reduces the Fibre Channel bandwidth available for production servers by 50%.

In addition it reduces the Fibre Channel bandwidth available between development and test servers, and the storage in the production fabrics, by 50%.

► DR site router failure

If the router at the DR site fails, the capacity of the Global Mirror connection will be reduced by 50%. However, since we rounded up our link speed, we still have about 300 Mbps or about 90% of the peak hour capacity available.

► DR site director failure

The director failure at DR site reduces the Fibre Channel bandwidth available for DR servers by 50%. However, in normal situations those servers are idle, so this reduction affects the system only in a case where the production workload is already running at the DR site.

► Primary site DS8100 port failure

If a port used for host access in the DS8100 at the primary site fails, the Fibre Channel bandwidth available for host access is reduced by 25%.

If a port used for Global Mirror in the DS8100 at the primary site fails, the remaining Fibre Channel ports can sustain the full Global Mirror performance.

- ▶ Primary site DS8100 failure

If the DS8100 at the primary site fails, all hosts lose access to it. This event can be promoted to site failure, and production can resume at the DR site.

- ▶ Primary site DS8100 port failure

If a port used for host access in the DS8100 at the DR site fails, the Fibre Channel bandwidth available for host access is reduced by 25%. However, in normal operation those servers are idle, so this reduction affects the system only in the case where the production workload is already running at the DR site.

If a port used for Global Mirror in the DS8100 at the primary site fails, the remaining Fibre Channel ports can sustain full Global Mirror performance.

- ▶ DR site DS8100 failure

If the DS8100 at the DR site fails, the Global Mirror connections change to Suspended state. The DS8100 at the primary site will accumulate changes to the data, and copy the changed data over to the DR site, when the DS8100 becomes available.

- ▶ Primary site failure

If the complete primary site fails, the IBM eRCMF software starts the production at the DR site automatically. While manual failover is also possible, it is difficult to manually reach the RTO target.

Glossary

8b/10b A data encoding scheme developed by IBM, translating byte-wide data to an encoded 10-bit format. The Fibre Channel (FC) FC-1 level defines this as the method to use to encode and decode data transmissions over the Fibre Channel.

active configuration In an ESCON® environment, the ESCON Director configuration determined by the status of the current set of connectivity attributes. Contrast with *saved configuration*.

adapter A hardware unit that aggregates other input/output (I/O) units, devices, or communications links to a system bus.

ADSM ADSTAR Distributed Storage Manager.

Advanced Intelligent Tape (AIT) A magnetic tape format by Sony that uses 8 mm cassettes, but is only used in specific drives.

agent In the client-server model, the part of the system that performs information preparation and exchange on behalf of a client or server application. In the Simple Network Management Protocol (SNMP), the managed system. See also *management agent*.

aggregation In the Storage Networking Industry Association Storage Model (SNIA), *virtualization* is known as *aggregation*. This aggregation can take place at the file level or at the level of individual blocks that are transferred to disk.

AIT See *Advanced Intelligent Tape*.

AL See *arbitrated loop*.

allowed In an ESCON Director, the attribute that, when set, establishes dynamic connectivity capability. Contrast with *prohibited*.

AL_PA Arbitrated Loop Physical Address.

American National Standards Institute (ANSI) The primary organization for fostering the development of technology standards in the United States. The ANSI family of Fibre Channel documents provides the standards basis for the Fibre Channel architecture and technology. See also *FC-PH*.

ANSI See *American National Standards Institute*.

APAR See *authorized program analysis report*.

arbitrated loop (AL) A Fibre Channel interconnection technology that allows up to 126 participating node ports and one participating fabric port to communicate.

arbitration The process of selecting one respondent from a collection of several candidates that request service concurrently.

Asynchronous Transfer Mode (ATM) A type of packet switching that transmits fixed-length units of data.

ATL See *Automated Tape Library*.

ATM See *Asynchronous Transfer Mode*.

authorized program analysis report (APAR) A report of a problem caused by a suspected defect in a current, unaltered release of a program.

Automated Tape Library (ATL) Large scale tape storage system, which uses multiple tape drives and mechanisms to address 50 or more cassettes.

backup A copy of computer data, or the act of copying such data, that is used to recreate data that has been lost, mislaid, corrupted, or erased.

bandwidth A measure of the information capacity of a transmission channel.

basic mode An S/390® or IBM® server zSeries® central processing mode that does not use logical partitioning. Contrast with *logically partitioned mode*.

blocked In an ESCON and FICON Director, the attribute that, when set, removes the communication capability of a specific port. Contrast with *unblocked*.

bridge A component used to attach more than one I/O unit to a port. Also a data communications device that connects two or more networks and forwards packets between them. The bridge may use similar or dissimilar media and signaling systems. It operates at the data link level of the OSI model. Bridges read and filter data packets and frames.

bridge/router A device that can provide the functions of a bridge, router, or both concurrently. A bridge/router can route one or more protocols, such as TCP/IP, and bridge all other traffic. See also *bridge* and *router*.

broadcast To send a transmission to all N_Ports on a fabric.

byte 1) In Fibre Channel, an eight-bit entity prior to encoding or after decoding, with its least significant bit denoted as bit 0 and most significant bit as bit 7. The most significant bit is shown on the left side in FC-FS unless otherwise shown. 2) In S/390 architecture or z/Architecture™ for zSeries (and FICON), an eight-bit entity prior to encoding or after decoding, with its least significant bit denoted as bit 7 and most significant bit as bit 0. The most significant bit is shown on the left side in S/390 architecture and z/Architecture for zSeries.

cascaded switches The connecting of one Fibre Channel switch to another Fibre Channel switch, creating a cascaded switch route between two N_Nodes connected to a Fibre Channel fabric.

chained In an ESCON environment, pertaining to the physical attachment of two ESCON Directors (ESCDs) to each other.

channel 1) A processor system element that controls one channel path, whose mode of operation depends on the type of hardware to which it is attached. In a channel subsystem, each channel controls an I/O interface between the channel control element and the logically attached control units. 2) In ESA/390 or z/Architecture, the part of a channel subsystem that manages a single I/O interface between a channel subsystem and a set of controllers (control units).

channel to channel See *CTC*.

channel to converter See *CVC*.

channel-attached Devices attached directly by data channels (I/O channels) to a computer. Also refers to devices attached to a controlling unit by cables rather than by telecommunication lines.

channel I/O A form of I/O where request and response correlation is maintained through a form of source, destination, and request identification.

channel path (CHP) A single interface between a central processor and one or more control units along which signals and data can be sent to perform I/O requests.

channel path identifier (CHPID) In a channel subsystem, a value assigned to each installed channel path of the system that uniquely identifies that path to the system.

channel subsystem (CSS) Relieves the processor of direct I/O communication tasks, and performs path management functions. Uses a collection of subchannels to direct a channel to control the flow of information between I/O devices and main storage.

CHP See *channel path*.

CHPID See *channel path identifier*.

CIFS Common Internet File System.

cladding In an optical cable, the region of low refractive index surrounding the core. See also *core* and *optical fiber*.

Class of Service A Fibre Channel frame delivery scheme that exhibit a specified set of delivery characteristics and attributes.

Class-1 A class of service that provides dedicated connection between two ports with confirmed delivery or notification of nondeliverability.

Class-2 A class of service that provides a frame switching service between two ports with confirmed delivery or notification of nondeliverability.

Class-3 A class of service that provides frame switching datagram service between two ports or a multicast service between a multicast originator and one or more multicast recipients.

Class-4 A class of service that provides a fractional bandwidth virtual circuit between two ports with confirmed delivery or notification of nondeliverability.

Class-6 A class of service that provides a multicast connection between a multicast originator and one or more multicast recipients with confirmed delivery or notification of nondeliverability.

client A software program used to contact and obtain data from a *server* software program on another computer, often across a great distance. Each *client* program is designed to work specifically with one or more kinds of server programs, and each server requires a specific kind of client program.

client/server The relationship between machines in a communications network. The client is the requesting machine, and the server is the supplying machine. Also used to describe the information management relationship between software components in a processing system.

cluster A type of parallel or distributed system that consists of a collection of interconnected whole computers and is used as a single, unified computing resource.

CNC A mnemonic for an ESCON channel used to communicate to an ESCON-capable device.

coaxial cable A transmission media (cable) used for high-speed transmission. It is called *coaxial* because it includes one physical channel that carries the signal surrounded (after a layer of insulation) by another concentric physical channel, both of which run along the same axis. The inner channel carries the signal and the outer channel serves as a ground.

configuration matrix In an ESCON environment or FICON, an array of connectivity attributes that appear as rows and columns on a display device and can be used to determine or change active and saved ESCON or FICON director configurations.

connected In an ESCON Director, the attribute that, when set, establishes a dedicated connection between two ESCON ports. Contrast with *disconnected*.

connection In an ESCON Director, an association established between two ports that provides a physical communication path between them.

connectivity attribute In an ESCON and FICON Director, the characteristic that determines a particular element of a port's status. See *allowed*, *prohibited*, *blocked*, *unblocked*, *as well as connected* and *disconnected*.

control unit A hardware unit that controls the reading, writing, or displaying of data at one or more I/O units.

controller A component that attaches to the system topology through a channel semantic protocol that includes some form of request/response identification.

core In an optical cable, the central region of an optical fiber through which light is transmitted and that has an index of refraction greater than the surrounding cladding material. See also *cladding* and *optical fiber*.

coupler In an ESCON environment, link hardware used to join optical fiber connectors of the same type. Contrast with *adapter*.

CRC See *Cyclic Redundancy Check*.

CSS See *channel subsystem*.

CTC Channel-to-channel. A mnemonic for an ESCON channel attached to another ESCON channel, where one of the two ESCON channels is defined as an ESCON CTC channel and the other ESCON channel is defined as a ESCON CNC channel. Also a mnemonic for a FICON channel supporting a CTC Control Unit function logically or physically connected to another FICON channel that also supports a CTC Control Unit function. FICON channels supporting the FICON CTC control unit function are defined as normal FICON native (FC) mode channels.

CVC A mnemonic for an ESCON channel attached to an IBM 9034 convertor. The 9034 converts ESCON CVC signals to parallel channel interface (OEMI) communication operating in block multiplex mode (Bus and Tag). Contrast with CBY.

Cyclic Redundancy Check (CRC) An error-correcting code used in Fibre Channel.

DASD See *direct access storage device*.

DAT See *Digital Audio Tape*.

data sharing A SAN solution in which files on a storage device are shared between multiple hosts.

datagram Refers to the Class 3 Fibre Channel Service that allows data to be sent rapidly to multiple devices attached to the fabric, with no confirmation of delivery.

DDM See *disk drive module*.

dedicated connection In an ESCON Director, a connection between two ports that is not affected by information contained in the transmission frames. This connection, which restricts those ports from communicating with any other port, can be

established or removed only as a result of actions performed by a host control program or at the ESCD console. Contrast with *dynamic connection*.

Note: The two links having a dedicated connection appear as one continuous link.

default Pertaining to an attribute, value, or option that is assumed when none is explicitly specified.

