

IBM System Storage SAN Volume Controller and Storwize V7000 Best Practices and Performance Guidelines

Jon Tate

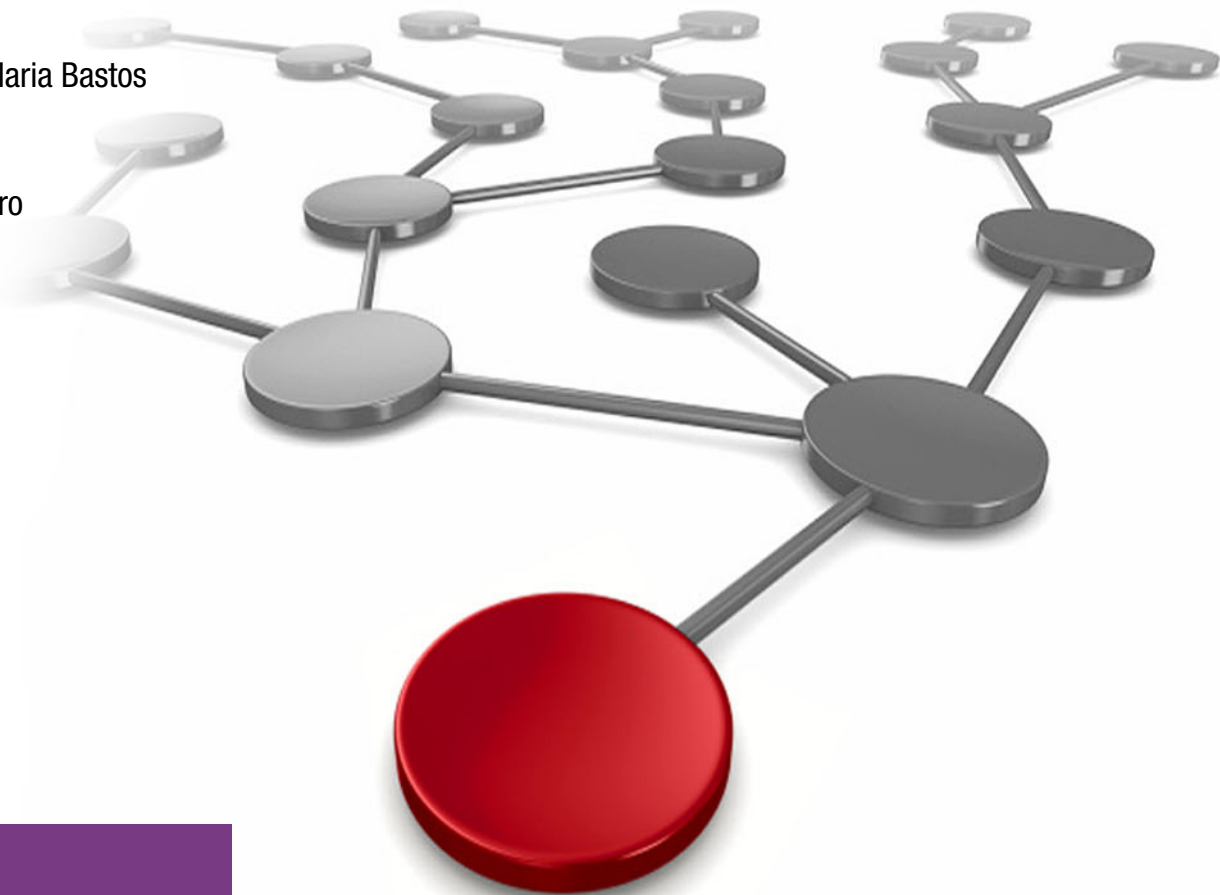
An Chen

Tiago Moreira Candelaria Bastos

Jana Jamsek

Danilo Morelli Miyasiro

Antonio Rainero



Storage



International Technical Support Organization

IBM System Storage SVC and Storwize V7000 Best Practices and Performance Guidelines

May 2018

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

Sixth Edition (May 2018)

This edition applies to IBM Spectrum Virtualize V8.1, and the associated hardware and software detailed within. Note that screen captures might differ from the generally available (GA) version, because parts of this book were written with pre-GA code.

© Copyright International Business Machines Corporation 2008, 2018. 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

Notices	xi
Trademarks	xii
Preface	xiii
Authors	xiii
Now you can become a published author, too!	xv
Comments welcome	xvi
Stay connected to IBM Redbooks	xvi
Summary of changes	xvii
April 2018, Sixth Edition	xvii
Chapter 1. Storage area network	1
1.1 SAN topology general guidelines	2
1.1.1 SAN performance and scalability	2
1.1.2 ISL considerations	3
1.2 SAN topology-specific guidelines	4
1.2.1 Single switch SAN Volume Controller/Storwize SANs	5
1.2.2 Basic core-edge topology	6
1.2.3 Edge-core-edge topology	6
1.2.4 Full mesh topology	8
1.2.5 IBM Spectrum Virtualize and IBM Storwize as a SAN bridge	8
1.2.6 Device placement	9
1.3 SAN Volume controller ports	11
1.3.1 Slots and ports identification	12
1.3.2 Port naming and distribution	12
1.4 Zoning	16
1.4.1 Types of zoning	16
1.4.2 Prezoning tips and shortcuts	18
1.4.3 SAN Volume Controller/Storwize internode communications zones	19
1.4.4 SAN Volume Controller/Storwize storage zones	20
1.4.5 SAN Volume Controller/Storwize host zones	35
1.4.6 Hot Spare Node zoning considerations	40
1.4.7 Zoning with multiple SAN Volume Controller/Storwize clustered systems	41
1.4.8 Split storage subsystem configurations	42
1.5 Distance extension for remote copy services	42
1.5.1 Optical multiplexors	42
1.5.2 Long-distance SFPs or XFPs	42
1.5.3 Fibre Channel over IP	42
1.5.4 SAN extension with Business Continuity configurations	44
1.5.5 Native IP replication	46
1.6 Tape and disk traffic that share the SAN	47
1.7 Switch interoperability	47
Chapter 2. Back-end storage	49
2.1 Storage controller path selection	50
2.1.1 Round robin	50
2.1.2 MDisk group balanced and controller balanced	51

2.2	Considerations for DS8000 series	52
2.2.1	Connectivity considerations	53
2.2.2	Defining storage	53
2.3	Considerations for IBM XIV Storage System	59
2.3.1	Connectivity considerations	59
2.3.2	Host options and settings for XIV systems	59
2.3.3	Volume considerations	60
2.3.4	Additional considerations	62
2.4	Considerations for IBM FlashSystem A9000/A9000R	63
2.4.1	Connectivity considerations	63
2.4.2	Volume considerations	63
2.4.3	Additional considerations	65
2.5	Considerations for IBM Storwize V7000/V5000/V3700	66
2.5.1	Connectivity considerations	66
2.5.2	Defining internal storage	67
2.5.3	Volume considerations	69
2.6	Considerations for IBM FlashSystem 900	70
2.6.1	Connectivity considerations	70
2.6.2	Defining storage	70
2.6.3	Volume considerations	71
2.7	Considerations for storage subsystem compression and deduplication capability	73
2.8	Considerations for third-party storage with EMC VMAX and Hitachi Data Systems	74
Chapter 3. Storage pools and managed disks		75
3.1	Availability considerations for storage pools	76
3.2	Selecting storage subsystems	77
3.3	Selecting the storage pool	77
3.3.1	Capacity planning consideration	78
3.3.2	Selecting the number of arrays per storage pool	78
3.3.3	Selecting LUN attributes	79
3.4	Quorum disk considerations	80
3.4.1	IP Quorum	82
3.4.2	IP Quorum requirements	83
3.5	Tiered storage pool	85
3.6	Adding MDisks to existing storage pools	89
3.6.1	Checking access to new MDisks	89
3.6.2	Persistent reserve	89
3.6.3	Renaming MDisks	89
3.7	Rebalancing extents across a storage pool	90
3.8	Removing MDisks from existing storage pools	90
3.8.1	Migrating extents from the MDisk to be deleted	90
3.8.2	Verifying the identity of an MDisk before removal	91
3.8.3	Correlating the back-end volume with the MDisk	91
3.9	Remapping managed MDisks	96
3.10	Controlling extent allocation order for volume creation	97
3.11	Considerations when using Encryption	97
3.11.1	General considerations	97
3.11.2	Hardware and software encryption	98
3.11.3	Encryption at rest with USB keys	101
3.11.4	Encryption at rest with key servers	102

Chapter 4. Volumes	109
4.1 Overview of volumes	110
4.2 Guidance for creating volumes	110
4.3 Striped versus sequential volumes	114
4.3.1 Use cases of sequential volumes	114
4.4 Thin-provisioned volumes	114
4.4.1 Compressed volumes	115
4.4.2 Space allocation	118
4.4.3 Thin Provisioning considerations	119
4.4.4 Limits on virtual capacity of thin-provisioned volumes	123
4.5 Volume migration	123
4.5.1 Image-type to striped-type migration	124
4.5.2 Migrating to image-type volume	124
4.5.3 Migrating with volume mirroring	125
4.6 Preferred paths to a volume	126
4.7 Changing the preferred node within or across I/O groups	127
4.8 Volume throttling	128
4.8.1 Managing throttles for volumes	129
4.9 Volume cache mode	130
4.9.1 Changing the cache mode of a volume	130
4.9.2 Underlying controller remote copy with IBM Spectrum Virtualize and Storwize cache-disabled volumes	131
4.9.3 Using underlying controller FlashCopy with IBM Spectrum Virtualize and Storwize cache-disabled volumes	133
4.9.4 Using IBM Spectrum Virtualize or Storwize with FlashSystem	133
4.10 VMware Virtual Volumes	135
4.11 Additional considerations	136
4.11.1 Volume protection	136
4.11.2 Volume resize	136
Chapter 5. Copy services	139
5.1 Introduction to copy services	140
5.1.1 FlashCopy	140
5.1.2 Metro Mirror and Global Mirror	140
5.1.3 Global Mirror with Change Volumes	140
5.1.4 Volume Mirroring function	140
5.2 FlashCopy	141
5.2.1 FlashCopy use cases	141
5.2.2 FlashCopy capabilities overview	143
5.2.3 FlashCopy functional overview	149
5.2.4 FlashCopy planning considerations	154
5.3 Remote Copy services	163
5.3.1 Remote copy functional overview	164
5.3.2 Remote Copy network planning	177
5.3.3 Remote Copy services planning	187
5.3.4 Remote Copy use cases	198
5.3.5 1920 error	202
5.4 Native IP replication	214
5.4.1 Native IP replication technology	214
5.4.2 IP partnership limitations	216
5.4.3 VLAN support	217
5.4.4 IP Compression	218
5.4.5 Remote copy groups	219

5.4.6	Supported configurations	221
5.4.7	Native IP replication performance consideration	229
5.5	Volume Mirroring	231
5.5.1	Read and write operations	232
5.5.2	Volume mirroring use cases	232
5.5.3	Mirrored volume components	234
5.5.4	Volume Mirroring synchronization options	235
5.5.5	Volume Mirroring performance considerations	236
5.5.6	Bitmap space for out-of-sync volume copies	238
Chapter 6.	Hosts	241
6.1	Configuration guidelines	242
6.1.1	Host levels and host object name	242
6.1.2	Host cluster	242
6.1.3	The number of paths	243
6.1.4	Host ports	244
6.1.5	Port masking	245
6.1.6	Host to I/O group mapping	246
6.1.7	Volume size as opposed to quantity	246
6.1.8	Host volume mapping	247
6.1.9	Server adapter layout	251
6.2	N-Port ID Virtualization	251
6.3	Host pathing	254
6.3.1	Multipathing Software	254
6.3.2	Preferred path algorithm	255
6.3.3	Path selection	255
6.3.4	Path management	256
6.3.5	Non-disruptive volume migration between I/O groups	256
6.4	I/O queues	259
6.4.1	Queue depths	259
6.5	Host clustering and reserves	259
6.5.1	Clearing reserves	260
6.5.2	IBM Spectrum Virtualize MDisk reserves	261
6.6	AIX hosts	261
6.6.1	HBA parameters for performance tuning	261
6.6.2	Configuring for fast fail and dynamic tracking	263
6.6.3	SDDPCM	263
6.7	Virtual I/O Server	264
6.7.1	Methods to identify a disk for use as a virtual SCSI disk	265
6.7.2	UDID method for MPIO	265
6.8	Windows hosts	265
6.8.1	Clustering and reserves	266
6.8.2	Tunable parameters	266
6.8.3	Guidelines for disk alignment using Microsoft Windows with IBM Spectrum Virtualize volumes	266
6.9	Linux hosts	266
6.9.1	Tunable parameters	267
6.10	Solaris hosts	267
6.10.1	Solaris MPxIO	268
6.10.2	Symantec Veritas Volume Manager	268
6.10.3	DMP multipathing	269
6.10.4	Troubleshooting configuration issues	269