Dense Wavelength Division Multiplexing (DWDM) The concept of packing multiple signals tightly together in separate groups, and transmitting them simultaneously over a common carrier wave.

destination Any point or location, such as a node, station, or a particular terminal, to which information is to be sent. An example is a Fibre Channel fabric F_Port; when attached to a Fibre Channel N_port, communication to the N_port via the F_port is said to be to the F_Port destination identifier (D_ID).

device A mechanical, electrical, or electronic contrivance with a specific purpose.

device address 1) In ESA/390 architecture and z/Architecture for zSeries, the field of an ESCON device-level frame that selects a specific device on a control unit image. 2) In the FICON channel FC-SB-2 architecture, the device address field in an SB-2 header that is used to select a specific device on a control unit image.

device number 1) In ESA/390 and z/Architecture for zSeries, a four-hexadecimal character identifier (for example, 19A0) that you associate with a device to facilitate communication between the program and the host operator. 2) The device number that you associate with a subchannel that uniquely identifies an I/O device.

dB Decibel. A ratio measurement distinguishing the percentage of signal attenuation (loss) between the I/O power. Attenuation is expressed as dB/km.

Digital Audio Tape (DAT) A tape media technology designed for very high quality audio recording and data backup. DAT cartridges look like

audio cassettes and are often used in mechanical auto-loaders. Typically, a DAT cartridge provides 2 GB of storage, but new DAT systems have much larger capacities.

Digital Linear Tape (DLT) A magnetic tape technology originally developed by Digital Equipment Corporation (DEC) and now sold by Quantum. DLT cartridges provide storage capacities from 10 GB to 35 GB.

direct access storage device (DASD) A mass storage medium on which a computer stores data. any online storage device: a disc, drive or CD-ROM.

disconnected In an ESCON Director, the attribute that, when set, removes a dedicated connection. Contrast with *connected*.

disk A mass storage medium on which a computer stores data.

disk drive module (DDM) A disk storage medium that you use for any host data that is stored within a disk subsystem.

disk mirroring A fault-tolerant technique that writes data simultaneously to two hard disks using the same hard disk controller.

disk pooling A SAN solution in which disk storage resources are pooled across multiple hosts rather than dedicated to a specific host.

distribution panel In an ESCON and FICON environment, a panel that provides a central location for the attachment of trunk and jumper cables and can be mounted in a rack, wiring closet, or on a wall.

DLT See *Digital Linear Tape*.

duplex Pertaining to communication in which data or control information can be sent and received at the same time, from the same node. Contrast with *half duplex*.

duplex connector In an ESCON environment, an optical fiber component that terminates both jumper cable fibers in one housing and provides physical keying for attachment to a duplex receptacle.

duplex receptacle In an ESCON environment, a fixed or stationary optical fiber component that provides a keyed attachment method for a duplex connector.

DWDM See *Dense Wavelength Division Multiplexing*.

dynamic connection In an ESCON Director, a connection between two ports, established or removed by the ESCD and that, when active, appears as one continuous link. The duration of the connection depends on the protocol defined for the frames transmitted through the ports and on the state of the ports. Contrast with *dedicated connection*.

dynamic connectivity In an ESCON Director, the capability that allows connections to be established and removed at any time.

Dynamic I/O Reconfiguration An S/390 and z/Architecture function that allows I/O configuration changes to be made nondisruptively to the current operating I/O configuration.

ECL See *Emitter Coupled Logic*.

ELS See *Extended Link Services*.

EMIF See *ESCON Multiple Image Facility*.

Emitter Coupled Logic (ECL) The type of transmitter used to drive copper media such as Twinax, Shielded Twisted Pair, or Coax.

enterprise network A geographically dispersed network under the auspices of one organization.

Enterprise Systems Architecture/390® (ESA/390) An IBM architecture for mainframe computers and peripherals. Processors that follow this architecture include the S/390 Server family of processors.

Enterprise System Connection (ESCON) 1) An ESA/390 computer peripheral interface. The I/O interface uses ESA/390 logical protocols over a serial interface that configures attached units to a communication fabric. 2) A set of IBM products and services that provide a dynamically connected environment within an enterprise.

entity In general, a real or existing object from the Latin *ens*, or being, which makes the distinction between an object's existence and its qualities. In programming, engineering and probably many other contexts, the word is used to identify units, whether concrete items or abstract ideas, that have no ready name or label.

E_Port Expansion Port. A port on a switch used to link multiple switches together into a Fibre Channel switch fabric.

ESA/390 See *Enterprise Systems Architecture/390*.

ESCD Enterprise Systems Connection (ESCON) Director.

ESCD console The ESCON Director display and keyboard device used to perform operator and service tasks at the ESCD.

ESCON See *Enterprise System Connection*.

ESCON channel A channel having an Enterprise Systems Connection channel-to-control-unit I/O interface that uses optical cables as a transmission medium. May operate in CBY, CNC, CTC or CVC mode. Contrast with *parallel channel*.

ESCON Director An I/O interface switch that provides the interconnection capability of multiple ESCON interfaces (or FICON Bridge (FCV) mode - 9032-5) in a distributed-star topology.

ESCON Multiple Image Facility (EMIF) In the ESA/390 architecture and z/Architecture for zSeries, a function that allows logical partitions (LPARs) to share an ESCON and FICON channel path (and other channel types) by providing each LPAR with its own channel-subsystem image.

exchange A group of sequences which share a unique identifier. All sequences within a given exchange use the same protocol. Frames from multiple sequences can be multiplexed to prevent a single exchange from consuming all the bandwidth. See also *sequence*.

Extended Link Services (ELS) Via a command request, solicits a destination port (N_Port or F_Port) to perform a function or service. Each ELS request consists of an Link Service (LS) command; the N_Port ELS commands are defined in the FC-FS architecture.

fabric Fibre Channel employs a fabric to connect devices. A fabric can be as simple as a single cable connecting two devices. The term is most often used to describe a more complex network using hubs, switches, and gateways.

Fabric Login (FLOGI) Used by an N_Port to determine if a fabric is present and, if so, to initiate a session with the fabric by exchanging service parameters with the fabric. Fabric Login is performed by an N_Port following link initialization and before communication with other N_Ports is attempted.

Fabric Shortest Path First (FSPF) An intelligent path selection and routing standard and is part of the Fibre Channel Protocol.

FC 1) A short form when referring to something that is part of the Fibre Channel standard. Used by the IBM I/O definition process when defining a FICON channel (using IOCP or HCD) that will be used in FICON native mode (using the FC-SB-2 communication protocol. See also *Fibre Channel*.

FC-0 Lowest level of the Fibre Channel Physical standard, covering the physical characteristics of the interface and media.

FC-1 Middle level of the Fibre Channel Physical standard, defining the 8b/10b encoding and decoding and transmission protocol.

FC-2 Highest level of the Fibre Channel Physical standard, defining the rules for signaling protocol and describing transfer of frame, sequence, and exchanges.

FC-3 The hierarchical level in the Fibre Channel standard that provides common services such as striping definition.

FC-4 The hierarchical level in the Fibre Channel standard that specifies the mapping of upper-layer protocols to levels below.

FCA See *Fibre Channel Association*.

FC-AL See *Fibre Channel Arbitrated Loop*.

FC-CT Fibre Channel Common Transport Protocol

FC-FG See *Fibre Channel Fabric Generic*.

FC-FP See *Fibre Channel HIPPI Framing Protocol*.

FC-FS See *Fibre Channel-Framing and Signaling*.

FC-GS See *Fibre Channel Generic Services*.

FCLC See *Fibre Channel Loop Association*.

FC-LE See *Fibre Channel Link Encapsulation*.

FCP See *Fibre Channel Protocol*.

FC-PH See *Fibre Channel Physical and Signaling*.

FC-PLDA Fibre Channel Private Loop Direct Attach. See *Private Loop Direct Attach*.

FCS See *Fibre Channel standard*.

FC-SB See *Fibre Channel Single Byte Command Code Set*.

FC Storage Director SAN Storage Director.

FC-SW See *Fibre Channel Switch Fabric*.

fiber See *optical fiber*.

Fibre Channel A technology for transmitting data between computer devices at a data rate of up to 4 Gbps. It is especially suited for connecting computer servers to shared storage devices and for interconnecting storage controllers and drives.

Fibre Channel Arbitrated Loop (FC-AL) A reference to the FC-AL standard, a shared gigabit media for up to 127 nodes, one of which may be attached to a switch fabric. See also *arbitrated loop*.

Fibre Channel Association (FCA) A Fibre Channel industry association that works to promote awareness and understanding of the Fibre Channel technology and its application, and provides a means for implementers to support the standards committee activities.

Fibre Channel Fabric Generic (FC-FG) A reference to the document (ANSI X3.289-1996) which defines the concepts, behavior, and characteristics of the Fibre Channel fabric along with suggested partitioning of the 24-bit address space to facilitate the routing of frames.

Fibre Channel-Framing and Signaling (FC-FS) The term used to describe the FC-FS architecture.

Fibre Channel Generic Services (FC-GS) A reference to the document (ANSI X3.289-1996) that describes a common transport protocol used to communicate with the server functions, a full X500-based directory service, mapping of the SNMP directly to the Fibre Channel, a time server, and an alias server.

Fibre Channel HIPPI Framing Protocol (FCFP) A reference to the document (ANSI X3.254-1994) that defines how the HIPPI framing protocol is transported via the Fibre Channel.

Fibre Channel Link Encapsulation (FC-LE) A reference to the document (ANSI X3.287-1996) which defines how IEEE 802.2 Logical Link Control (LLC) information is transported via the Fibre Channel.

Fibre Channel Loop Association (FCLC) An independent working group of the FCA focused on the marketing aspects of the Fibre Channel loop technology.

Fibre Channel Physical and Signaling (FC-PH) A reference to the ANSI X3.230 standard, that contains the definition of the three lower levels (FC-0, FC-1, and FC-2) of the Fibre Channel.

Fibre Channel Protocol (FCP) The mapping of SCSI-3 operations to Fibre Channel.

Fibre Channel Service Protocol (FSP) The common FC-4 level protocol for all services, transparent to the fabric type or topology.

Fibre Channel Single Byte Command Code Set (FC-SB) A reference to the document (ANSI X.271-1996) which defines how the ESCON command set protocol is transported using the Fibre Channel.

Fibre Channel standard (FCS) An ANSI standard for a computer peripheral interface. The I/O interface defines a protocol for communication over a serial interface that configures attached units to a communication fabric. The protocol has four layers. The lower of the four layers defines the physical media and interface, the upper of the four layers defines one or more Upper Layer Protocols (ULP), for example, FCP for SCSI command protocols and FC-SB-2 for FICON protocol supported by ESA/390 and z/Architecture. Refer to ANSI X3.230.1999x.

Fibre Channel Switch Fabric (FC-SW) A reference to the ANSI standard under development that further defines the fabric behavior described in FC-FG and defines the communications between different fabric elements required for those elements to coordinate their operations and management address assignment.

fiber optic cable See *optical cable*.

fiber optics The branch of optical technology concerned with the transmission of radiant power through fibers made of transparent materials such as glass, fused silica, and plastic.

Note: Telecommunication applications of fiber optics use optical fibers. Either a single discrete fiber or a non-spatially aligned fiber bundle can be used for each information channel. Such fibers are often called “optical fibers” to differentiate them from fibers used in non-communication applications.

FICON 1) An ESA/390 and zSeries computer peripheral interface. The I/O interface uses ESA/390 and zSeries FICON protocols (FC-FS and FC-SB-2) over a Fibre Channel serial interface that configures attached units to a FICON supported Fibre Channel communication fabric. 2) An FC4 proposed standard that defines an effective mechanism for the export of the SBCCS-2 (FC-SB-2) command protocol via Fibre Channels.

FICON channel A channel having a Fibre Channel connection (FICON) channel-to-control-unit I/O interface that uses optical cables as a transmission medium. May operate in either FC or FCV mode.

FICON Director A Fibre Channel switch that supports the ESCON-like “control unit port” (CUP function) that is assigned a 24-bit Fibre Channel port address to allow FC-SB-2 addressing of the CUP function to perform command and data transfer. (In the Fibre Channel world, it is a means of in-band management using a FC-4 ULP.)

field replaceable unit (FRU) An assembly that is replaced in its entirety when any one of its required components fails.

F_Node Fabric Node. A fabric attached node.

FLOGI See *Fabric Login*.

F_Port Fabric Port. A port used to attach a Node Port (N_Port) to a switch fabric.

frame A linear set of transmitted bits that define the basic transport unit. The frame is the most basic element of a message in Fibre Channel communications, consisting of a 24-byte header and zero to 2112 bytes of data. See also *sequence*.

FRU See *field replaceable unit*.

FSP See *Fibre Channel Service Protocol*.

FSPF See *Fabric Shortest Path First*.

full duplex A mode of communications allowing simultaneous transmission and reception of frames.

gateway A node on a network that interconnects two otherwise incompatible networks.

Gbps Gigabits per second. Also sometimes referred to as Gb/s. In computing terms, it is approximately 1 000 000 000 bits per second. Most precisely it is 1 073 741 824 (1024 x 1024 x 1024) bits per second.

GBps Gigabytes per second. Also sometimes referred to as GB/s. In computing terms, it is approximately 1 000 000 000 bytes per second. Most precisely it is 1 073 741 824 (1024 x 1024 x 1024) bytes per second.

GBIC See *Gigabit Interface Converter*.

Gigabit One billion bits or one thousand megabits.

Gigabit Interface Converter (GBIC) Industry standard transceivers for connection of Fibre Channel nodes to arbitrated loop hubs and fabric switches.

Gigabit Link Module (GLM) A generic Fibre Channel transceiver unit that integrates the key functions necessary for the installation of a Fibre channel media interface on most systems.

GLM See *Gigabit Link Module*.

G_Port Generic Port. A generic switch port that is either an F_Port or E_Port. The function is automatically determined during login.

half duplex In data communication, pertaining to transmission in only one direction at a time. Contrast with *duplex*.

hard disk drive Storage media within a storage server used to maintain information that the storage server requires. Also a mass storage medium for computers that is typically available as a fixed disk or a removable cartridge.

hardware The mechanical, magnetic, and electronic components of a system, such as computers, telephone switches, and terminals.

HBA Host bus adapter.

HCD Hardware configuration dialog.

HDA See *head and disk assembly*.

HDD See *hard disk drive*.

head and disk assembly (HDA) The portion of an HDD associated with the medium and the read/write head.

hierarchical storage management (HSM) A software and hardware system that moves files from disk to slower, less expensive storage media based on rules and observation of file activity. Modern HSM systems move files from magnetic disk to optical disk to magnetic tape.

High Performance Parallel Interface (HPPI) An ANSI standard that defines a channel that transfers data between CPUs and from a CPU to disk arrays and other peripherals.

HPPI See *High Performance Parallel Interface*.

HMMP HyperMedia Management Protocol.

HMMS See *HyperMedia Management Schema*.

hop An Fibre Channel frame may travel from a switch to a director, a switch to a switch, or a director to a director, which in this case is one hop.

HSM See *Hierarchical Storage Management*.

hub A Fibre Channel device that connects nodes into a logical loop by using a physical star topology. Hubs will automatically recognize an active node

and insert the node into the loop. A node that fails or is powered off is automatically removed from the loop.

hub topology See *loop topology*.

Hunt Group A set of associated N_Ports attached to a single node, assigned a special identifier that allows any frames containing this identifier to be routed to any available N_Port in the set.

HyperMedia Management Schema (HMMS) The definition of an implementation-independent, extensible, common data description/schema, that allows data from a variety of sources to be described and accessed in real time regardless of the source of the data. See also *WEBM* and *HMMP*.

ID See *identifier*.

identifier A unique name or address that identifies such items as programs, devices, or systems.

in-band signaling Signaling that is carried in the same channel as the information. Also referred to as in-band.

in-band virtualization An implementation in which the virtualization process takes place in the data path between servers and disk systems. The virtualization can be implemented as software running on servers or in dedicated engines.

information unit A unit of information defined by an FC-4 mapping. Information units are transferred as a Fibre Channel sequence.

initial program load (IPL) 1) The initialization procedure that causes an operating system to commence operation. 2) The process by which a configuration image is loaded into storage at the beginning of a work day or after a system malfunction. (3) The process of loading system programs and preparing a system to run jobs.

input/output (I/O) 1) Pertaining to a device whose parts can perform an input process and an output process at the same time. 2) Pertaining to a functional unit or channel involved in an input

process, output process, or both, concurrently or not, and to the data involved in such a process. (3) Pertaining to input, output, or both.

input/output configuration data set (IOCDS) The data set in the S/390 and zSeries processor (in the support element) that contains an I/O configuration definition built by the I/O configuration program (IOCP).

input/output configuration program (IOCP) An S/390 program that defines to a system the channels, I/O devices, paths to the I/O devices, and the addresses of the I/O devices. The output is normally written to a S/390 or zSeries IOCDS.

interface 1) A shared boundary between two functional units, defined by functional characteristics, signal characteristics, or other characteristics as appropriate. The concept includes the specification of the connection of two devices having different functions. 2) Hardware, software, or both, that link systems, programs, or devices.

intermix A mode of service defined by Fibre Channel that reserves the full Fibre Channel bandwidth for a dedicated Class 1 connection, but allows connection-less Class 2 traffic to share the link if the bandwidth is available.

inter-switch link (ISL) An Fibre Channel connection between switches and directors.

I/O See *input/output*.

I/O configuration The collection of channel paths, control units, and I/O devices that attaches to the processor. This may also include channel switches (for example, an ESCON Director).

IOCDS See *input/output configuration data set*.

IOCP See *input/output configuration control program*.

IODF The data set that contains the S/390 or zSeries I/O configuration definition file produced during the definition of the S/390 or zSeries I/O configuration by HCD. Used as a source for IPL, IOCP, and Dynamic I/O Reconfiguration.

IP Internet Protocol

IPI Intelligent Peripheral Interface

IPL See *initial program load*.

ISL See *inter-switch link*.

isochronous transmission Data transmission which supports network-wide timing requirements. A typical application for isochronous transmission is a broadcast environment which needs information to be delivered at a predictable time.

JBOD Just a bunch of disks.

jukebox A device that holds multiple optical disks and one or more disk drives, and can swap disks in and out of the drive as needed.

jumper cable In an ESCON and FICON environment, an optical cable having two conductors that provide physical attachment between a channel and a distribution panel or an ESCON/FICON Director port or a control unit/device, between an ESCON/FICON Director port and a distribution panel or a control unit/device, or between a control unit/device and a distribution panel. Contrast with *trunk cable*.

LAN See *local area network*.

laser A device that produces optical radiation using a population inversion to provide *light amplification by stimulated emission of radiation* and (generally) an optical resonant cavity to provide positive feedback. Laser radiation can be highly coherent temporally, spatially, or both.

latency A measurement of the time it takes to send a frame between two locations.

LC Lucent Connector. A registered trademark of Lucent Technologies.

LCU See *logical control unit*.

LED See *light emitting diode*.

licensed internal code (LIC) Microcode that IBM does not sell as part of a machine, but instead, licenses it to the client. LIC is implemented in a part of storage that is not addressable by user programs. Some IBM products use it to implement functions as an alternate to hard-wire circuitry.

light emitting diode (LED) A semiconductor chip that gives off visible or infrared light when activated. Contrast with *laser*.

link 1) In an ESCON environment or FICON environment (Fibre Channel environment), the physical connection and transmission medium used between an optical transmitter and an optical receiver. A link consists of two conductors, one used for sending and the other for receiving, thereby providing a duplex communication path. 2) In an ESCON I/O interface, the physical connection and transmission medium used between a channel and a control unit, a channel and an ESCD, a control unit and an ESCD, or at times between two ESCDs. 3) In a FICON I/O interface, the physical connection and transmission medium used between a channel and a control unit, a channel and a FICON Director, a control unit and a Fibre Channel FICON Director, or at times between two Fibre Channels switches.

link address 1) On an ESCON interface, the portion of a source or destination address in a frame that ESCON uses to route a frame through an ESCON director. ESCON associates the link address with a specific switch port that is on the ESCON director. 2) On a FICON interface, the port address (1-byte link address), or domain and port address (2-byte link address) portion of a source (S_ID) or destination address (D_ID) in a Fibre Channel frame that the Fibre Channel switch uses to route a frame through a Fibre Channel switch or Fibre Channel switch fabric. See also *port address*.

Link_Control_Facility A termination card that handles the logical and physical control of the Fibre Channel link for each mode of use.

LIP See *loop initialization primitive sequence*.

local area network (LAN) A computer network located in a user's premises within a limited geographic area, usually not larger than a floor or small building. Transmissions within a LAN are mostly digital, carrying data among stations at rates usually above one Mbps.

logical control unit (LCU) A separately addressable control unit function within a physical control unit. Usually a physical control unit that supports several LCUs. For ESCON, the maximum number of LCUs that can be in a control unit (and addressed from the same ESCON fiber link) is 16. They are addressed from x'0' to x'F'. For FICON architecture, the maximum number of LCUs that can be in a control unit (and addressed from the same FICON fibre link) is 256. They are addressed from x'00' to x'FF'. For both ESCON and FICON, the actual number supported, and the LCU address value, is both processor- and control unit implementation-dependent.

logical partition (LPAR) A set of functions that create a programming environment that is defined by the ESA/390 architecture or z/Architecture for zSeries. The ESA/390 architecture or z/Architecture for zSeries uses the term LPAR when more than one LPAR is established on a processor. An LPAR is conceptually similar to a virtual machine environment except that the LPAR is a function of the processor. Also, LPAR does not depend on an operating system to create the virtual machine environment.

logical switch number (LSN) A two-digit number used by the IOCP to identify a specific ESCON or FICON Director. This number is separate from the director's "switch device number" and, for FICON, it is separate from the director's "Fibre Channel switch address".

logically partitioned mode A central processor mode, available on the configuration frame when

using the PR/SM™ facility, that allows an operator to allocate processor hardware resources among LPARs. Contrast with *basic mode*.

login server An entity within the Fibre Channel fabric that receives and responds to login requests.

loop circuit A temporary point-to-point like path that allows bidirectional communications between loop-capable ports.

loop initialization primitive (LIP) sequence A special Fibre Channel sequence that is used to start loop initialization. Allows ports to establish their port addresses.

loop topology An interconnection structure in which each point has physical links to two neighbors resulting in a closed circuit. In a loop topology, the available bandwidth is shared.