6.11 VMware server	270
6.11.1 Multipathing solutions supported	270
6.11.2 Multipathing configuration maximums	271
6.12 Monitoring	271
6.12.1 Load measurement and stress tools	272
Chapter 7. IBM Easy Tier function	273
7.1 Easy Tier	274
7.1.1 Easy Tier concepts	274
7.1.2 Four tiers Easy Tier and Read Intensive flash drive	276
7.1.3 SSD arrays and Flash MDisks	277
7.1.4 Disk tiers	280
7.1.5 Easy Tier process	282
7.1.6 Easy Tier operating modes	284
7.2 Easy Tier implementation considerations	285
7.2.1 Implementation rules	286
7.2.2 Limitations	286
7.2.3 Easy Tier settings	287
7.3 Monitoring tools	291
7.3.1 Offloading statistics	291
7.3.2 Interpreting the STAT tool output	294
7.3.3 IBM STAT Charting Utility	301
Chapter 8. Monitoring	305
8.1 Generic monitoring	306
8.1.1 Monitoring with the GUI	306
8.1.2 Monitoring using quotas and alert	307
8.2 Performance Monitoring	307
8.2.1 Performance monitoring with the GUI	308
8.2.2 Performance monitoring with IBM Spectrum Control	310
8.2.3 Important metrics for debugging	313
8.2.4 Performance support package	314
8.3 Metro and Global Mirror monitoring with IBM Copy Services Manager and scripts	315
8.3.1 Monitoring MM and GM with scripts	316
8.4 Monitoring Tier1 SSD	317
Chapter 9. Maintenance	319
9.1 Documenting IBM Spectrum Virtualize and SAN environment	320
9.1.1 Naming conventions	320
9.1.2 SAN fabric documentation	323
9.1.3 IBM Spectrum Virtualize documentation	325
9.1.4 Storage documentation	327
9.1.5 Technical support information	327
9.1.6 Tracking incident and change tickets	328
9.1.7 Automated support data collection	329
9.1.8 Subscribing to IBM Spectrum Virtualize support	329
9.2 Storage management users	329
9.3 Standard operating procedures	330
9.3.1 Allocating and deallocating volumes to hosts	330
9.3.2 Adding and removing hosts	331
9.4 IBM Spectrum Virtualize code update	332
9.4.1 Current and target IBM Spectrum Virtualize code level	332
9.4.2 IBM Spectrum Virtualize Upgrade Test Utility	332
9.4.3 IBM Spectrum Virtualize hardware considerations	334

9.4.4	Attached hosts preparation	335
9.4.5	Storage controllers preparation	335
9.4.6	SAN fabrics preparation	335
9.4.7	SAN components update sequence	336
9.4.8	IBM Spectrum Virtualize participating in Metro Mirror or Global Mirror	336
9.4.9	IBM Spectrum Virtualize update	337
9.4.10	IBM Spectrum Virtualize disk drive update	337
9.5	SAN modifications	338
9.5.1	Cross-referencing HBA WWPNs	338
9.5.2	Cross-referencing LUN IDs	340
9.5.3	HBA replacement	340
9.6	Hardware upgrades for IBM Spectrum Virtualize	341
9.6.1	Adding IBM Spectrum Virtualize nodes to an existing cluster	341
9.6.2	Upgrading IBM Spectrum Virtualize nodes in an existing cluster	343
9.6.3	Moving to a new IBM Spectrum Virtualize cluster	343
9.6.4	Splitting an IBM Spectrum Virtualize cluster	345
9.7	Adding expansion enclosures	345
9.8	I/O Throttling	346
9.8.1	General information on I/O Throttling	347
9.8.2	I/O Throttling on front end I/O control	347
9.8.3	I/O Throttling on backend I/O control	347
9.8.4	Overall benefits of using I/O Throttling	348
9.8.5	Considerations for I/O Throttling	348
9.8.6	Configuring I/O Throttling using the CLI	349
9.8.7	Configuring I/O Throttling using the GUI	349
Chapter 10.	Troubleshooting and diagnostics	353
10.1	Starting troubleshooting	354
10.1.1	Recommended actions and fix procedure	356
10.2	Remote Support Assistance	357
10.3	Common issues	358
10.3.1	Host problems	359
10.3.2	SAN events	361
10.3.3	Storage subsystem issues	361
10.3.4	Port masking issues	364
10.3.5	Interoperability	364
10.4	Collecting data and isolating the problem	365
10.4.1	Collecting data from IBM Spectrum Virtualize	365
10.4.2	SDDPCM and SDDDSM data collection	367
10.4.3	Additional data collection	368
10.5	Recovering from problems	369
10.5.1	Solving IBM Spectrum Virtualize events	369
10.5.2	Solving host problems	372
10.5.3	Solving SAN issues	374
10.5.4	Solving back-end storage issues	375
10.5.5	Common error recovery using IBM Spectrum Virtualize CLI	377
10.6	Health status during upgrade and known error	377
10.7	Call Home Web and Health Checker feature	378
10.7.1	Health Checker	379
Chapter 11.	IBM Real-time Compression	381
11.1	Evaluate compression savings using Comprestimator	382
11.2	Evaluate workload using Disk Magic	384

11.3	Verify available CPU resources	384
11.4	Configure a balanced system	385
11.5	Standard benchmark tools	386
11.6	Compression with FlashCopy	386
11.7	Compression with Easy Tier	386
11.8	Compression on the backend	387
11.9	Migrating generic volumes	387
11.10	Mixed volumes in the same MDisk group	388
Appendix A. IBM i considerations		389
	IBM i Storage management	390
	Single-level storage	390
	IBM i response time	390
	Planning for IBM i capacity	391
	Connecting SAN Volume Controller or Storwize to IBM i	391
	Native connection	391
	Connection with VIOS_NPIV	392
	Connection with VIOS virtual SCSI	393
	Setting of attributes in VIOS	393
	FC adapter attributes	393
	Disk device attributes	393
	Disk drives for IBM i	394
	Defining LUNs for IBM i	394
	Data layout	395
	Fibre Channel adapters in IBM i and VIOS	396
	Zoning SAN switches	397
	IBM i Multipath	397
	Boot from SAN	397
	IBM i mirroring	398
	Copy services considerations	398
	HyperSwap considerations	398
	Outage of Storwize I/O group at site 1	399
	Disaster at site 1	399
	Planned outage with Live Partition Mobility and Storwize HyperSwap	400
	Implementation with IASP: disaster at site 1	400
Appendix B. Business continuity		401
	Business continuity with Stretched Cluster	402
	Business continuity with Enhanced Stretched Cluster	402
	Business continuity with HyperSwap	402
	Third site and IP quorum	403
	Comparison of business continuity solutions	404
Appendix C. Scripting examples		407
	Secure Shell (SSH)	408
	Bash	408
	Python	410
	SMI-S	411
	HTTPS and RESTful API on IBM Spectrum Control	416
	HTTPS on IBM Spectrum Virtualize	418
	Conclusions	420

Related publications	421
IBM Redbooks	421
Other publications	422
Online resources	422
Help from IBM	423

Notices

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

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

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

IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

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

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

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

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

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

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

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


COPYRIGHT LICENSE:

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

Trademarks

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

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

AIX®	IBM Spectrum™	Redbooks®
DB2®	IBM Spectrum Control™	Redpaper™
developerWorks®	IBM Spectrum Protect™	Redbooks (logo)  ®
DS4000®	IBM Spectrum Virtualize™	Service Request Manager®
DS8000®	MicroLatency®	Storwize®
Easy Tier®	POWER®	System i®
FlashCopy®	POWER6®	System Storage®
Global Technology Services®	POWER7®	SystemMirror®
GPFS™	POWER8®	Tivoli®
HyperSwap®	PowerHA®	XIV®
IBM®	PowerVM®	z/OS®
IBM Cloud™	ProtecTIER®	
IBM FlashSystem®	Real-time Compression™	

The following terms are trademarks of other companies:

ITIL is a Registered Trade Mark of AXELOS Limited.

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

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

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

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

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

Preface

This IBM® Redbooks® publication captures several of the preferred practices and describes the performance gains that can be achieved by implementing the IBM System Storage® SAN Volume Controller and IBM Storwize® V7000 powered by IBM Spectrum™ Virtualize V8.1. These practices are based on field experience.

This book highlights configuration guidelines and preferred practices for the storage area network (SAN) topology, clustered system, back-end storage, storage pools and managed disks, volumes, remote copy services, and hosts. Then it provides performance guidelines for SAN Volume Controller, back-end storage, and applications. It explains how you can optimize disk performance with the IBM System Storage Easy Tier® function. It also provides preferred practices for monitoring, maintaining, and troubleshooting SAN Volume Controller and Storwize V7000.

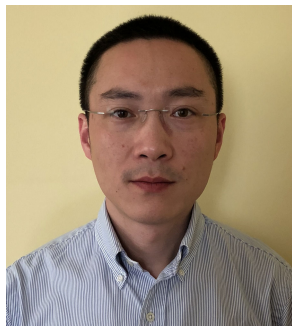
This book is intended for experienced storage, SAN, and SAN Volume Controller administrators and technicians. Understanding his book requires advanced knowledge of the SAN Volume Controller and Storwize V7000 and SAN environments.

Authors

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



Jon Tate is a Project Manager for IBM System Storage SAN Solutions at the ITSO, San Jose Center. Before joining the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2/3 support for IBM mainframe storage products. Jon has 32 years of experience in storage software and management, services, and support. He is an IBM Certified IT Specialist, an IBM SAN Certified Specialist, and is Project Management Professional (PMP) certified. He is also the UK Chairman of the Storage Networking Industry Association (SNIA).



An Chen is an IBM Lab Service storage specialist based in Australia. He joined IBM in 2007 and has more than 10 years of experience in design and delivery storage solutions for Open Systems. His areas of expertise include file and block storage implementation, SAN, IBM ProtecTIER®, performance assessment, disaster recovery and high availability solutions. He holds a Master of Science degree in Internetworking from University of Technology, Sydney.



Tiago Moreira Candelaria Bastos is a SAN and Storage Disk specialist for IBM Brazil. He has over 16 years experience in the IT arena, and is an IBM Certified Master IT Specialist, as well as certified on the Storwize portfolio. He works on Storage as a Service (SaaS) implementation projects and his areas of expertise include planning, configuring and troubleshooting IBM DS8000®, Storwize V5000 and V7000, FlashSystem 900, SVC and IBM XIV®.



Jana Jamsek was an IT specialist for IBM Slovenia at the time of writing. She worked in Storage Advanced Technical Skills for Europe as a specialist for IBM Storage Systems and IBM i systems. Jana had 8 years of experience in the IBM System i® and AS/400 areas, and 13 years of experience in Storage. She has a Master's degree in computer science and a degree in mathematics from she University of Ljubljana, Slovenia. Jana worked on complex customer cases that involved IBM i and Storage systems in various European and Middle East countries. She presented at IBM Storage and Power universities and runs workshops for IBM employees and customers. She is the author or co-author of many IBM publications in this area.



Danilo Morelli Miyasiro is a SAN and Disk Storage Specialist for IBM Global Technology Services® in Brazil. He graduated in Computer Engineering at State University of Campinas, Brazil, and has more than 10 years of experience in IT. As a storage subject matter expert (SME) for several international customers, he works on designing, implementing and supporting storage solutions. He is an IBM Certified Specialist for DS8000 and the Storwize family, and also holds certifications from the ITIL Foundation and other storage products.



Antonio Rainero is a Consulting IT Specialist working for the IBM Global Technology Services organization in IBM Italy. He joined IBM in 1998, and has more than 16 years of experience in the delivery of storage services for Open Systems and IBM z/OS® clients. His areas of expertise include storage systems implementation, SANs, storage virtualization, performance analysis, disaster recovery, and high availability solutions. He has co-authored several IBM Redbooks publications. Antonio holds a degree in Computer Science from University of Udine, Italy.