LPAR See *logical partition*.

L_Port Loop Port. A node or fabric port capable of performing arbitrated loop functions and protocols. NL_Ports and FL_Ports are loop-capable ports.

LSN See *logical switch number*.

Lucent Connector (LC) A registered trademark of Lucent Technologies

LVD Low Voltage Differential.

management agent A process that exchanges a managed node's information with a management station.

managed node A computer, a storage system, a gateway, a media device such as a switch or hub, a control instrument, a software product such as an operating system or an accounting package, or a machine on a factory floor, such as a robot.

managed object A variable of a managed node. This variable contains one piece of information about the node. Each node can have several objects.

Management Information Block (MIB) A formal description of a set of network objects that can be managed using the SNMP. The format is defined as part of SNMP and is a hierarchical structure of information relevant to a specific device, defined in object-oriented terminology as a collection of objects, relations, and operations among objects.

management station A host system that runs the management software.

MAR See *Media Access Rules*.

Mbps Megabits per second. Also sometimes referred to as Mb/s. In computing terms, it is approximately 1 000 000 bits per second. Most precisely it is 1 048 576 (1024 x 1024) bits per second.

MBps Megabytes per second. Also sometimes referred to as MB/s. In computing terms, it is approximately 1 000 000 bytes per second. Most precisely it is 1 048 576 (1024 x 1024) bytes per second.

media Plural of medium. The physical environment through which transmission signals pass. Common media include copper and fiber optic cable.

Media Access Rules (MAR) Enable systems to self-configure themselves in a SAN environment.

Media Interface Adapter (MIA) Enables optic-based adapters to interface with copper-based devices, including adapters, hubs, and switches.

metadata server In Storage Tank™, servers that maintain information (metadata) about the data files and grant permission for application servers to communicate directly with disk systems.

meter Equal to 39.37 inches, or just slightly larger than a yard (36 inches)

MIA See *Media Interface Adapter*.

MIB See *Management Information Block*.

mirroring The process of writing data to two separate physical devices simultaneously.

MM Multi-Mode. See *Multi-Mode Fiber*.

MMF See *Multi-Mode Fiber*.

multicast Sending a copy of the same transmission from a single source device to multiple destination devices on a fabric. This includes sending to all N_Ports on a fabric (broadcast) or to only a subset of the N_Ports on a fabric (multicast).

Multi-Mode Fiber (MMF) In optical fiber technology, an optical fiber that is designed to carry multiple light rays or modes concurrently, each at a slightly different reflection angle within the optical core. Multi-Mode fiber transmission is used for relatively short distances because the modes tend to disperse over longer distances. See also *Single-Mode Fiber*.

multiplex The ability to intersperse data from multiple sources and destinations onto a single transmission medium. Refers to delivering a single transmission to multiple destination N_Ports.

name server Provides translation from a given node name to one or more associated N_Port identifiers.

NAS See *Network Attached Storage*.

ND See *node descriptor*.

NDMP Network Data Management Protocol

NED See *node-element descriptor*.

network An aggregation of interconnected nodes, workstations, file servers, and peripherals, with its own protocol that supports interaction.

Network Attached Storage (NAS) A term used to describe a technology where an integrated storage system is attached to a messaging network that uses common communications protocols, such as TCP/IP.

Network File System (NFS) A distributed file system in UNIX developed by Sun™ Microsystems. It allows a set of computers to cooperatively access each other's files in a transparent manner.

Network Management System (NMS) A system responsible for managing at least part of a network. NMSs communicate with agents to help keep track of network statistics and resources.

network topology Physical arrangement of nodes and interconnecting communications links in networks based on application requirements and geographical distribution of users.

NFS See *Network File System*.

NL_Port Node Loop Port. A node port that supports arbitrated loop devices.

NMS See *Network Management System*. A system responsible for managing at least part of a network. NMSs communicate with agents to help keep track of network statistics and resources.

node An entity with one or more N_Ports or NL_Ports.

node descriptor (ND) In an ESCON and FICON environment, a 32-byte field that describes a node, channel, ESCON Director or FICON Director port, or a control unit.

node-element descriptor (NED) In an ESCON and FICON environment, a 32-byte field that describes a node element, such as a disk (DASD) device.

non-blocking Indicates that the capabilities of a switch are such that the total number of available transmission paths is equal to the number of ports. Therefore, all ports can have simultaneous access through the switch.

Non-L_Port A Node or Fabric port that is not capable of performing the arbitrated loop functions and protocols. N_Ports and F_Ports are not loop-capable ports.

N_Port Node Port. A Fibre Channel-defined hardware entity at the end of a link which provides the mechanisms necessary to transport information units to or from another node.

N_Port Login (PLOGI) Allows two N_Ports to establish a session and exchange identities and service parameters. It is performed following completion of the FLOGI process and prior to the FC-4 level operations with the destination port. May be either explicit or implicit.

OEMI See *original equipment manufacturer information*.

open system A system whose characteristics comply with standards made available throughout the industry and that can be connected to other systems that comply with the same standards.

operation A term defined in FC-2 that refers to one of the Fibre Channel *building blocks* composed of one or more, possibly concurrent, exchanges.

optical cable A fiber, multiple fibers, or a fiber bundle in a structure built to meet optical, mechanical, and environmental specifications. See also *jumper cable*, *optical cable assembly*, and *trunk cable*.

optical cable assembly An optical cable that is connector-terminated. Generally, an optical cable that has been connector-terminated by a manufacturer and is ready for installation. See also *jumper cable* and *optical cable*.

optical fiber Any filament made of dielectric materials that guides light, regardless of its ability to send signals. See also *fiber optics* and *optical waveguide*.

optical fiber connector A hardware component that transfers optical power between two optical fibers or bundles and is designed to be repeatedly connected and disconnected.

optical waveguide A structure capable of guiding optical power. In optical communications, generally a fiber designed to transmit optical signals. See *optical fiber*.

ordered set A Fibre Channel term referring to four 10-bit characters (a combination of data and special characters) providing low-level link functions, such as frame demarcation and signaling between two ends of a link.

original equipment manufacturer information (OEMI) A reference to an IBM guideline for a computer peripheral interface. More specifically, it refers to IBM S/360™ and S/370™ Channel to Control Unit OEMI. The interface uses ESA/390 logical protocols over an I/O interface that configures attached units in a multi-drop bus environment. This OEMI interface is also supported by the zSeries 900 processors.

originator A Fibre Channel term referring to the initiating device.

out-of-band signaling Signaling that is separated from the channel carrying the information. Also referred to as *out-of-band*.

out-of-band virtualization An alternative type of virtualization in which servers communicate directly with disk systems under control of a virtualization function that is not involved in the data transfer.

parallel channel A channel having a System/360™ and System/370™ channel-to-control-unit I/O interface that uses bus and tag cables as a transmission medium. Contrast with *ESCON channel*.

path In a channel or communication network, any route between any two nodes. For ESCON and FICON, this is the route between the channel and the control unit/device, or sometimes from the operating system control block for the device and the device itself.

path group The ESA/390 and zSeries architecture (z/Architecture) term for a set of channel paths that are defined to a controller as being associated with

a single S/390 image. The channel paths are in a group state and are online to the host.

path-group identifier ESA/390 and z/Architecture term for the identifier that uniquely identifies a given LPAR. The path-group identifier is used in communication between the system image program and a device. The identifier associates the path group with one or more channel paths, defining these paths to the control unit as being associated with the same system image.

PCICC (IBM) PCI Cryptographic Coprocessor.

peripheral Any computer device that is not part of the essential computer (the processor, memory and data paths) but is situated relatively close by. A near synonym is I/O device.

petard A device that is small and sometimes explosive.

PLDA See *Private Loop Direct Attach*.

PLOGI See *N_Port Login*.

point-to-point topology An interconnection structure in which each point has physical links to only one neighbor resulting in a closed circuit. In point-to-point topology, the available bandwidth is dedicated.

policy-based management Management of data on the basis of business policies (for example, “all production database data must be backed up every day”), rather than technological considerations (for example, “all data stored on this disk system is protected by remote copy”).

port An access point for data entry or exit. A receptacle on a device to which a cable for another device is attached. See also *duplex receptacle*.

port address In an ESCON Director, an address used to specify port connectivity parameters and to assign link addresses for attached channels and control units. In a FICON director or Fibre Channel switch, it is the middle 8 bits of the full 24-bit Fibre Channel port address. This field is also referred to

as the *area field* in the 24-bit Fibre Channel port address. See also *link address*.

port bypass circuit A circuit used in hubs and disk enclosures to automatically open or close the loop to add or remove nodes on the loop.

port card In an ESCON and FICON environment, a field-replaceable hardware component that provides the optomechanical attachment method for jumper cables and performs specific device-dependent logic functions.

port name In an ESCON or FICON Director, a user-defined symbolic name of 24 characters or less that identifies a particular port.

Private Loop Direct Attach (PLDA) A technical report which defines a subset of the relevant standards suitable for the operation of peripheral devices such as disks and tapes on a private loop.

Private NL_Port An NL_Port which does not attempt login with the fabric and only communicates with other NL Ports on the same loop.

processor complex A system configuration that consists of all the machines required for operation; for example, a processor unit, a processor controller, a system display, a service support display, and a power and coolant distribution unit.

program temporary fix (PTF) A temporary solution or bypass of a problem diagnosed by IBM in a current unaltered release of a program.

prohibited In an ESCON or FICON Director, the attribute that, when set, removes dynamic connectivity capability. Contrast with *allowed*.

protocol 1) A set of semantic and syntactic rules that determine the behavior of functional units in achieving communication. 2) In Fibre Channel, the meaning of, and sequencing rules for, requests and responses used for managing the switch or switch fabric, transferring data, and synchronizing states of Fibre Channel fabric components. 3) A specification for the format and relative timing of information exchanged between communicating parties.

PTF See *program temporary fix*.

Public NL_Port An NL_Port that attempts login with the fabric and can observe the rules of either public or private loop behavior. A public NL_Port may communicate with both private and public NL_Ports.

QoS See *Quality of Service*.

Quality of Service (QoS) A set of communications characteristics required by an application. Each QoS defines a specific transmission priority, level of route reliability, and security level.

Quick Loop A unique Fibre Channel topology that combines arbitrated loop and fabric topologies. It is an optional licensed product that allows arbitrated loops with private devices to be attached to a fabric.

RAID See *Redundant Array of Inexpensive or Independent Disks*.

RAID 0 Level 0 RAID support. Striping, no redundancy.

RAID 1 Level 1 RAID support. Mirroring, complete redundancy.

RAID 5 Level 5 RAID support. Striping with parity.

Redundant Array of Inexpensive or Independent Disks (RAID) A method of configuring multiple disk drives in a storage subsystem for high availability and high performance.

repeater A device that receives a signal on an electromagnetic or optical transmission medium, amplifies the signal, and then retransmits it along the next leg of the medium.

responder A Fibre Channel term referring to the answering device.

route The path that an ESCON frame takes from a channel through an ESCD to a control unit/device.

router 1) A device that can decide which of several paths network traffic will follow based on some optimal metric. Routers forward packets from one network to another based on network-layer information. 2) A dedicated computer hardware or software package which manages the connection between two or more networks. See also *bridge* and *bridge/router*.

SAF-TE SCSI Accessed Fault-Tolerant Enclosures.

SAN See *storage area network*.

SAN See *System Area Network*.

SANSymphony In-band block-level virtualization software made by DataCore Software Corporation and resold by IBM.

saved configuration In an ESCON or FICON Director environment, a stored set of connectivity attributes whose values determine a configuration that can be used to replace all or part of the ESCD's or FICON's active configuration. Contrast with *active configuration*.

SC connector A fiber optic connector standardized by ANSI TIA/EIA-568A for use in structured wiring installations.

scalability The ability of a computer application or product (hardware or software) to continue to function because of a change in size or volume. For example, the ability to retain performance levels when adding additional processors, memory, and storage.

SCSI See *Small Computer System Interface*.

SCSI-3 SCSI-3 consists of a set of primary commands and additional specialized command sets to meet the needs of specific device types. The SCSI-3 command sets are used not only for the SCSI-3 parallel interface but for additional parallel and serial protocols, including Fibre Channel, Serial Bus Protocol (used with IEEE 1394 Firewire physical protocol), and the Serial Storage Protocol (SSP).

SCSI Enclosure Services (SES) ANSI SCSI-3 proposal that defines a command set for soliciting basic device status (temperature, fan speed, power supply status, etc.) from a storage enclosures.

SCSI-FCP The term used to refer to the ANSI Fibre Channel Protocol for SCSI document (X3.269-199x) that describes the FC-4 protocol mappings and the definition of how the SCSI protocol and command set are transported using a Fibre Channel interface.

SE See *service element*.

sequence A series of frames strung together in numbered order which can be transmitted over a Fibre Channel connection as a single operation. See also *exchange*.

SERDES Serializer Deserializer.

Serial Storage Architecture (SSA) A high speed serial loop-based interface developed as a high speed point-to-point connection for peripherals, particularly high speed storage arrays, RAID, and CD-ROM storage by IBM.

server A computer which is dedicated to one task.

service element (SE) A dedicated service processing unit used to service a S/390 machine (processor).

SES See *SCSI Enclosure Services*.

Simple Network Management Protocol (SNMP) The Internet network management protocol that provides a means to monitor and set network configuration and run-time parameters.

Single-Mode Fiber (SMF) In optical fiber technology, an optical fiber that is designed for the transmission of a single ray or mode of light as a carrier. It is a single light path used for long-distance signal transmission. See also *Multi-Mode Fiber*.

Small Computer System Interface (SCSI) 1) A set of evolving ANSI standard electronic interfaces that allow personal computers to communicate with

peripheral hardware such as disk drives, tape drives, CD_ROM drives, printers, and scanners faster and more flexibly than previous interfaces. The interface uses a SCSI logical protocol over an I/O interface that configures attached targets and initiators in a multidrop bus topology. The following table identifies the major characteristics of the different SCSI versions.

SCSI version	Signal rate (MHz)	BusWidth (bits)	Maximum DTR (MBps)	Maximum no. devices	Maximum cable length (m)
SCSI-1	5	8	5	7	6
SCSI-2	5	8	5	7	6
Wide SCSI-2	5	16	10	15	6
Fast SCSI-2	10	8	10	7	6
Fast Wide SCSI-2	10	16	20	15	6
Ultra™ SCSI	20	8	20	7	1.5
Ultra SCSI-2	20	16	40	7	12
Ultra2 LVD SCSI	40	16	80	15	12

SM Single Mode. See *Single-Mode Fiber*.

SMART Self Monitoring and Reporting Technology.

SMF See *Single-Mode Fiber*.

SNIA See *Storage Networking Industry Association*.

SN storage network. See also *SAN*.

SNMP See *Simple Network Management Protocol*.

SNMWG See *Storage Network Management Working Group*.

SSA See *Serial Storage Architecture*.

star The physical configuration used with hubs in which each user is connected by communications links radiating out of a central hub that handles all communications.

storage area network (SAN) A dedicated, centrally managed, secure information infrastructure, which enables any-to-any interconnection of servers and storage systems.

storage media The physical device onto which data is recorded. Magnetic tape, optical disks, and floppy disks are all storage media.

Storage Network Management Working Group (SNMWG) Chartered to identify, define, and support open standards needed to address the increased management requirements imposed by storage area network environments.

Storage Networking Industry Association (SNIA) A non-profit organization comprised of more than 77 companies and individuals in the storage industry.

Storage Tank An IBM file aggregation project that enables a pool of storage, and even individual files, to be shared by servers of different types. In this way, Storage Tank can greatly improve storage utilization and enables data sharing.

StorWatch Expert StorWatch applications that employ a three-tiered architecture that includes a management interface, a StorWatch manager and agents that run on the storage resource or resources being managed. Products employ a StorWatch database that can be used for saving key management data, such as capacity or performance metrics. Products also use the agents and analysis of storage data saved in the database to perform higher value functions including the reporting of capacity and performance over time (trends), configuration of multiple devices based on policies, monitoring of capacity and performance, automated responses to events or conditions, and storage related data mining.

StorWatch Specialist A StorWatch interface for managing an individual Fibre Channel device or a limited number of like devices (that can be viewed as a single group). Typically provide simple, point-in-time management functions such as configuration, reporting on asset and status information, simple device and event monitoring, and some service utilities.

STP Shielded Twisted Pair.

striping A method for achieving higher bandwidth using multiple N_Ports in parallel to transmit a single information unit across multiple levels.

subchannel A logical function of a channel subsystem associated with the management of a single device.

subsystem A secondary or subordinate system, or programming support, usually capable of operating independently of or asynchronously with a controlling system.

SWCH In ESCON Manager, the mnemonic used to represent an ESCON Director.

switch A component with multiple entry and exit points (ports) that provides dynamic connection between any two of these points.

switch topology An interconnection structure in which any entry point can be dynamically connected to any exit point. The available bandwidth is scalable.

system area network (SAN) Term originally used to describe a particular symmetric multiprocessing (SMP) architecture in which a switched interconnect is used in place of a shared bus. Server area network refers to a switched interconnect between multiple SMPs.

T11 A technical committee of the National Committee for Information Technology Standards, titled T11 I/O Interfaces. Develops standards for moving data into and out of computers.

tape backup Making magnetic tape copies of hard disk and optical disc files for disaster recovery.

tape pooling A SAN solution in which tape resources are pooled and shared across multiple hosts rather than being dedicated to a specific host.

TCP See *Transmission Control Protocol*.

TCP/IP See *Transmission Control Protocol/ Internet Protocol*.

time server A Fibre Channel-defined service function that allows for the management of all timers used within a Fibre Channel system.

topology An interconnection scheme that allows multiple Fibre Channel ports to communicate. For example, point-to-point, arbitrated loop, and switched fabric are all Fibre Channel topologies.

TL_Port A private to public bridging of switches or directors, referred to as Translative Loop.

T_Port An ISL port more commonly known as an E_Port, referred to as a Trunk port and used by INRANGE.

Transmission Control Protocol (TCP) A reliable, full duplex, connection-oriented end-to-end transport protocol running on top of IP.

Transmission Control Protocol/ Internet Protocol (TCP/IP) A set of communications protocols that support peer-to-peer connectivity functions for both LAN and WANs.

trunk cable In an ESCON and FICON environment, a cable consisting of multiple fiber pairs that do not directly attach to an active device. This cable usually exists between distribution panels (or sometimes between a set processor channels and a distribution panel) and can be located within, or external to, a building. Contrast with *jumper cable*.

twinax A transmission media (cable) consisting of two insulated central conducting leads of coaxial cable.

twisted pair The most common type of transmission media (cable), that consists of two insulated copper wires twisted around each other to reduce the induction (interference) from one wire to another. The twists, or lays, are varied in length to reduce the potential for signal interference between pairs. Several sets of twisted pair wires may be enclosed in a single cable.

ULP Upper Level Protocols,

unblocked In an ESCON and FICON Director, the attribute that, when set, establishes communication capability for a specific port. Contrast with *blocked*.

Under-The-Covers (UTC) A term used to characterize a subsystem in which a small number of hard drives are mounted inside a higher function unit. The power and cooling are obtained from the system unit. Connection is by parallel copper ribbon cable or pluggable backplane, using IDE or SCSI protocols.

unit address The ESA/390 and zSeries term for the address associated with a device on a given controller. On ESCON and FICON interfaces, the unit address is the same as the device address. On OEMI interfaces, the unit address specifies a controller and device pair on the interface.

UTC See *Under-The-Covers*.

UTP Unshielded Twisted Pair

virtual circuit A unidirectional path between two communicating N_Ports that permits fractional bandwidth.

virtualization An abstraction of storage where the representation of a storage unit to the operating system and applications on a server is divorced from the actual physical storage where the information is contained.

virtualization engine Dedicated hardware and software that are used to implement virtualization.

WAN See *wide area network*.

Wave Division Multiplexing (WDM) A technology that puts data from different sources together on an optical fiber, with each signal carried on its own separate light wavelength. Using WDM, up to 80 (and theoretically more) separate wavelengths or channels of data can be multiplexed into a stream of light transmitted on a single optical fiber.

WDM See *Wave Division Multiplexing*.

Web-Based Enterprise Management (WEBM) A consortium working on the development of a series of standards to enable active management and monitoring of network-based elements.

WEBM See *Web-Based Enterprise Management*.

wide area network (WAN) A network which encompasses inter-connectivity between devices over a wide geographic area. A WAN may be privately owned or rented, but the term usually indicates the inclusion of public (shared) networks.

z/Architecture An IBM architecture for mainframe computers and peripherals. Processors that follow this architecture include the zSeries family of processors.

zoning In Fibre Channel environments, the grouping together of multiple ports to form a virtual private storage network. Ports that are members of a group or zone can communicate with each other but are isolated from ports in other zones.

zSeries A family of IBM mainframe servers that support high performance, availability, connectivity, security, and integrity.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

IBM Redbooks

- ▶ *IBM Storage Solutions for Server Consolidation*, SG24-5355
- ▶ *Implementing the Enterprise Storage Server in Your Environment*, SG24-5420
- ▶ *IBM Enterprise Storage Server*, SG24-5465
- ▶ *Introduction to Storage Area Networks*, SG24-5470
- ▶ *IBM Tape Solutions for Storage Area Networks and FICON*, SG24-5474
- ▶ *IBM TotalStorage: Implementing an Open IBM SAN*, SG24-6116
- ▶ *IP Storage Networking: IBM NAS and iSCSI Solutions*, SG24-6240
- ▶ *Implementing Linux with IBM Disk Storage*, SG24-6261
- ▶ *Implementing the IBM TotalStorage NAS 300G: High Speed Cross Platform Storage and Tivoli SANergy!*, SG24-6278
- ▶ *Using iSCSI Solutions' Planning and Implementation*, SG24-6291
- ▶ *IBM TotalStorage: SAN Product, Design, and Optimization Guide*, SG24-6384
- ▶ *Implementing the Cisco MDS 9000 in an Intermix FCP, FCIP, and FICON Environment*, SG24-6397
- ▶ *Introduction to SAN Distance Solutions*, SG24-6408
- ▶ *Introducing Hosts to the SAN fabric*, SG24-6411
- ▶ *The IBM TotalStorage NAS Integration Guide*, SG24-6505
- ▶ *iSCSI Performance Testing & Tuning*, SG24-6531

Other resources

These publications are also relevant as further information sources:

- ▶ Clark, Tom. *IP SANs: An Introduction to iSCSI, iFCP, and FCIP Protocols for Storage Area Network*. Addison-Wesley Professional, first edition, December 2001. ISBN 0201752778.
- ▶ Farley, Marc. *Building Storage Networks*. McGraw-Hill/Osborne Media, first edition, January 2000. ISBN 0072120509.
- ▶ Judd, Josh. *Multiprotocol Routing for SANs*. Infinity Publishing, October 2004. ISBN 0741423065.