Authors of the previous edition:

Angelo Bernasconi, Tiago Moreira Candelaria Bastos, Giulio Fiscella, Bruno Garcia Galle, Jana Jamsek, Antonio Rainero

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

James Whitaker
IBM Systems, Manchester, UK

David Green
IBM Systems, US

Mikhail Zakharov
IBM Czech Republic

Senaka Meegama
IBM Australia

Alex Ainscow
Christopher Bulmer
Paul Cashman
Carlos Fuente
Katja Gebuhr
Warren Hawkins
Gareth Jones
Andrew Martin
Evelyn Perez
Mark Visser
Stephen Wright
IBM Systems, Hursley, UK

Nick Clayton
IBM Systems, UK

Barry Whyte
IBM Systems, New Zealand

Nelson Monteiro da Silva Neto
Marcos Ferreira da Silva
IBM Brazil

Diogo Henrique Padovani
Oracle

Special thanks to the Brocade Communications Systems staff in San Jose, California for their support of this residency in terms of equipment and support in many areas:

Silviano Gaona
Sangam Racherla
Brian Steffler
Marcus Thordal
Brocade Communications Systems (an indirect subsidiary of Broadcom Limited)

Now you can become a published author, too!

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

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

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

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

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

ibm.com/redbooks

- ▶ Send your comments in an email to:

redbooks@us.ibm.com

- ▶ Mail your comments to:

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

Stay connected to IBM Redbooks

- ▶ Find us on Facebook:

<http://www.facebook.com/IBMRedbooks>

- ▶ Follow us on Twitter:

<http://twitter.com/ibmredbooks>

- ▶ Look for us on LinkedIn:

<http://www.linkedin.com/groups?home=&gid=2130806>

- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>

- ▶ Stay current on recent Redbooks publications with RSS Feeds:

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

Summary of changes

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

Summary of Changes for *IBM System Storage SVC and Storwize V7000 Best Practices and Performance Guidelines*, SG24-7521-05, as created or updated on September 7, 2018.

May 2018, Sixth Edition

This revision includes the following new and changed information.

New information

- ▶ Remote Support Assistance
- ▶ Support Package Upload
- ▶ Call Home Web and Health Checker feature
- ▶ Considerations for A9000/A9000R
- ▶ Considerations for storage subsystem with compression and data deduplication capability
- ▶ Added recommendations for Solaris newer versions, and IBM AIX® SDDPCM
- ▶ Added considerations for Hot Spare Node
- ▶ Added new examples and suggestions for naming standards
- ▶ Added I/O Throttling
- ▶ Zoning and port assignment considerations for A9000 and A9000R systems
- ▶ Hot Spare Node zoning considerations
- ▶ Zoning best practices with multi core SANs
- ▶ Resiliency Best Practices for business continuity solutions
- ▶ 1920 recovery options with the Consistency Protection feature
- ▶ Resize volumes with active MM/GM relationships

Changed information

- ▶ Consolidated recommendations for storage controllers from the storage pool chapter to the storage controller chapter and revised/simplified some of the recommendations
- ▶ Updated encryption recommendations for V8.1 and added configuration details
- ▶ Included a script to start the IP quorum app for RHEL and SUSE when the server reboots or it quits unexpectedly
- ▶ Included descriptions and recommendations for MDisk group balanced and controller balanced supported storage systems
- ▶ Removed some 2145-CG8 references
- ▶ Removed some scripting examples
- ▶ Updated IBM FlashCopy® volume placement, background, and cleaning best practices sections
- ▶ Updated copy services limits and restrictions
- ▶ Updated GUI panels



Storage area network

The storage area network (SAN) is one of the most important aspects when implementing and configuring IBM Spectrum Virtualize™ and IBM Storwize storage. Due to their unique behavior and the interaction with other storage, there are specific SAN design and zoning recommendations that differ from classic storage practices.

This chapter does not describe how to design and build a flawless SAN from the beginning. Rather, it provides guidance to connect IBM Spectrum Virtualize and Storwize in an existing SAN to achieve a stable, redundant, resilient, scalable, and performance-likely environment. However, you can take the principles here into account when building your SAN.

This chapter includes the following sections:

- ▶ SAN topology general guidelines
- ▶ SAN topology-specific guidelines
- ▶ SAN Volume controller ports
- ▶ Zoning
- ▶ Distance extension for remote copy services
- ▶ Tape and disk traffic that share the SAN
- ▶ Switch interoperability

1.1 SAN topology general guidelines

The SAN topology requirements for IBM Spectrum Virtualize and Storwize do not differ too much from any other SAN. Remember that a well-sized and designed SAN allows you to build a redundant and failure-proof environment, as well as minimizing performance issues and bottlenecks. Therefore, before installing any of the products covered by this book, ensure that your environment follows an actual SAN design and architecture preferred practices for the storage industry.

For more SAN design and preferred practices, see the *SAN Fabric Resiliency and Administration Best Practices* white paper at the following link:

<https://ibm.biz/BdsFjG>

A topology is described in terms of how the switches are interconnected. There are several different SAN topologies, such as core-edge, edge-core-edge, or full mesh. Each topology has its utility, scalability, and also its cost, so one topology will be a better fit for some SAN demands than others. Independent of the environment demands, there are a few best practices that must be followed to keep your SAN working correctly, performing well, redundant, and resilient.

1.1.1 SAN performance and scalability

Regardless of the storage and the environment, planning and sizing of the SAN makes a difference when growing your environment and when troubleshooting problems.

Because most SAN installations continue to grow over the years, the main SAN industry-lead companies design their products in a way to support a certain growth. Keep in mind that your SAN must be designed to accommodate both short-term and medium-term growth.

From the performance standpoint, the following topics must be evaluated and considered:

- ▶ Host-to-storage fan-in fan-out ratios
- ▶ Host to inter-switch link (ISL) oversubscription ratio
- ▶ Edge switch to core switch oversubscription ratio
- ▶ Storage to ISL oversubscription ratio
- ▶ Size of the trunks
- ▶ Monitor for slow drain device issues

From the scalability standpoint, ensure that your SAN will support the new storage and host traffic. Make sure that the chosen topology will also support a growth not only in performance, but also in port density.

If new ports need to be added to the SAN, you might need to drastically modify the SAN to accommodate a larger-than-expected number of hosts or storage. Sometimes these changes increase the number of hops on the SAN, and so cause performance and ISL congestion issues. For additional information, see 1.1.2, “ISL considerations” on page 3.

Consider the use of SAN director-class switches. They reduce the number of switches in a SAN and provide the best scalability available. Most of the SAN equipment vendors provide high port density switching devices. With MDS 9718 Multilayer Director, Cisco offers the industry’s highest port density single chassis with up to 768 16/32 Gb ports. The IBM b-type UltraScale Inter-Chassis Links (ICL) technology enables you to create multichassis configurations with up to 4608 16/32 Gb ports.

Therefore, if possible, plan for the maximum size configuration that you expect your IBM SAN Volume Controller (SAN Volume Controller or SVC) and Storwize installation to reach. Planning for the maximum size does not mean that you must purchase all of the SAN hardware initially. It only requires you to design the SAN to be able to reach the expected maximum size.

1.1.2 ISL considerations

ISLs are responsible for interconnecting the SAN switches, creating SAN flexibility and scalability. For this reason, they can be considered as the core of a SAN topology. Consequently, they are sometimes the main cause of issues that can affect a SAN. For this reason it is important to take extra caution when planning and sizing the ISL in your SAN.

Regardless of your SAN size, topology, or the size of your SAN Volume Controller/Storwize installation, consider the following practices to your SAN Inter-switch link design:

- ▶ Beware of the ISL oversubscription ratio
 - The standard recommendation is up to 7:1 (seven hosts using a single ISL). However, it can vary according to your SAN behavior. Most successful SAN designs are planned with an oversubscription ratio of 7:1 and some extra ports are reserved to support a 3:1 ratio. However, high-performance SANs start at a 3:1 ratio.
 - Exceeding the standard 7:1 oversubscription ratio requires you to implement fabric bandwidth threshold alerts. If your ISLs exceed 70%, schedule fabric changes to distribute the load further.
- ▶ Avoid unnecessary ISL traffic
 - Connect all SAN Volume Controller/Storwize node ports in a clustered system to the same SAN switches/Directors as all of the storage devices with which the clustered system of SAN Volume Controller/Storwize is expected to communicate. Conversely, storage traffic and internode traffic must never cross an ISL, except during migration scenarios.
 - Keep high-bandwidth utilization servers and I/O Intensive application on the same SAN switches as the SAN Volume Controller/Storwize host ports. Placing these servers on a separate switch can cause unexpected ISL congestion problems. Also, placing a high-bandwidth server on an edge switch wastes ISL capacity.
- ▶ Properly size the ISLs on your SAN. They must have adequate bandwidth and buffer credits to avoid traffic or frames congestion. A congested inter-switch link can affect the overall fabric performance.

When the Fibre Channel switches reach 4 Gbps and beyond, the number of issues that are related to traffic saturation dramatically decrease. It is rare for an ISL trunk to reach a sustained 100 percent bandwidth utilization. However, congestion related to buffer credit starvation remains common.

- ▶ Always deploy redundant ISLs on your SAN. Using an extra ISL avoids congestion if an ISL fails because of certain issues, such as a SAN switch line card or port blade failure.
- ▶ Use the link aggregation features, such as IBM b-type Trunking or Cisco Port Channel, to obtain better performance and resiliency.
- ▶ Avoid exceeding two hops between the SAN Volume Controller/Storwize and the hosts. More than two hops are supported. However, when ISLs are not sized properly, more than two hops can lead to ISL performance issues and buffer credit starvation (SAN congestion).

When sizing over two hops, consider that all of the ISLs going to the switch where the SAN Volume Controller/Storwize is connected will also handle the traffic coming from the switches on the edges, as shown in Figure 1-1.

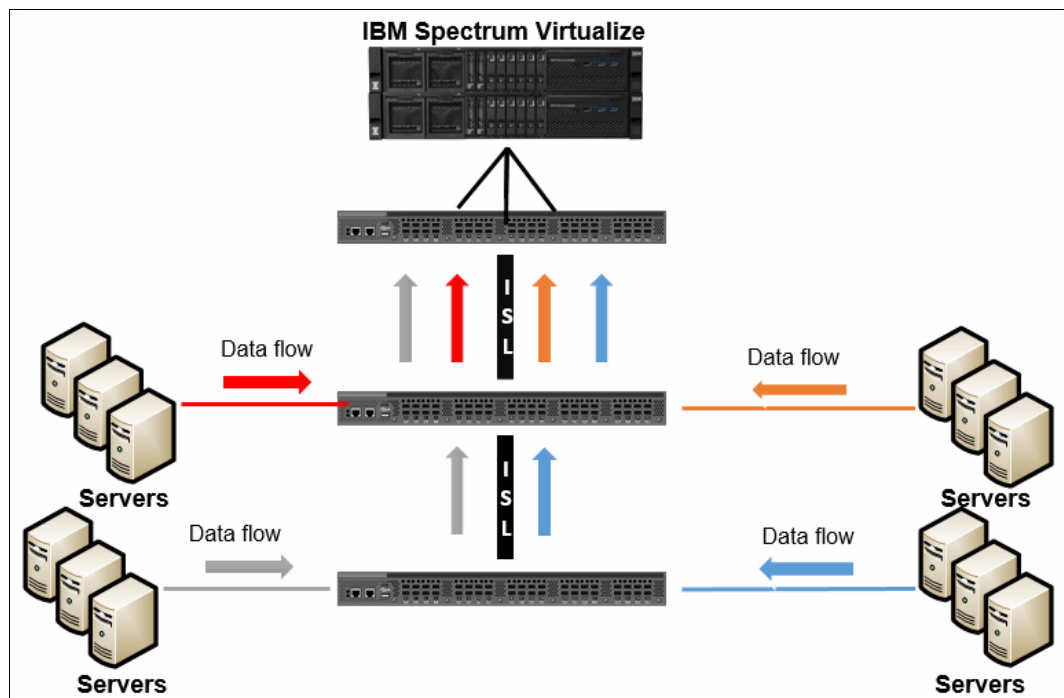


Figure 1-1 ISL data flow

- ▶ If possible, use SAN directors to avoid many ISL connections. Problems that are related to oversubscription or congestion are much less likely to occur within SAN director fabrics.
- ▶ When interconnecting SAN directors through ISL, spread the ISL cables across different directors blades. In a situation where an entire blade fails, the ISL will still be redundant through the links connected to other blades.
- ▶ Plan for the peak load, not for the average load.