Referenced Web sites

These Web sites are also relevant as further information sources:

- ▶ IBM TotalStorage hardware, software, and solutions
<http://www.storage.ibm.com>
- ▶ IBM TotalStorage storage area network
<http://www.storage.ibm.com/snetwork/index.html>
- ▶ Brocade
<http://www.brocade.com>
- ▶ Cisco
<http://www.cisco.com>
- ▶ McDATA
<http://www.inrange.com/>
- ▶ QLogic
<http://www.qlogic.com>
- ▶ Emulex
<http://www.emulex.com>
- ▶ Finisar
<http://www.finisar.com>
- ▶ Veritas
<http://www.veritas.com>
- ▶ Tivoli
<http://www.tivoli.com>

- ▶ JNI
<http://www.Jni.com>
- ▶ IEEE
<http://www.ieee.org>
- ▶ Storage Networking Industry Association
<http://www.snia.org>
- ▶ 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>
- ▶ xSeries 430 and NUMA-Q Information Center
<http://publib.boulder.ibm.com/xseries/>

How to get IBM Redbooks

You can order hardcopy Redbooks, as well as view, download, or search for Redbooks at the following Web site:

ibm.com/redbooks

You can also download additional materials (code samples or diskette/CD-ROM images) from that site.

IBM Redbooks collections

Redbooks are also available on CD-ROMs. Click the CD-ROMs button on the Redbooks Web site for information about all the CD-ROMs offered, as well as updates and formats.

Index

Numerics

2027-R04 151–152
2027-R16 151, 154, 156, 158
2062-D01 67–68
2062-D07 70–71
2062-T07 71–72
2109-A16 20, 24, 30, 33, 40
2109-A16 hardware components 21
2109-A16 internal clock 41
3DES 132
802.3ad Link Aggregation 162
9500 series directors 72

A

accelerated write request 13
access 73, 82, 96–97
access control list (ACL) 132
acknowledgement frames 11, 47, 196
ACL (access control list) 132
active state 95
active supervisor 70–71, 73
active supervisor module 82
active zoneset 106
Address Resolution Protocol (ARP) 83
addressing 34, 41
 schemes 3
administrative state 95
Advanced Encryption Standard (AES) 132
Advanced WEB TOOLS 24
AES (Advanced Encryption Standard) 132
aggregate bandwidth 74, 98, 100
airflow for cooling 67
any-to-any connectivity 152
API (application programming interface) 80
application integration 85
application programming interface (API) 80
arbitrated loop 89
architect 119
architecture 160
ARP (Address Resolution Protocol) 83
ASIC 22, 44
asynchronous replication 59, 110, 112, 145, 180–181, 199

auto mode 88
automatic negotiation 22
autosensing 68, 70–71, 74, 88
availability 200

B

backbone fabric 24, 30, 42
backplane 72
backup
 consolidation 208
 infrastructure 118
bandwidth 45, 152, 192
 management 192
 requirement 194
best practices 39, 125
block I/O 114
boot 21
booting iSCSI 8
broadcast 94
buffer 157
 credits 36, 74–75, 79, 100
 overflow 162, 168, 192
bus 83
business continuance 69

C

cabling 193
call home 80
capacity 145, 194
capacity planning 194
cascading 88
central arbiter 73
centralized management 85, 132
Challenge Handshake Authentication Protocol (CHAP) 44
CHAP (Challenge Handshake Authentication Protocol) 44
CHAP secret 44
chassis 70–71
checksum 165
Cisco
 best practices 125
 domain 102

- family routing products 65
- real-life solutions 137
- solutions 109
- Cisco Fabric Manager 77, 80–81, 83
- Cisco IOS CLI 80
- Cisco IVR 97
- Cisco Mainframe Package 88
- Cisco MDS 9000 75, 80–82, 89–90, 95, 99, 101–102, 106–107
- Cisco MDS 9000 Fabric Manager 80
- Cisco MDS 9000 Multilayer switches 65
- Cisco MDS 9200 switches 75
- Cisco MDS 9216 Multilayer Fabric Switch 88
- Cisco MDS 9216i Multilayer Switch 69
- Cisco MDS 9500 68
 - directors 75
- Cisco MDS 9506 Multilayer Director 70, 88
- Cisco MDS 9509 Multilayer Director 88
- Cisco SAN-OS 79
 - CLI 88
- Cisco Systems 66
- Cisco VSAN 93
- Class F frame 101
- Class F traffic 172
- CLI (command-line interface) 24, 80, 83, 132, 157
- client 80
- clock 41, 72
 - module 72
- cluster 36
- coarse wavelength division multiplexing (CWDM) 68, 70, 74–76, 88, 110
- command-line interface (CLI) 24, 80, 83, 132, 157
- communication 89
- compact 21
- CompactFlash 69, 140
- compression 3, 6, 31, 84, 127, 134, 152, 161, 181
 - algorithms 9
 - ratio 168
- configuration 80, 89, 91, 95–96, 105–106
- configure 75, 105
- congestion 101, 197
 - control 3, 5, 9–10
 - control methods 101
- connecting hosts 16
- connection 1
 - allegiance 166
 - costs 109, 126, 177
- connectivity 23, 72, 74, 82
- consolidate 52, 208
- consolidation 36, 51, 180
- control 2, 69–72, 101
- control engine 130
- Control Unit Port (CUP) 87
- cooling 21
- cooling fan 23
- core network 126
- Core PID 27, 35, 42, 179, 194
- corporate subsidiary separation 184
- cost 16, 75
- CRC 2
- credit flow 48
- crossbar
 - fabric 130
 - switches 75
 - switching fabric 72–73
- CUP (Control Unit Port) 87
- CWDM (coarse wavelength division multiplexing) 68, 70, 74–76, 88, 110

D

- dark fiber 37
- data integrity 165
- data streaming 14
- data traffic 96, 103
- datagram 163
- daughter board 21
- dedicated link 192
- default 88, 94–95, 101, 107
 - VSAN 94–95
 - zone 106
- deflate 168
- delay 10, 195
- deleted VSAN 96
- delimiter
 - end of frame 2
 - start of frame 2
- dense wavelength division multiplexing (DWDM) 16, 37, 110, 180
- destination ID (DID) 98, 102
- device 89
- device management tool 80
- Device Manager 88
- device view 81
- diagnostics 79
- DID (destination ID) 98, 102
- Differentiated Service Code Point (DSCP) 162
- director 66, 68, 72, 75, 93, 97, 106–107

- failure 220
- disaster recovery (DR) 58, 69, 215
- discovery 8
- disruptive domain reconfiguration 107
- distance 195
 - connections 45
 - disaster recovery over IP 215
 - limitations 5
 - limited capabilities 7
- Distributed Services Time-Out Value timers 106
- distributing traffic 98
- documentation 193
- domain
 - Cisco 102
 - manager 92–93, 107
 - reconfiguration disruptive 107
- domain ID 27, 106–107, 172
- downtime 57
- DR (disaster recovery) 58, 69, 215
- driver 193
- dropped packets 192, 197
- DSCP (Differentiated Service Code Point) 162
- dual physical fabrics 138
- dual redundant power supplies 54
- dual redundant supervisor modules 70–71
- DWDM (dense wavelength division multiplexing) 16, 37, 110, 180

E

- E_D_TOV 106
- E_Port 88–90, 94, 99, 104, 106–107, 170, 178
- E_Port mode 88
- edge fabric 24, 34
- edge quench control 101
- egress
 - direction 103
 - port 197
 - source 103
 - traffic 104, 168, 172
- EISL (extended ISL) 77, 90, 98
 - frame 99
- Element Manager, SAN Router 155, 159
- encapsulated 4
- encapsulating 3
- encapsulation 5, 45
- encryption 132
- end of frame (EOF) delimiter 2
- EOF (end of frame) delimiter 2

- equal cost 173
- Error Detect Time-Out Value timers 106
- error detection 5
- Ethernet (out-of-band) connection 81
- Ethernet connection 83
- Ethernet MAC 196
- EX_Port 24
- exchange-based load balancing 98
- exchange-level trunking 44
- expansion port 89–90
- Expedited Forwarding 162
- exported nodes 25
- extended distances 78
- extended ISL (EISL) 77, 90, 98–99
- extension 36, 126, 177

F

- F_Port 88–89, 104
- fabric 3, 24, 66, 68, 80, 89, 92–93, 95, 101, 106
 - addressing schemes 3
 - build 170
 - crossbar
 - switching 73
 - expansion port 89
 - extension with FC-FC routing 36
 - failover 131
 - integrated crossbar switching 72
 - isolation 116
 - level 170
 - login 6
 - planning 160
 - reconfiguration 161
 - services 92
 - view 81
- Fabric Application Interface Standard (FAIS) 78
- Fabric ASIC 22
- fabric ID 24
- fabric management 80, 89
- Fabric Manager 24, 77, 80–81, 83, 88
- fabric manager 170
- Fabric Manager Server (FMS) 84–85
- Fabric Shortest Path First (FSPF) 92–93, 170, 173, 175
- Fabric Stability Time-Out Value timers 106
- fabric-wide elements 83
- failover 82
- FAIS (Fabric Application Interface Standard) 78
- Fast Ethernet 152

- Fast LZO 167
- Fast Write 11, 31, 152, 161–162
- fault isolation 201
- FC fabric support 156
- FC frame 2
- FC tape acceleration 127
- FC_AL 87
- FC_XFER_RDY 164–165
- FC-AL (Fibre Channel Arbitrated Loop) 113
- FCC 79, 101
 - process 102
- FC-FC 4–5, 40
- FC-FC routing 15, 23, 34, 178
 - for fabric extension 36
 - performance 44
 - security 43
 - solution 178
- FCIP 3–5, 40, 69, 76–77, 84, 87–88, 127, 140
 - compression 111, 134
 - deployment 141
 - link 62
 - link sizing 61
 - performance 111
 - tunnel 27, 77
 - tunneling 16, 23, 27, 31, 69, 145, 150
 - tunneling performance 45
- FCIP Activation 77, 85
- FCIP Tape Acceleration 84
- FCIP Write Acceleration (FCIP-WA) 78, 84, 112, 127, 134
- FCIP-WA (FCIP Write Acceleration) 78, 84, 112, 127, 134
- FC-NAT (Fibre Channel network address translation) 3, 25
- FCP (Fibre Channel Protocol) 79, 87
- fctrace 90
- FC-WA (FCIP Write Acceleration) 78
- fd (front domain) 25
- FDMI 80
- features 153
- Fibre Channel 2, 7, 68–71, 76–77, 82, 88, 98, 102, 104, 106
 - analyzer 90
 - attached targets 77
 - director 92
 - frame 83
 - interface 90, 102
 - ports 157
 - protocol 82
 - router hardware 130
 - routers 3
 - secure router port 202
 - switch support 23
 - switching 3
 - tape acceleration 113
 - trace feature 90
 - traffic 90
 - tunnel 90
- Fibre Channel Arbitrated Loop (FC-AL) 113
- Fibre Channel network address translation (FC-NAT) 3, 25
- Fibre Channel over IP 4
- Fibre Channel Protocol (FCP) 79
- FICON 87–88
- FICON cascaded 87
- FICON Control Unit Port (CUP) 87
- FID 24
- firewalls 43
- firewire 7
- firmware 68–69, 193–194
- FL_Port 88–89, 104
- flexible fabric switch 68
- FLOGI 6
- flow control 102, 168, 192
- FMS (Fabric Manager Server) 84–85
- forward congestion control 101
- forwarding tables 3
- forwards packets 3
- frame 2, 90, 98, 101
 - Class F 101
 - E_Port 89
 - Fibre Channel 83
- frame level trunking 44
- frame size 10, 47
- frame structure 2
- frame-based algorithm 167
- frames 102
- front domain (fd) 25, 28
- front domain ID 28
- front to rear airflow for cooling 67
- FSPF (Fabric Shortest Path First) 92–93, 170, 173, 175
- FSPF routing 26
- FX_Port 89

G

- gateway 41, 77

- gateway-to-gateway 5
- geographically distributed 5
- Gigabit Ethernet 69, 152
 - ports 76
- Global Mirror 62
- Global Mirroring 147
- guaranteed bandwidth 46

H

- handshakes 164
- hardware 66, 76, 96
 - limitations 200
 - selection 133
- HBA (host bus adapter) 54, 209
- header 2–3
- heartbeat 200
- heterogeneous interconnection 129
- heterogeneous IVR 129
- high availability 70–71, 73, 80, 99
- high latency network 112
- high priority status 101
- higher compression 167
- historical performance 85
- hop count 26
- hops 129
- host bus adapter (HBA) 54, 209
- host optimized ports 67
- hot code activation 42
- hot code load 42
- hot swappable fan tray 68, 71
- hot-swappable FRU 21

I

- I/O block sizes 127
- I/O response time 127
- IBM TotalStorage
 - b-type family real-life routing solutions 51
 - b-type family routing best practices 39
 - b-type family routing products 19
 - b-type family routing solution 39
 - b-type family routing solutions 33
 - m-type family best practices 191
 - m-type family real-life routing solutions 207
 - m-type family routing products 151
 - m-type family solutions 177
- IBM TotalStorage SAN16B-R 20
- IBM TotalStorage SAN16M-R 151
- IEEE-1394 7

- iFCP 5, 152, 156–157, 160–162, 167, 196, 218, 220
 - compression 181
 - conversion 16
 - Fast Write 182
 - link failure 220
 - link sizing 218
 - path 200
- iFCP (Internet Fibre Channel Protocol) 3, 180, 196, 214
- IFL (inter-fabric link) 24–25, 34, 53, 56, 213
 - failure 54, 211
- implicit transfer 95
- in-band 82
- in-band management 81
- incoming transfers 7
- infrastructure simplification 20
- ingress
 - direction 103, 105
 - source 103
 - source port 104
 - traffic 104, 168
- initial IP address configuration 22
- initialization 88–89
- initiator 7, 77
- initiator session ID (ISID) 7, 166
- in-order delivery of Fibre Channel frames 49
- integrated crossbar switching fabric 72
- integration 1
- integrity control 13
- Intelligent Peripheral Interface (IPI) 7
- intelligent ports 157
- intelligent storage services 78
- interconnected 192
- interconnection 5
- inter-fabric link (IFL) 24–25, 34, 53–54, 56, 211, 213
- Internet Fibre Channel Protocol (iFCP) 3, 180, 196, 214
- Internet Storage Name Server (iSNS) 6
- Internet Storage Name Service (iSNS) 8
- internetworking 152, 156
- interoperability 35, 88, 109, 168, 177
 - matrix 107
 - mode 105–107, 168
- inter-switch link (ISL) 5, 24, 70, 74–75, 89, 98–99, 106, 131
- Inter-VSAN Routing (IVR) 4, 69, 80, 84, 96, 115, 119, 121, 126–127, 131, 147

- Inter-VSAN Routing with FCIP 111
- investment 52, 209
- IP address 41, 82–83
- IP backbone 141
- IP connectivity 157
- IP drivers 83
- IP line card 77
- IP packets 3
- IP routers 3
- IP services 88
- IP storage services 71–72
- IP-based Global Mirroring 20
- IPFC 83
- IPI (Intelligent Peripheral Interface) 7
- IPS 88, 111
 - ACLs 80
- IPSec 192, 201
- IQN 44
- IRL 160, 162, 170
- iSAN 160, 170, 192, 194
- iSCSI 6–7, 16, 40, 44, 76–77, 79, 88, 114, 127, 140, 152, 156, 182
 - adapter 23
 - booting 8
 - client 42
 - connection 16
 - deployment 141
 - discovery 8
 - drafts 8
 - driver 23
 - gateway 23, 37, 43, 165
 - gateway security 44
 - immediate data 115
 - initiator 44
 - initiator authentication 44
 - initiator name 171
 - low-cost connection 113
 - name 166
 - naming 8
 - packet 7
 - portal 45
 - protocol 7
 - qualified name 44
 - router 16
 - solution 127
- iSCSI (Small Computer System Interface over IP) 3
- ISID (initiator session ID) 7, 166
- ISID/TSID session pair 166
- ISL (inter-switch link) 5, 24, 70, 75, 89, 98–99, 106, 131
 - connections 74
- iSNS (Internet Storage Name Server) 6
- iSNS (Internet Storage Name Service) 8
- isolated 66, 94, 106
- isolated VSAN 94–95
- isolation 1, 109, 115, 126, 177
 - and interoperability using IVR 115
 - multivendor switches and modes 119
- IVR (Inter-VSAN Routing) 4, 69, 80, 84, 96, 119, 126–127, 147
 - and VSAN 131
 - isolation and interoperability 115
 - storage migration 121

J

- Java 80
- jumbo frame 11, 45, 47, 196–197
- jumbo IP packet 9
- jumbo packet 111, 181

K

- kernel 21

L

- LAN-free tape backups 52
- latency 8, 10–11, 16, 45, 47, 61, 79, 101, 110, 142–143, 161, 182, 192, 195
- license key 87
- limitations 203
- link
 - bandwidth 46, 195
 - bounce 111
 - latency 10, 46, 195
 - sizing 198
 - speed 10, 47
- Link Aggregation 162
- load balance 98, 131
- load balancing 95, 98
 - traffic 70–71
- local FC-FC routing 34
- location 80, 82
- locking 80
- logic control 70–71
- login negotiation 166
- long distance disaster recovery 58, 215
- long distance network 11

- longwave FC connection 141
- loop 89
- low-cost connection with iSCSI 113
- LSAN 24, 29, 40, 43, 57, 62
 - configuration 35
 - zone 54
- LUN 198
 - access 110
- LZO 167
 - algorithm 167
 - with history 168

M

- MAC (medium access control) 82
- management xvi, 69, 72, 82–83, 91, 95, 97
 - in-band 81
 - out-of-band 81
 - port 22, 153, 157
- maximum transmission unit (MTU) 162, 192
- McDATA Eclipse 1620 SAN Router 152
- McDATA Eclipse 2640 SAN Router 156
- McDATA Open Fabric 168
- McDATA Open Fabric Mode 168, 173
- MDS 9000 75, 80–82, 88–90, 95, 98–99, 101–102, 106–107
 - advanced management 79
- MDS 9216 66, 68, 73–74
- MDS 9506 66, 74
- MDS 9509 66, 71, 74, 97, 107
- medium access control (MAC) 82
- Meta SAN 24, 26–27, 30, 40
- Metro Fibre Channel Protocol (mFCP) 153, 157, 161, 200, 203
- Metro Simple Name Server (mSNS) 170–171
- mFCP (Metro Fibre Channel Protocol) 153, 157, 161, 200
 - link 203
- Microsoft iSCSI initiator 42
- migrate data 186
- migration 1, 57, 188
 - new storage environment 211
- mode 74, 82, 88–89, 102, 105, 107
 - settings 119, 185
- modular
 - basis 120
 - chassis 75
- monitoring 73, 80, 90, 102
- MPLS (Multiprotocol Label Switching) 49, 61, 218

- MPS (Multiprotocol Services) 88
- mSAN 153, 160, 168, 170, 178, 188–189, 194, 203
 - zone 178, 209
- mSAN zone 209
- mSNS (Metro Simple Name Server) 170
 - database 171
 - Keyed Query Service 171
 - registration service 171
 - State Change Notification service 171
- MTU (maximum transmission unit) 162, 192
- m-type router 200
- multicast 94
- multiple board design 21
- multiple paths 26, 200
- multiple paths on router level 200
- multiple VSANs 93
- multiplexers 110
- Multiprotocol Label Switching (MPLS) 49, 61, 218
- multiprotocol ports 22
- multiprotocol routing 20
- Multiprotocol Services (MPS) 88
- multiprotocol switch/router products 3

N

- N_PORTS 2
- name server 96, 107, 170
- naming 8
- NAS gateway 113
- NAT (network address translation) 3, 29, 173
- native mode 42, 173
- negotiation 20
- network address translation (NAT) 3, 29, 173
- network interface 45
- network interface card (NIC) 7
- network traffic 90
- NIC (network interface card) 7
- NL_Port 89
- node worldwide name (nWWN) 170, 202
- non-blocking connectivity 22
- nondisruptive restart 73
- nondisruptive software upgrade 130
- nondisruptive switchover 73
- nondisruptively 107
- nonintrusive 90
- non-jumbo router 196
- non-LSAN 57
- non-trunking ports 95
- NR_Port 25–26

nWWN (node worldwide name) 170, 202

O

Open Fabric Mode 168, 173
operational state of a VSAN 96
optimization 167
Organizational Unique Identifier (OUI) 170
originator exchange ID (OXID) 95, 98
OUI (Organizational Unique Identifier) 170
outgoing transfer 7
out-of-band (Ethernet) connection 81
out-of-band management 81
out-of-order 3
out-of-order packet delivery 49
oversubscription 133
OXID (originator exchange ID) 95, 98