1.2 SAN topology-specific guidelines

Some preferred practices, as mentioned in 1.1, “SAN topology general guidelines” on page 2, apply to all SANs. However, there are specific preferred practices requirements to each SAN topology available. The following topic shows the difference between the different kinds of topology and highlights the specific considerations for each of them.

This section covers the following topologies:

- ▶ Single switch fabric
- ▶ Core-edge fabric
- ▶ Edge-core-edge
- ▶ Full mesh

1.2.1 Single switch SAN Volume Controller/Storwize SANs

The most basic SAN Volume Controller/Storwize topology consists of a single switch per SAN fabric. This switch can range from a 24-port 1U switch for a small installation of a few hosts and storage devices, to a director with hundreds of ports. This is a low-cost design solution that has the advantage of simplicity and is a sufficient architecture for small-to-medium SAN Volume Controller/Storwize installations.

One of the advantages of a single switch SAN is that when all servers and storages are connected to the same switches, there is no hop.

Note: To meet redundancy and resiliency requirements, a single switch solution needs at least two SAN switches and directors, with one per different fabric.

The preferred practice is to use a multislotted director-class single switch over setting up a core-edge fabric that is made up solely of lower-end switches, as described in 1.1.1, “SAN performance and scalability” on page 2.

The single switch topology, as shown in Figure 1-2, has only two switches, so the SAN Volume Controller/Storwize ports must be equally distributed on both fabrics.

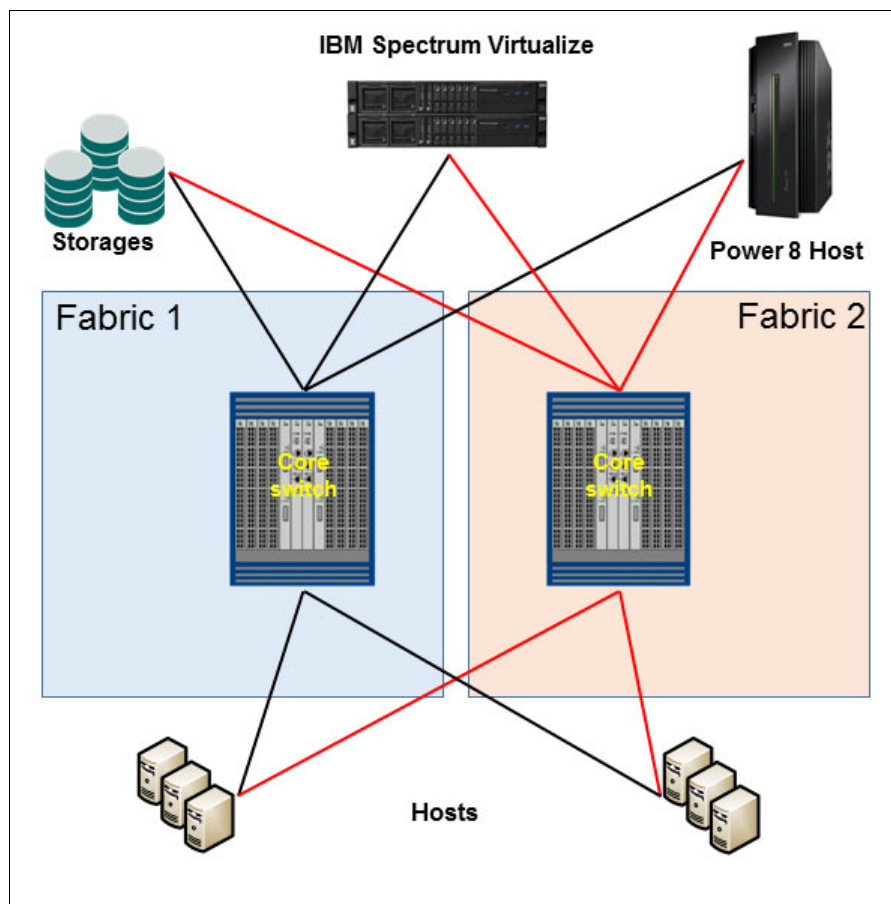


Figure 1-2 Single switch SAN

1.2.2 Basic core-edge topology

The core-edge topology (as shown in Figure 1-3) is easily recognized by most SAN architects. This topology consists of a switch in the center (usually, a director-class switch), which is surrounded by other switches. The *core switch* contains all SAN Volume Controller and Storwize ports, storage ports, and high-bandwidth hosts. It is connected by using ISLs to the edge switches. The edge switches can be of any size from 24 port switches up to multi-slot directors.

When the SAN Volume Controller, Storwize, and servers are connected to different switches, the hop count for this topology is one.

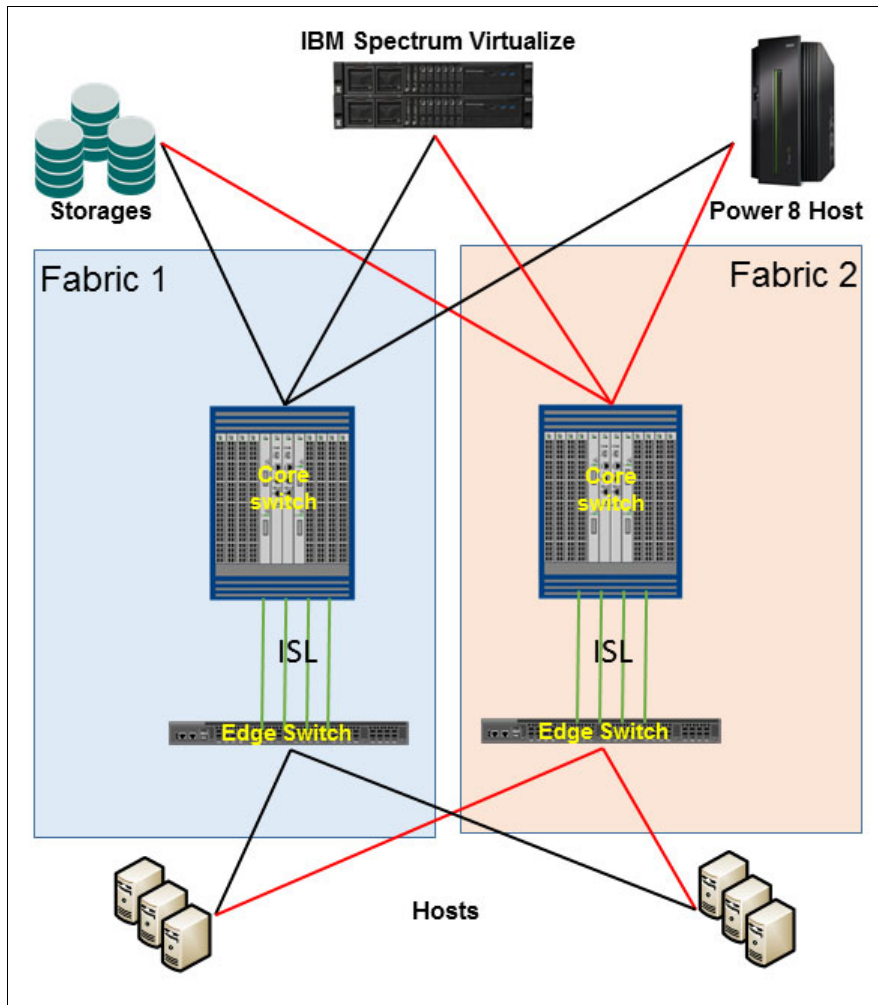


Figure 1-3 Core-edge topology

1.2.3 Edge-core-edge topology

Edge-core-edge is a topology that is used for installations where a core-edge fabric made up of multislot director-class SAN switches is insufficient. This design is useful for large, multiclustered system installations. Similar to a regular core-edge, the edge switches can be of any size, and multiple ISLs must be installed per switch.

Figure 1-4 shows an edge-core-edge topology with two different edges, one of which is exclusive for the storage, SAN Volume Controller, and high-bandwidth servers. The other pair is exclusively for servers.

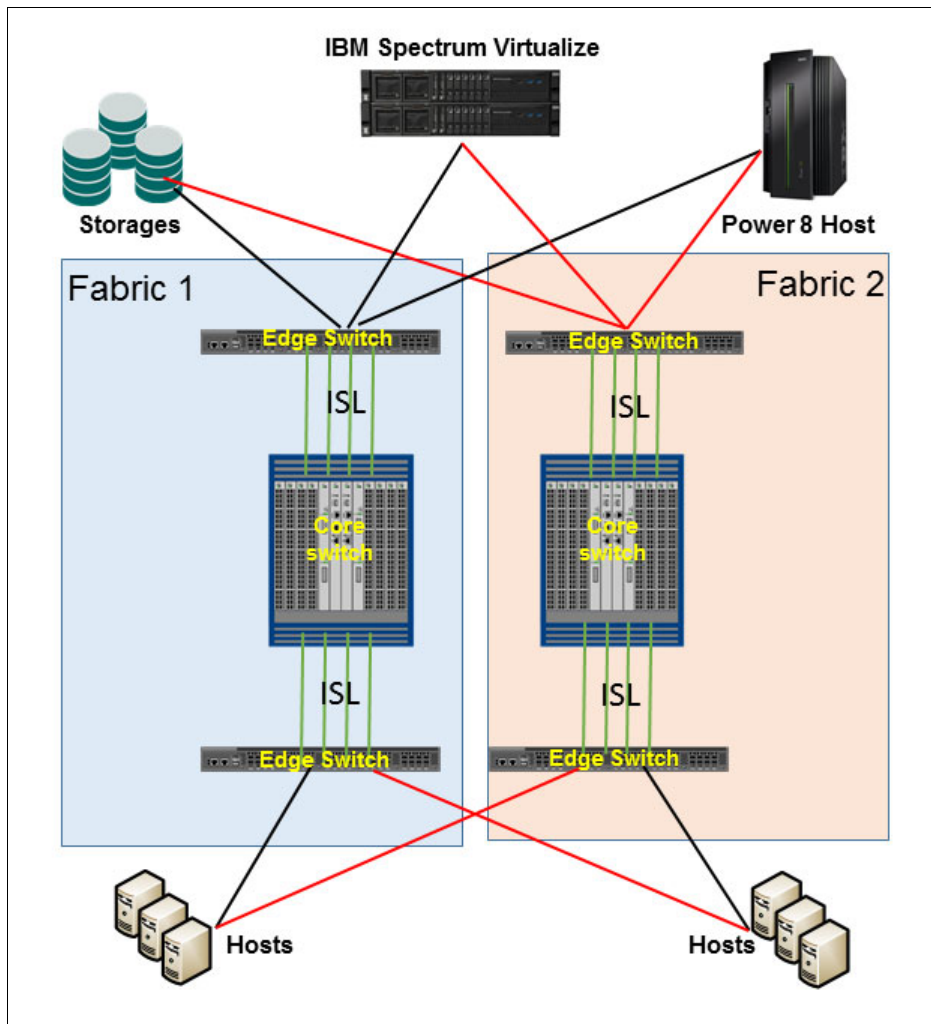


Figure 1-4 Edge-core-edge topology

Edge-core-edge fabrics allow better isolation between tiers. For additional information, see 1.2.6, “Device placement” on page 9.

1.2.4 Full mesh topology

In a full mesh topology, all switches are interconnected to all other switches on the same fabric. So the server and storage placement is not a concern after the number of hops is no more than one hop. Figure 1-5 shows a full mesh topology.

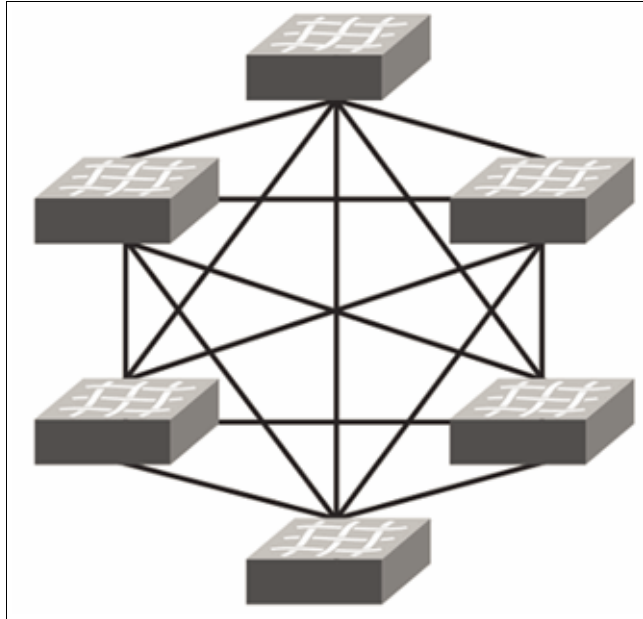


Figure 1-5 Full mesh topology

1.2.5 IBM Spectrum Virtualize and IBM Storwize as a SAN bridge

IBM SAN Volume Controller nodes now have a maximum of 16 ports. In addition to the increased throughput capacity, this number of ports enables new possibilities and allows different kinds of topologies and migration scenarios.

One of these topologies is the use of a SAN Volume Controller and IBM Storwize as a bridge between two isolated SANs. This configuration is useful for storage migration or sharing resources between SAN environments without merging them. Another use is if you have devices with different SAN requirements in your installation.

Figure 1-6 has an example of an IBM Spectrum Virtualize and Storwize as a SAN bridge.

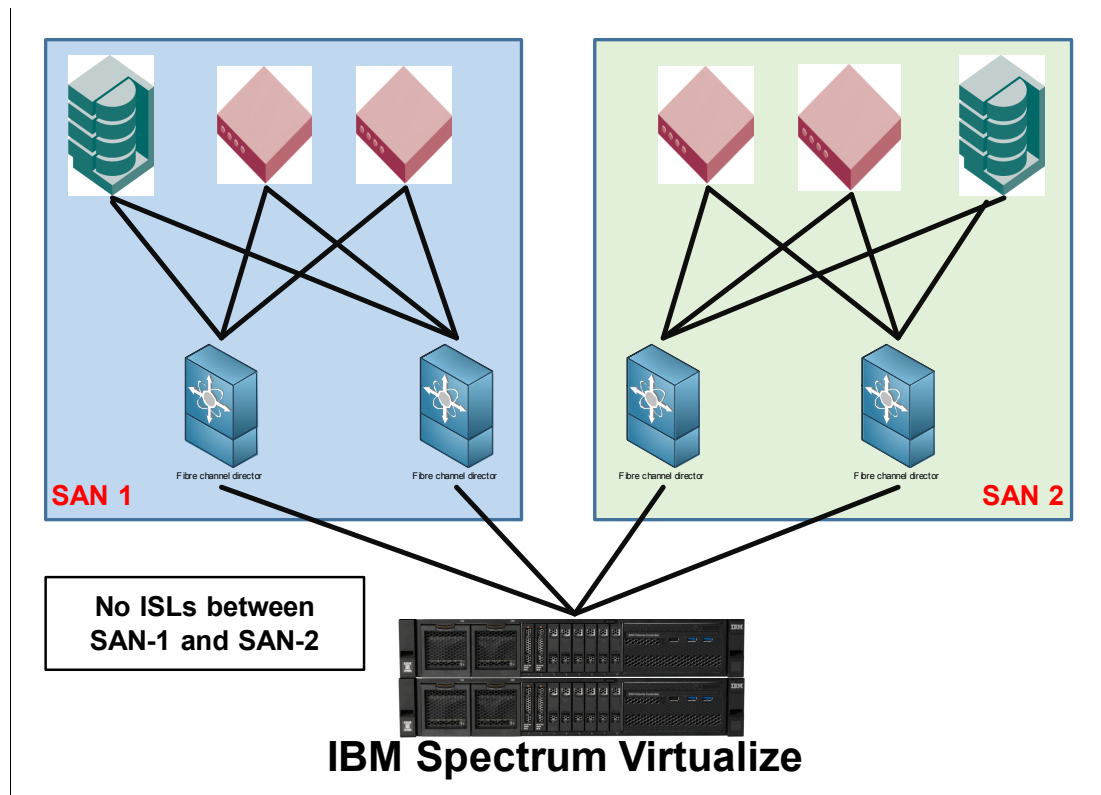


Figure 1-6 IBM Spectrum Virtualize and Storwize as SAN bridge

Notice in Figure 1-6 that both SANs (Blue and Green) are isolated and there is no communication through ISLs. When connected to both fabrics, SAN Volume Controller and Storwize are able to virtualize storages from either fabrics. They can provide disks from storage on the Green SAN (right), for example, to hosts on blue SAN (left).

1.2.6 Device placement

With the growth of virtualization, it is not usual to experience frame congestion on the fabric. Device placement seeks to balance the traffic across the fabric to ensure that the traffic is flowing in a certain way to avoid congestion and performance issues. The ways to balance the traffic consist of isolating traffic by using zoning, virtual switches, or traffic isolation zoning.

Keeping the traffic local to the fabric is a strategy to minimize the traffic between switches (and ISLs) by keeping the data flow local, as shown in Figure 1-7.

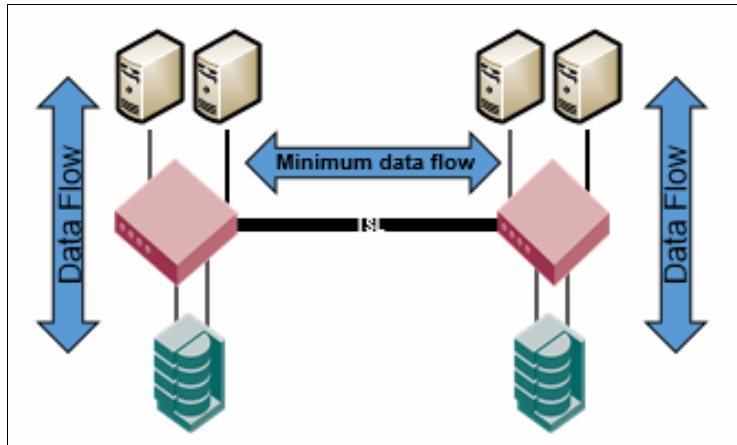


Figure 1-7 Storage and hosts attached to the same SAN switch

This solution can fit perfectly in small and medium SANs. However, it is not as scalable as other topologies available. As stated in 1.2, “SAN topology-specific guidelines” on page 4, the most scalable SAN topology is the edge-core-edge. Besides scalability, this topology provides different resources to isolate the traffic and reduce possible SAN bottlenecks. Figure 1-8 shows an example of traffic segregation on the SAN using edge-core-edge topology.

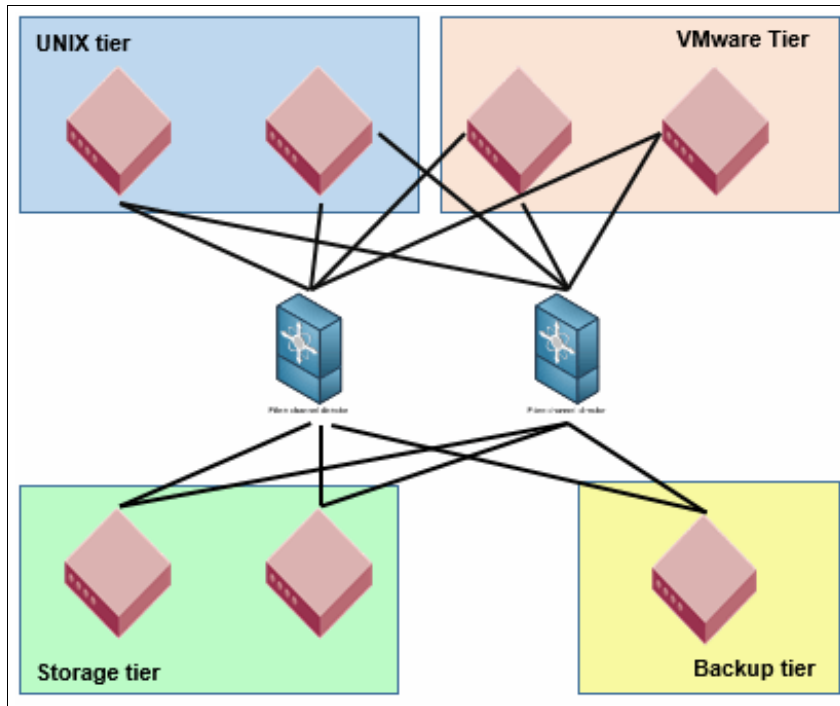


Figure 1-8 Edge-core-edge segregation

Even when sharing the same core switches, it is possible to use virtual switches (see “SAN partitioning” on page 11 for details) to isolate one tier from the other. This configuration helps avoid traffic congestion caused by slow drain devices that are connected to the backup tier switch.

SAN partitioning

SAN partitioning is a hardware-level feature that allows SAN switches to share hardware resources by partitioning its hardware into different and isolated virtual switches. Both IBM b-type and Cisco provide SAN partitioning features called, respectively, *Virtual Fabric* and *Virtual SAN (VSAN)*.

Hardware-level fabric isolation is accomplished through the concept of switch virtualization, which allows you to partition physical switch ports into one or more “virtual switches.” Virtual switches are then connected to form virtual fabrics. As the number of available ports on a switch continues to grow, partitioning switches allow storage administrators to take advantage of high port density switches by dividing physical switches into different virtual switches. Without SAN partitioning an FC switch is limited to 256 ports.

From a device perspective, SAN partitioning is completely transparent and so the same guidelines and practices that apply to physical switches apply also to the virtual ones.

While the main purposes of SAN partitioning are port consolidation and environment isolation, this feature is also instrumental in the design of a business continuity solution based on IBM Spectrum Virtualize/Storwize.

For a description of the IBM Spectrum Virtualize/Storwize business continuity solutions, see Appendix B, “Business continuity” on page 401.

1.3 SAN Volume controller ports

IBM SAN Volume controller hardware has significantly increased port connectivity options. Models 2145-DH8 and 2145-SV1 deliver up to 16x 16 Gb FC ports per node as shown in Table 1-1.

Table 1-1 SVC connectivity

Feature	2145-DH8	2145-SV1
Fibre Channel HBA	4x Quad 8 Gb 4x Dual 16 Gb 4x Quad 16 Gb	4x Quad 16 Gb
Ethernet I/O	4x Quad 10 Gb iSCSI/FCoE	4x Quad 10 Gb iSCSI/FCoE
Built in ports	4x 1 Gb	4x 10 Gb
SAS expansion ports	4x 12 Gb SAS	4x 12 Gb SAS

This new port density expands the connectivity options and provides new ways to connect the SVC to the SAN. This sections describes some preferred practices and use cases that show how to connect a SAN volume controller on the SAN to use this increased capacity.

1.3.1 Slots and ports identification

The SAN volume controller can have up to four quad Fibre Channel (FC) HBA cards (16 FC ports) per node. Figure 1-9 shows the port location in the rear view of the 2145-SV1 node.

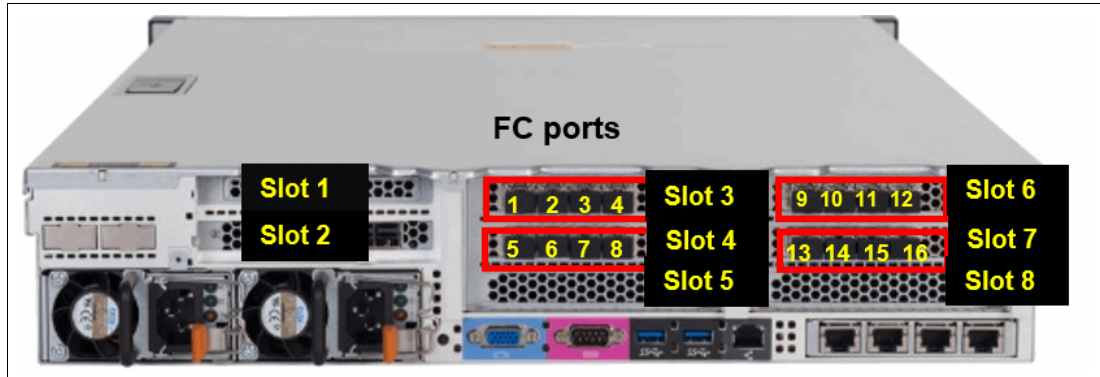


Figure 1-9 SAN Volume Controller 2145-SV1 rear port view

Figure 1-10 shows the 2145-DH8 node port and slot locations.

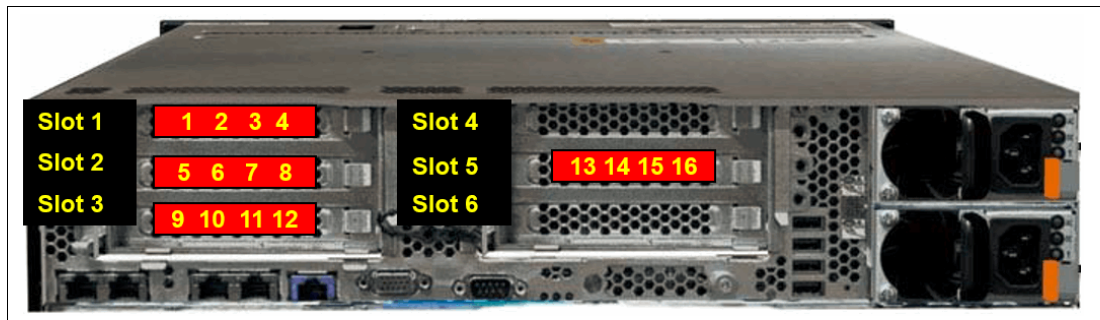


Figure 1-10 SAN Volume controller 2145-DH8 rear port view

For maximum redundancy and resiliency, spread the ports across different fabrics. Because the port count varies according to the number of cards included in the solution, try to keep the port count equal on each fabric.

1.3.2 Port naming and distribution

In the field, fabric naming conventions vary. However, it is common to find fabrics that are named, for example, PROD_SAN_1 and PROD_SAN_2, or PROD_SAN_A and PROD_SAN_B. This type of naming convention is used to simplify the SAN Volume Controller, after their denomination followed by *1* and *2* or *A* and *B*, which specifies that the devices connected to those fabrics contains the redundant paths of the same servers and SAN devices.

To simplify the SAN connection identification and troubleshooting, keep all odd ports on the odd fabrics, or “A” fabrics and the even ports on the even fabric or “B” fabrics, as shown in Figure 1-11.

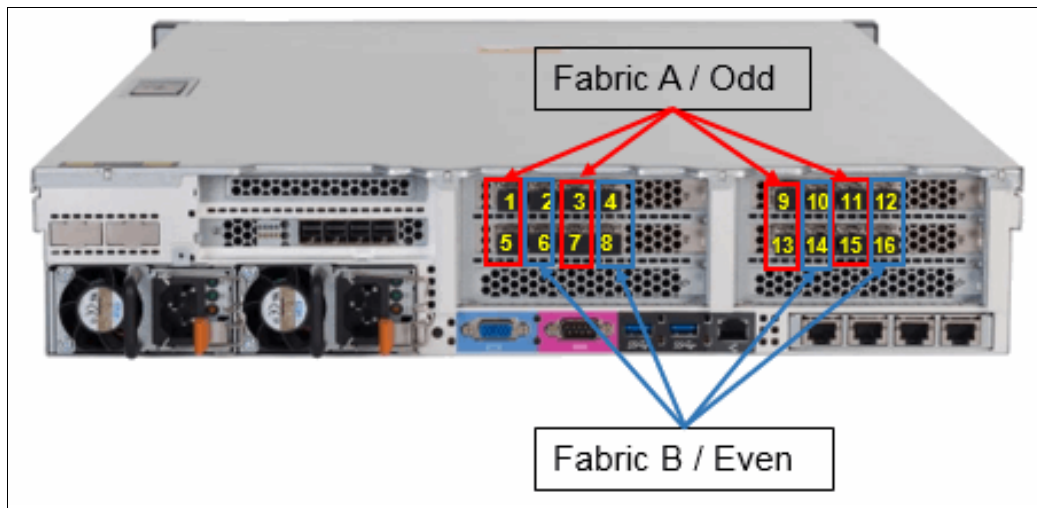


Figure 1-11 SAN Volume controller model 2145-SV1 Port distribution

SAN Volume Controller model 2145-DH8 follows the same arrangement, odd ports to odd or “A” Fabric, and even ports attached to even fabrics or “B” fabric.

As a preferred practice, assign specifics uses to specifics SAN volume controller ports. This technique helps optimize the port utilization by aligning the internal allocation of hardware CPU cores and software I/O threads to those ports. Apart from that guideline, configuring the ports for specific uses will ensure the ability to replace nodes non-disruptively in the future.

Figure 1-12 shows the specific port use guidelines for the 2145-CG8 and 2145-DH8.

Slot/Port	Port #	SAN	4-port Nodes	8-port Nodes with 2 port cards	8-port Nodes with 4 port cards	12-port Nodes	16-port Nodes
S1P1	1	A / 1	Host/Storage/Inter-node	Host/Storage	Host/Storage	Host/Storage	Host/Storage
S1P2	2	B / 2	Host/Storage/Inter-node	Host/Storage	Host/Storage	Host/Storage	Host/Storage
S1P3	3	A / 1	Host/Storage/Replication*	--	Inter-node	Host/Storage	Host/Storage
S1P4	4	B / 2	Host/Storage/Replication*	--	Host/Storage or Replication**	Host/Storage	Host/Storage
S2P1	5	A / 1	Host/Storage/Replication*	Host/Storage	Host/Storage	Inter-node	Inter-node
S2P2	6	B / 2	Host/Storage/Replication*	Host/Storage	Host/Storage	Inter-node	Inter-node
S2P3	7	A / 1	Host/Storage/Replication*	--	Host/Storage or Replication**	Host/Storage or Replication**	Host/Storage or Replication**
S2P4	8	B / 2	Host/Storage/Replication*	--	Inter-node	Host/Storage	Host/Storage
S3P1	9	A / 1	Host/Storage/Replication*	Inter-node	Inter-node	Host/Storage	Host/Storage
S3P2	10	B / 2	Host/Storage/Replication*	Host/Storage or Replication**	Host/Storage or Replication**	Host/Storage or Replication**	Host/Storage or Replication**
S3P3	11	A / 1	Host/Storage/Replication*	--	Inter-node or Host/Storage	Inter-node or Host/Storage	Inter-node or Host/Storage
S3P4	12	B / 2	Host/Storage/Replication*	--	Inter-node or Host/Storage	Inter-node or Host/Storage	Inter-node or Host/Storage
S5P1	13	A / 1	Host/Storage/Replication*	Host/Storage or Replication**	Host/Storage or Replication**	Host/Storage	Host/Storage
S5P2	14	B / 2	Host/Storage/Replication*	Inter-node	Inter-node	Host/Storage	Host/Storage
S5P3	15	A / 1	Host/Storage/Replication*	--	Host/Storage	Host/Storage	Host/Storage
S5P4	16	B / 2	Host/Storage/Replication*	--	Host/Storage	Host/Storage	Host/Storage
localfcportmask			With Rep 0011 / No Rep 1111	10010000	10000100	110000110000	0000110000110000
remotefcportmask			1100	01100000	01001000	001001000000	0000001001000000
* Inter-node if no replication planned							
** Use for Host/Storage in case no replication is in place.							

Figure 1-12 Port masking configuration on 2145-CG8 and 2145-DH8 nodes

The preferred practice for the 2145-SV1 is similar. The significant change is on the slot placement, and that there is no dual port HBA configuration option, as shown in Figure 1-13.

Slot/Port	Port #	SAN	4-port Nodes	8-port Nodes	12-port Nodes	16-port Nodes
S3P1	1	A / 1	Inter-node or Host/Storage	Inter-node	Inter-node	Inter-node
S3P2	2	B / 2	Inter-node or Host/Storage	Host/Storage or Replication*	Host/Storage or Replication*	Host/Storage or Replication*
S3P3	3	A / 1	Host/Storage or Replication	Host/Storage	Host/Storage	Host/Storage
S3P4	4	B / 2	Host/Storage or Replication	Host/Storage	Host/Storage	Host/Storage
S4P1	5	A / 1		Host/Storage or Replication*	Host/Storage or Replication*	Host/Storage or Replication*
S4P2	6	B / 2		Host/Storage	Host/Storage	Host/Storage
S4P3	7	A / 1		Host/Storage	Host/Storage	Host/Storage
S4P4	8	B / 2		Inter-node	Inter-node	Inter-node
S6P1	9	A / 1			Host/Storage	Host/Storage
S6P2	10	B / 2			Host/Storage	Host/Storage
S6P3	11	A / 1			Inter-node or Host/Storage	Inter-node or Host/Storage
S6P4	12	B / 2			Inter-node or Host/Storage	Inter-node or Host/Storage
S7P1	13	A / 1				Host/Storage
S7P2	14	B / 2				Host/Storage
S7P3	15	A / 1				Host/Storage
S7P4	16	B / 2				Host/Storage
localfcportmask			With Rep 0011/ No Rep 1111	10000010	110010000001	0000110010000001
remotefcportmask			1100	00010010	000000010010	0000000000010010

* Use for Host/Storage in case no replication is in place.

Figure 1-13 2145-SV1 port assignment

Due to the new port availability on the 2145-DH8 and 2145-SV1, and the increased bandwidth with the 16 Gb ports, it is possible to segregate the port assignment between hosts and storage, thus isolating their traffic.

Host and storage ports have different traffic behavior, so keeping host and storage ports together produces maximum port performance and utilization by benefiting from its full duplex bandwidth. For this reason, sharing host and storage traffic in the same ports is generally the preferred practice. However, traffic segregation can also provide some benefits in terms of troubleshooting and host zoning management. Consider, for instance, SAN congestion conditions due to a slow draining device.

In this case, segregating the ports simplifies the identification of the device causing the problem. At the same time, it limits the effects of the congestion to the hosts or backend ports only. Furthermore, dedicating ports for host traffic reduces the possible combinations of host zoning and simplifies SAN management. It is advised to implement the port traffic segregation with configurations with a suitable number of ports (that is, 12 ports or more) only.

Important: Use port masking to assign specific uses to the SAN Volume Controller ports. For additional information, see Chapter 6, “Hosts” on page 241.

Buffer credits

SAN Volume Controller and Storwize ports have a predefined number of buffer credits. The amount of buffer credits determines the available throughput over distances as follows:

- ▶ All 8 Gbps adapters have 41 credits available per port, saturating links at up to 10 km at 8 Gbps
- ▶ 2-port 16 Gbps (DH8 only nodes) adapters have 80 credits available per port, saturating links at up to 10 km at 16 Gbps
- ▶ 4-port 16 Gbps adapters have 40 credits available per port, saturating links at up to 5 km at 16 Gbps

Switch port buffer credit: For stretched cluster and IBM HyperSwap® configurations not using ISLs for the internode communication, it is advised to set the switch port buffer credits to match the IBM Spectrum Virtualize/Storage port.

Port designation and CPU cores utilization

The ports assignment/designation recommendation is based on the relationship between a single port to a CPU and core.

Figure 1-14 shows the Port to CPU core mapping for a 2145-SV1 node.

Uncompressed-Physical port to CPU Core								
CPU1 CORE	0	1	2	3	4	5	6	7
Port	15	14	13	12	11	10	9	8
CPU2 Core	8	9	10	11	12	13	14	XX
Port	7	6	5	4	3	2	1	XX
					16			

Compressed-Physical port to CPU Core								
CPU1 CORE	0	1	2	3	4	5	6	7
Port	15	14	1	12	11	10	9	8
	7	6	3	4	3	2	1	
			5		16			

Slot 3	Slot 4	Slot 6	Slot 7
--------	--------	--------	--------

Figure 1-14 Port to CPU core mapping for SV1 nodes

Figure 1-15 shows the Port to CPU core mapping for a 2145-DH8 node.

CPU1 CORE	0	1	2	3	4	5	6	7
WWPN	1	2	3	4	5	6	7	8
WWPN	9	10	11	12	13	14	15	16

Slot 1	Slot 2	Slot 3	Slot 5
--------	--------	--------	--------

Figure 1-15 Port to CPU core mapping for DH8 nodes

N_Port ID Virtualization feature: On N_Port ID Virtualization (NPIV)-enabled systems, the port to CPU core assignment for the virtual WWPN is the same as the physical WWPN.

1.4 Zoning

Because of the nature of storage virtualization and cluster scalability, the SVC/Storwize zoning differs from traditional storage devices. Zoning a SVC/Storwize cluster into a SAN fabric requires planning and following specific guidelines.

Important: Errors that are caused by improper SAN Volume Controller/Storwize zoning are often difficult to isolate and the steps to fix them can impact the SAN environment. Therefore, create your zoning configuration carefully.

The initial configuration for SAN Volume Controller and Storwize requires the following different zones:

- ▶ Internode and intra-cluster zones
- ▶ Replication zones (if using replication)
- ▶ Back-end storage to SAN Volume Controller/Storwize zoning
- ▶ Host to SAN Volume Controller/Storwize zoning

For each zoning type, there are different guidelines, which are detailed later in this chapter.

Note: Although internode/intra-cluster zone is not necessary for non clustered Storwize family systems, it is generally preferred to have one.

1.4.1 Types of zoning

Modern SAN switches have two types of zoning available: Port zoning, and worldwide port name (WWPN) zoning. The preferred method is to use only WWPN zoning. A common misconception is that WWPN zoning provides poorer security than port zoning, which is not the case. Modern SAN switches enforce the zoning configuration directly in the switch hardware. Also, you can use port binding functions to enforce a WWPN to be connected to a particular SAN switch port.

Zoning types and NPIV: Avoid the use of a zoning configuration that has a mix of port and WWPN zoning. For NPIV configurations, host zoning must use the WWPN zoning type.

Traditional zone design preferred practice calls for *single initiator* zoning. This means that a zone can consist of many target devices but only one initiator. This is because target devices will usually wait for an initiator device to connect to them, while initiators will actively attempt to connect to each device that they are zoned to. The single initiator approach removes the possibility of a misbehaving initiator affecting other initiators.

The drawback to single initiator zoning is that on a large SAN having many zones can make the SAN administrators job more difficult, and the number of zones on a large SAN can exceed the zone database size limits.

Cisco and IBM b-type have both developed features that can reduce the number of zones by allowing the SAN administrator to control which devices in a zone can talk to other devices in the zone. The features are called Cisco Smart Zoning and IBM b-type Peer Zoning. Both Cisco Smart Zoning and IBM b-type Peer Zoning are supported with IBM Spectrum Virtualize and Storwize systems. A brief overview of both is provided below.

Cisco Smart Zoning

Cisco Smart Zoning is a feature that, when enabled, restricts the initiators in a zone to communicating only with target devices in the same zone. For our cluster example, this would allow a SAN administrator to zone all of the host ports for a VMware cluster in the same zone with the storage ports that all the hosts need access to. Smart Zoning configures the access control lists in the fabric routing table to only allow the hosts to communicate with target ports.

You can read more about Smart Zoning here:

<https://ibm.biz/Bdjuu2>

Other relevant implementation information can be found here:

<https://ibm.biz/Bdjuuq>

IBM b-type Peer Zoning

IBM b-type Peer Zoning is a feature that provides a similar functionality of restricting what devices can see other devices within the same zone. However, Peer Zoning is implemented such that some devices in the zone are designated as principal devices. The non-principal devices can only communicate with the principal device, not with each other. As with Cisco, the communication is enforced in the fabric routing table. You can see a comparison of Peer Zoning to traditional zoning here:

<https://ibm.biz/Bdjuuz>

Recommendation: Use Smart and Peer zoning for the host zoning only. For intracluster, backend, and intercluster zoning, use traditional zoning instead.

1.4.2 Prezoning tips and shortcuts

Several tips and shortcuts are available for SAN Volume Controller/Storwize zoning.

Naming convention and zoning scheme

When you create and maintaining a SAN Volume Controller/Storwize zoning configuration, you must have a defined naming convention and zoning scheme. If you do not define a naming convention and zoning scheme, your zoning configuration can be difficult to understand and maintain.

Remember that environments have different requirements, which means that the level of detailing in the zoning scheme varies among environments of various sizes. Therefore, ensure that you have an easily understandable scheme with an appropriate level of detail. Then make sure that you use it consistently and adhere to it whenever you change the environment.

For more information about SAN Volume Controller/Storwize naming convention, see 9.1.1, “Naming conventions” on page 320.

Aliases

Use zoning aliases when you create your SAN Volume Controller/Storwize zones if they are available on your particular type of SAN switch. Zoning aliases makes your zoning easier to configure and understand, and causes fewer possibilities for errors.

One approach is to include multiple members in one alias because zoning aliases can normally contain multiple members (similar to zones). This approach can help avoid some common issues that are related to zoning and make it easier to maintain the port balance in a SAN.

Create the following zone aliases:

- ▶ One zone alias for each SAN Volume Controller/Storwize port
- ▶ Zone an alias group for each storage subsystem port pair (the SAN Volume controller/Storwize must reach the same storage ports on both I/O group nodes)

You can omit host aliases in smaller environments, as we did in the lab environment that was used for this publication. Figure 1-16 shows some alias examples.

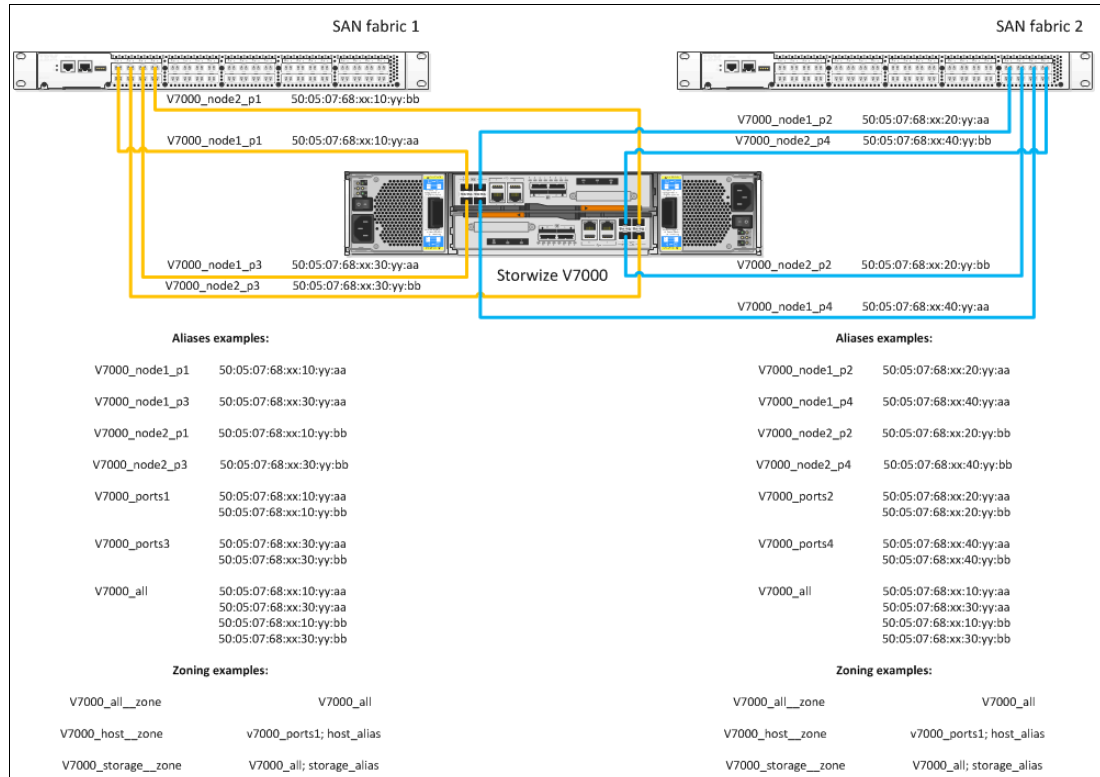


Figure 1-16 Different SAN Volume Controller/Storwize aliasing examples

1.4.3 SAN Volume Controller/Storwize internode communications zones

Internode (or intra-cluster) communication is critical to the stable operation of the cluster. The ports that carry internode traffic are used for mirroring write cache and metadata exchange between nodes/canisters. In Storwize systems, internode communication primarily take place through the internal PCI connectivity between the two canisters of a control enclosure. However for the clustered Storwize systems, the internode communication requirements are very similar to the SAN Volume Controller ones.

To establish efficient, redundant, and resilient intracluster communication, the intracluster zone must contain at least two ports from each node/canister. For SVC nodes with eight ports or more, generally isolate the intracluster traffic by dedicating node ports specifically to internode communication. The ports to be used for intracluster communication varies according to the machine type-model number and port count. See Figure 1-12 on page 13 (2145-DH8) and Figure 1-13 on page 14 (2145-SV1) for port assignment recommendations.

NPIV configurations: On NPIV-enabled configurations, use the physical WWPN for the intracluster zoning.

Only 16 port logins are allowed from one node to any other node in a SAN fabric. Ensure that you apply the proper port masking to restrict the number of port logins. Without port masking any SAN Volume Controller/Storwize port, and any member of the same zone can be used for intracluster communication, even the port members of SVC to host and SVC to storage zoning.

Note: To check whether the login limit is exceeded, count the number of distinct ways by which a port on node X can log in to a port on node Y. This number must not exceed 16. For more port masking information, see Chapter 6, “Hosts” on page 241.

1.4.4 SAN Volume Controller/Storage storage zones

The zoning between SAN Volume Controller/Storage and other storage is necessary to allow the virtualization of any storage space under the SAN Volume Controller/Storage. This storage is referred to as back-end storage.

A zone for each back-end storage to each SAN Volume controller/Storage node/canister must be created in both fabrics, as shown in Figure 1-17. Doing so reduces the overhead that is associated with many logins. The ports from the storage subsystem must be split evenly across the dual fabrics.

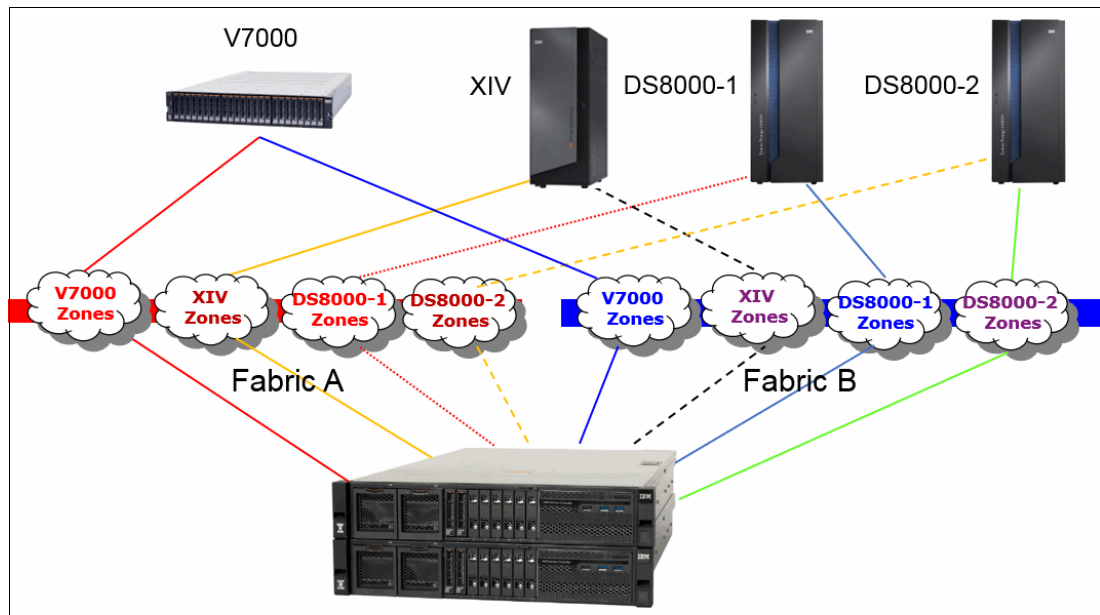


Figure 1-17 Back-end storage zoning

Usually all nodes/canisters in a SAN Volume Controller/Storage system should be zoned to the same ports on each back-end storage system, with the following exceptions:

- ▶ When implementing Enhanced Stretched Cluster or HyperSwap configurations where the backend zoning can be different for the nodes/canisters accordingly to the site definition (see *IBM Spectrum Virtualize and SAN Volume Controller Enhanced Stretched Cluster with VMware*, SG24-8211 and *IBM Storwize V7000, Spectrum Virtualize, HyperSwap, and VMware Implementation*, SG24-8317 for further details)
- ▶ When the SAN has a multi-core design that requires special zoning considerations as described in “Zoning to storage best practice” on page 21

NPIV configurations: On NPIV enabled systems, use the physical WWPN for the zoning to the backend controller.

When two nodes/canisters are zoned to different set of ports for the same storage system, the SAN Volume Controller/Storwize operation mode is considerate degraded.The system then logs errors that request a repair action. This situation can occur if inappropriate zoning is applied to the fabric.

Figure 1-18 shows a zoning example (that uses generic aliases) between a two node SVC and a Storwize V5000. Notice that both SAN volume controller nodes have access to the same set of Storwize V5000 ports.

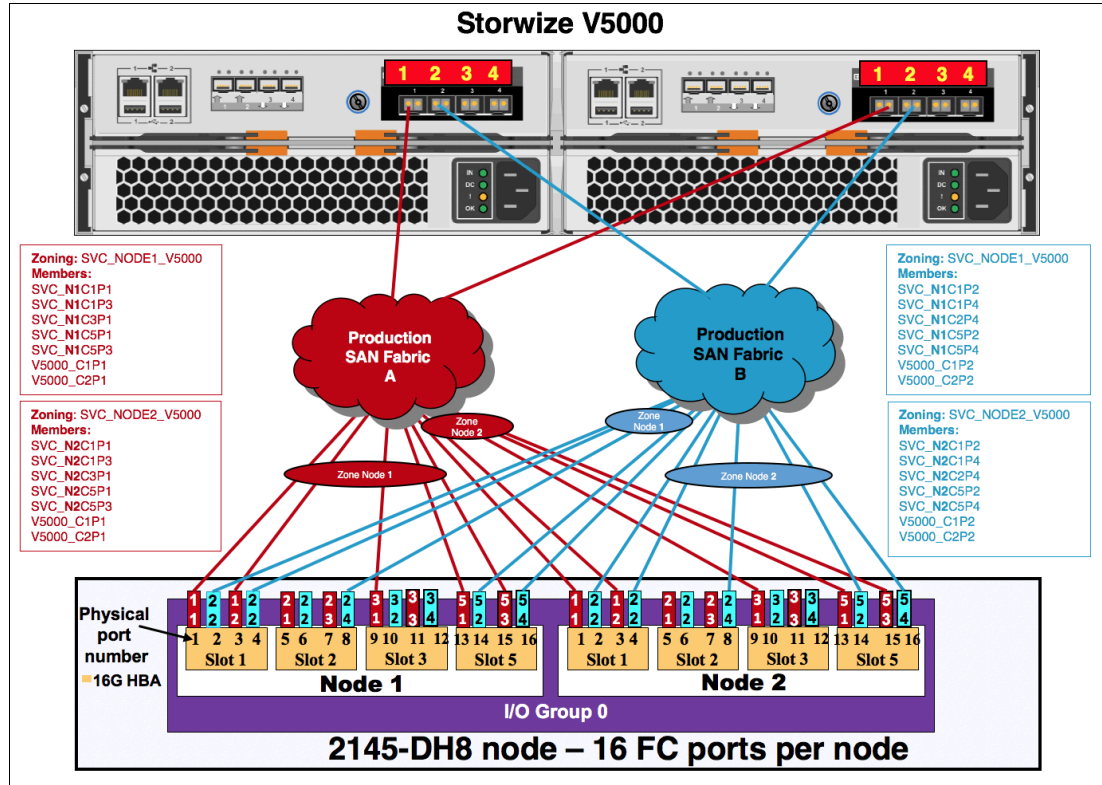


Figure 1-18 Storwize V5000 to SAN Volume Controller zoning

Each storage controller/model has its own preferred zoning and port placement practices. The generic guideline for all storage is to use the ports that are distributed between the redundant storage components, such as nodes, controllers, canisters, and FA adapters (respecting the port count limit described in “Back-end storage port count” on page 23). The following chapters describe the IBM Storage-specific zoning guide lines. Storage vendors other than IBM might have similar preferred practices. For more information, contact your vendor.

Zoning to storage best practice

1.1.2, “ISL considerations” on page 3 of this book details ISL considerations of ensuring that the SAN Volume Controller/Storwize is connected to the same physical switches as the back-end storage ports. 1.2, “SAN topology-specific guidelines” on page 4 reviewed some of the options for SAN Design. This section describes preferred practices for zoning SAN Volume Controller/Storwize ports to controller ports on each of the different SAN designs.

The high-level best-practice is to configure zoning such that the SAN Volume Controller/Storwize ports are zoned only to the controller ports that are attached to the same switch. For single-core designed fabrics, this is not an issue because there is only one switch on each fabric that the SAN Volume Controller/Storwize and controller ports should be connected to. For the mesh/dual-core and other designs where the SAN Volume Controller is connected to multiple switches in the same fabric, then zoning may become an issue.

Figure 1-19 depicts preferred practice zoning on a dual-core fabric. You can see that there are two zones. Zone1 includes only the SAN Volume Controller and back-end ports attached to the core switch on the left. Zone2 includes only the SAN Volume Controller and back-end ports attached to the core switch on the right.

Mesh fabric designs that have the SAN Volume Controller and controller ports connected to multiple switches would follow the same general guidelines. Failure to follow this preferred practice recommendation might result in SAN Volume Controller performance impacts to the fabric. Potential impacts will be covered in the next section.

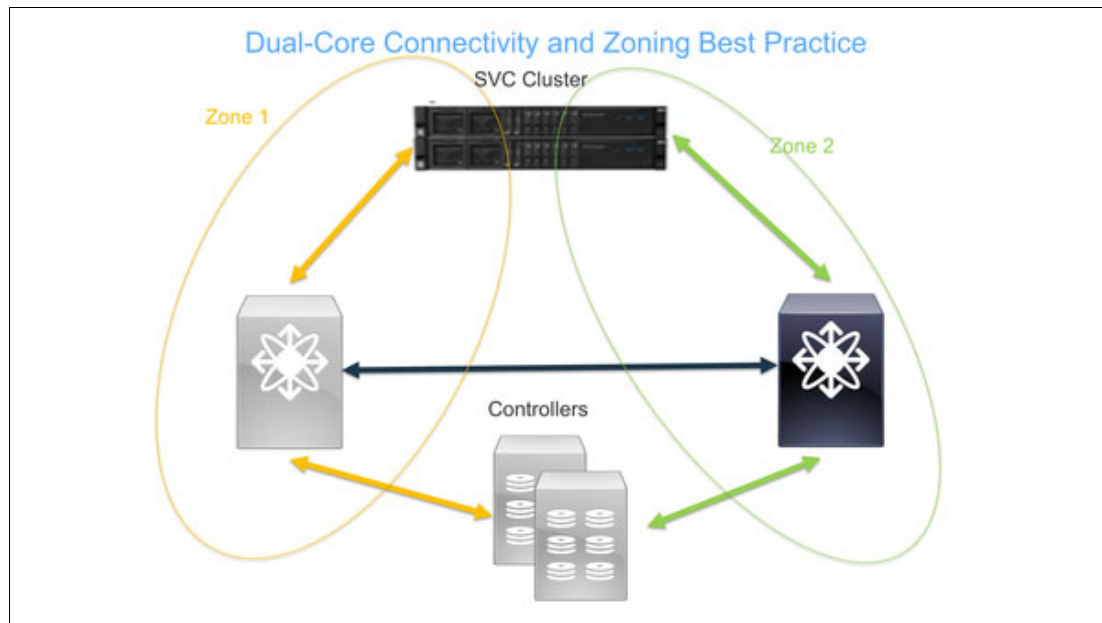


Figure 1-19 Dual core zoning schema

Real-life potential impacts of deviation from best practice zoning

Figure 1-20 on page 23 depicts a design consisting of a dual-core IBM b-type fabric with the SAN Volume Controller cluster attached to one switch and controllers attached to the other. An IBM GPFS™ cluster is attached to the same switch as the controllers. This is a real-world design for a client that was experiencing extreme performance problems on its SAN. The client had dual fabrics, each fabric had this same flawed design.

The design violates both the best practices of ensuring SAN Volume Controller and storage ports are connected to the same switches, and zoning the ports as depicted in Figure 1-19. It also violates the best practice of connecting the host ports (the GPFS cluster) to the same switches as the SAN Volume Controller where possible.

This creates an issue with traffic traversing the ISL unnecessarily as seen in Figure 1-20. I/O requests from the GPFS cluster must traverse the ISL four times. This design should be corrected such that the SAN Volume Controller, controller, and GPFS cluster ports are all connected to both core switches, and zoning is redone to be in accordance with Figure 1-19 on page 22.

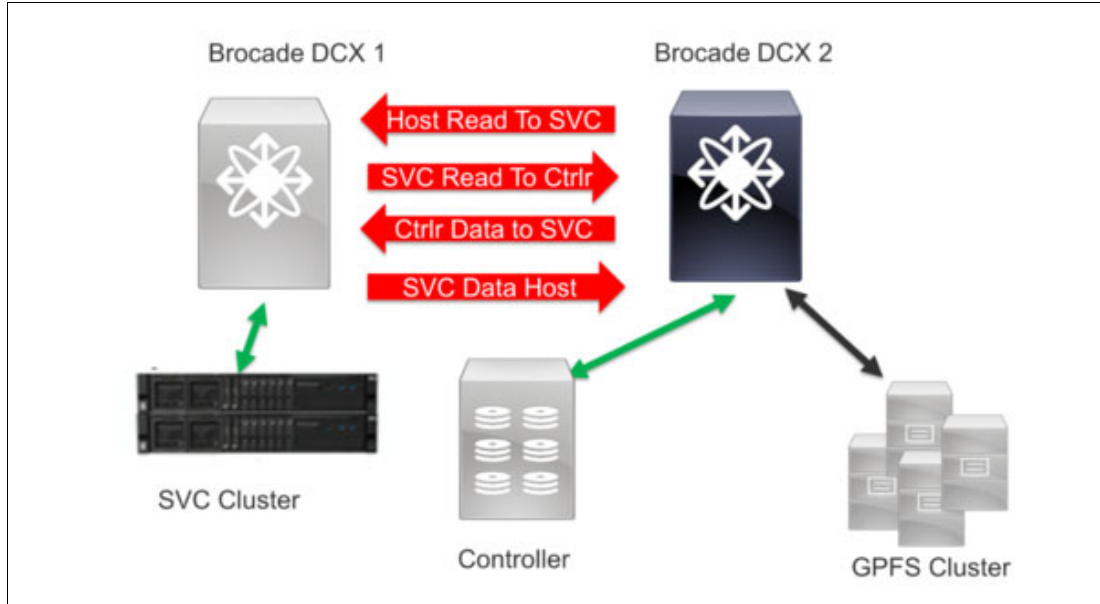


Figure 1-20 ISL traffic overloading

As stated before, Figure 1-20 depicts an actual customer SAN design. The impact of the additional traffic on the ISL between the core switches from this design caused significant delays in command response time from the GPFS cluster to the SAN Volume Controller and from the SAN Volume Controller to the Controller.

The SAN Volume Controller cluster also logged nearly constant errors against the controller including disconnecting from controller ports. The SAN switches logged frequent link time-outs and frame drops on the ISL between the switches. Finally, the customer had other devices sharing the ISL that were not zoned to the SAN Volume Controller. These devices were affected as well.

Back-end storage port count

The current firmware available (V8.1 at the time of writing), sets the limitation of 1024 worldwide node names (WWNNs) per SAN Volume Controller/Storwize cluster and up to 1024 WWPNS. The rule is that each port represents a WWPN count on the SVC cluster. However, the WWNN count differs based on the type of storage.

For example, at the time of writing, EMC DMX/Symmetrix, all HDS storage, and SUN/HP use one WWNN per port. This configuration means that each port appears as a separate controller to the SAN Volume Controller/Storwize. So each port connected to the SAN Volume Controller and Storwize means one WWPN and a WWNN increment.

IBM storage and EMC Clariion/VNX use one WWNN per storage subsystem, so each appears as a single controller with multiple port WWPNS.

The preferred practice is to assign up to sixteen ports from each back-end storage to the SAN Volume