P

packet 3–4, 6, 9, 162
 delivery out of order 49
 drop 197
 loss 45, 48, 192
 segments 9
 size 3, 9
 transmission times 11
parallel SCSI 7
parameters 90, 95
partitioning 171
passive optical mux 70, 74–76, 88
path 3, 95, 102
 cost 173
 failover 200
 selection 170, 175
path quench control 101
payload 2, 10, 47, 115, 134, 196
 compression 6
PCB (printed circuit board) 21
PCI bus 21
peak workloads 133
performance 44, 75, 81, 133, 145
 degradation 48, 57
peripheral local bus 21
phantom domains 25
phantom link 25
phase collapse 7
PID (port ID) 26
pilot solution 193
piloting new technology 127

port 67–72, 74, 76–77, 79, 88, 90, 93–94, 97–98, 104, 201
 addresses 25
 addressing 88
 density 75
 failure 221
 groups 67, 75
 modes 88
 speeds 133
 types 88, 90
 VSAN membership 95
port ID (PID) 26
PortChannel 80, 89, 98–99, 104
PortChanneling 98–99, 131
power 21
 failure 54, 210, 219
 supplies 71–72
primary mSNS 171, 174
principal domains 25
printed circuit board (PCB) 21
priority 101
private loop 87, 89
privileges 83
probe 90
problem determination 43
propagation delay 10
protocol conversion 3
public arbitrated loop 89
public loop 89
pWWN zoning 202

Q

QoS (Quality of Service) 6, 80, 98, 100, 128, 162, 192
 priority 100
Quality of Service (QoS) 6, 80, 98, 100, 128, 162, 192
quench message 102

R

R_A_TOV 106
R_Port 152, 156, 160, 170, 173, 178
R_RDY 100, 113
radius 80
rate limiting 162, 168
RBAC 79
ready signal 11
receiver ready 100

- recovery 5
- recovery point 147
- recovery point objective (RPO) 216
- recovery time objective (RTO) 59, 64, 216
- Redbooks Web site 245
 - Contact us xix
- redundancy 27, 70–72, 92, 130, 156
- redundant 2109-A16s 34
- redundant fabrics 130
- redundant Fibre Channel connections 178
- redundant ISLs 131
- redundant power supplies 68, 71–72
- Registered State Change Notification (RSCN) 92, 94, 120, 161, 171, 201
- reliability 110, 138, 192
- remote fabric 25
- remote locations 192
- remote mirroring 78
- remote site connection over IP 16
- remote SPAN (RSPAN) 90
- renegotiate 5
- replication 1, 69, 143
- Resource Allocation Time-Out Value timers 106
- resource sharing 1
- retransmission 197
- retry-tolerant 5
- right-to-left cooling 67
- RMON 80
- role-based administration 132
- role-based management 83
- role-based security methodology 132
- round 143
- round trip 46, 61
 - delay 8
 - latency 46
- round-robin algorithm 173
- round-robin database (RRD) 86
- round-trip
 - delay 10
 - link latency 10
- round-trip time (RTT) 46, 148, 154, 158, 195
- route frames 2
- routed
 - connection 41
 - fabric 30
 - FCIP 41
 - SAN 160
- routed networks 2
- router 3, 116

- configuration 194
- failure 211, 220
- internal network architecture 169
- management 193
- router hardware 130
- routing 1, 170
 - capabilities 20
 - concepts 19
 - domains 172
 - services 168
 - tables 99
- RPO (recover point objective) 216
- RRD (round-robin database) 86
- RSCN (Registered State Change Notification) 92, 94, 120, 161, 171, 201
- RSPAN (remote SPAN) 90
- RTO 221
- RTO (recovery time objective) 59, 64, 216
- RTT (round-trip time) 46, 148, 154, 158, 195
- running configuration 96

S

- SACK (Selective ACKnowledgment) 167
- SAN 129
 - availability 200
 - islands 58
 - router 156
 - routing 1
 - routing architecture 168
- SAN (SAN Volume Controller) 36, 129, 140
- SAN extension 109, 181
 - with FCIP 110
 - with iFCP 180
- SAN Extension over IP Package 85
- SAN Extension Tuner 85, 111, 134, 148
- SAN Router Element Manager 155, 159
- SAN Volume Controller (SVC) 36, 126, 129, 140, 180
- SAN04M-R 151–152
- SAN16B-R 20
- SAN16M-R 154, 156, 158
- SAN-OS 79
- SANTap 78
- SANvergence Management 152
- Saturn processor 167
- scalability 34, 177–178, 203
- SCSI 7
 - commands 7

- packets 7
- protocol 7
- write operation 164
- SCSI (Small Computer Systems Interface) 7
- SD_Port 90
- SDH 110, 180
- SDRAM 21
 - controller 21
- secondary mSNS 170, 174
- secondary supervisor 73
- security 76, 88, 131, 192, 201
 - centralized management 132
 - features 43
 - mechanism 132
- Selective ACKnowledgment (SACK) 167
- sendtargets command 8
- separate fabric services 119, 185
- separate SAN fabrics 52
- serial I/O bus 2
- serial port 22
- server 80, 88, 97, 107
- serverless backup 78
- Service Location Protocol (SLP) 8
- service-level agreement (SLA) 10, 40, 47, 196
- SFP (small form-factor pluggable) fiber optic transceiver 67–68, 70–71, 88
- SFTP 79
- share resources 20
- shared access 118
- shared link 192
- SID (source ID) 98, 102
- SilkWorm 42
- simple name server (SNS) 170
- Simple Network Management Protocol (SNMP) 80
- single point of failure 73
- sizing 147
- SLA (service-level agreement) 10, 40, 47, 196
- slots 69–72, 74
- SLP (Service Location Protocol) 8
- Small Computer System Interface over IP (iSCSI) 3
- Small Computer Systems Interface (SCSI) 7
- small form-factor pluggable (SFP) fiber optic transceiver 67–68, 70–71, 88
- SMI-S 80
- SNMP (Simple Network Management Protocol) 80
- SNMPv3 79
- SNS (simple name server) 170
- socket 4
- SOF (start of frame) delimiter 2
- SONET 110, 180
- source ID (SID) 98, 102
- source interface 103
 - types 104
- SPAN 79, 90, 94, 102–104
- span destination 102
- SPAN source 103
- speed 20, 73
- speed of light 46
- SSH 79
- SSM (Storage Services Module) 78
- standard write request 12
- standards 2
- standby supervisor 70–71, 82
- start of frame (SOF) delimiter 2
- statistics monitoring 85
- status, high priority 101
- storage environment migration 211
- storage migration using IVR 121
- Storage Services Module (SSM) 78
- streaming 13
- subfabrics 15
- subnet mask 41
- Summary View 81
- supervisor 73, 79, 104
- supervisor module 69–73, 81–82, 101, 104
- suspended state 95
- SVC (SAN Volume Controller) 36, 126, 129, 180
- SW_RSCN 94
- switch 3
 - fabric 103–104
 - failure 220
 - interoperability 105
 - RSCNs 94
- switch port 90, 102
 - analyzer 90
- switchover nondisruptive 73
- Symmetric Flow Control 162
- synchronous replication 110, 142
- system design 133

T

- TACACS+ 80
- tape acceleration 13, 84
- tape data backup 199
- target 7
- Target Portal Group Tag 7
- target session ID (TSID) 166

- target-optimized ports 140
- TCP congestion 9
- TCP receive window 48
- TCP Selective ACKnowledgment (SACK) 167
- TCP stack 161
- TCP/IP offload engine (TOE) 17, 114, 127, 183
- TE_Port 88, 90, 99, 104, 107
- thresholds 85
- time of frame 10, 47
- time of frame in transit 196
- time server 41
- Time-Out Value timers 106
- Tivoli Storage Manager 209
 - server 211
- TL_Port 89, 104
- TOE (TCP/IP offload engine) 17, 114, 127, 183
- tools 1
- topology 77
- topology discovery 80
- total latency 195
- trace feature 90
- traffic 71, 73, 77, 90–91, 93–94, 96, 102–104
 - congestion 10
 - Fibre Channel 90
 - hotspot 85
 - isolation 98
 - load balancing 70
 - profile 112
 - shaping 45–46
- traffic load balancing 71
- transfer ready 13
- transfer ready message 164
- transit 196
- transit VSAN 111, 142, 150
- translate domain 25–26
- translate domain (xlate) 25
- translate domain IDs 40
- translative loop 89
- transmission times 197
- transport 7
- transport latency 112
- Triple Data Encryption Standard 132
- tri-rate SFPs 76
- trunking 90, 99, 106, 152, 156
 - E_Port 90
 - port 95, 99
- TSID (target session ID) 166
- tuning 127
- tunnel 4–5

- tunneling 3–4
 - services 3
 - storage 9

U

- unanswered packets 9
- uncertain state 113
- unicast 94
- unplanned access 35
- upstream 101
- user-defined VSAN 95

V

- value proposition for SAN routing 1
- VE_Port 27
- virtual E_Port 27
- virtual fabrics 4, 15
- virtual ISL 77
 - connections 77
- virtual links 25
- virtual output queuing 79
- virtual SAN 66, 91, 93
- virtual slot 26
- virtual slot numbers 40
- virtual target 111, 165
- virtualization 138
- voltages 21
- VPN 201
- VSAN 4, 66, 79, 81, 88, 90–91, 93–97, 99, 101–102, 104–107, 127, 131, 141, 147
 - administration 132
 - as a SPAN source 103
 - attributes 95
 - deleted 96
 - manager 96
 - name 95
 - operational state 96
 - trunking 90, 99
- VSAN 4094 94–95
- VSAN ID 95
- VSAN-based runtime 96

W

- WAN (wide area network) 69, 192
- wide area network (WAN) 69, 192
- wire speeds 5
- worldwide name (WWN) 43, 54, 77, 94, 168, 171,

209
write acceleration 11, 13, 31, 85
write acknowledgement 13
WWN (worldwide name) 43, 54, 77, 94, 209
168, 171
WWPN 106

X

xlate (translate domain) 25
XML-CIM 80
XPath Fabric ASIC 21–22
XPath OS 23
XPath OS version 7.3 30

Z

zone 24, 28, 94, 106
zoned 66
zoning 15, 62, 77, 79–80, 92–94, 96, 106, 171, 202
zoning definitions 96



IBM TotalStorage: Introduction to SAN Routing

(0.5" spine)
0.475" <-> 0.875"
250 <-> 459 pages



IBM TotalStorage: Introduction to SAN Routing



Redbooks

**Uncover the basics
of the IBM approach
to multiprotocol
SANs**

**Learn about the IBM
TotalStorage
multiprotocol
portfolio**

**Explore routing
products and
solutions**

The rapid spread and adoption of production storage area networks (SANs) has fuelled the need for multiprotocol routers. The routers provide improved scalability, security, and manageability by enabling devices in separate SAN fabrics to communicate *without* merging fabrics into a single, large SAN fabric. This capability allows clients to deploy separate SAN solutions at the departmental and data center levels. Then clients can consolidate these separate solutions into large enterprise SAN solutions as their experience and requirements grow and change.

Alternatively multiprotocol routers can help to connect existing enterprise SANs. For instance, the introduction of Small Computer System Interface over IP (iSCSI) provides connection of low-end, low-cost hosts to enterprise SANs. Using an Internet Protocol (IP) in a Fibre Channel (FC) environment provides resource consolidation and disaster recovery planning over long distance. And using FC-FC routing services provides connectivity between two or more fabrics without merging them into a single SAN.

This IBM Redbook targets storage network administrators, system designers, architects and IT professionals who are engaged in the selling, designing, or administration of SANs. It introduces you to the products, concepts, and technology in the IBM TotalStorage SAN Routing portfolio. It shows the features of each product and examples of how you can deploy and use them.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks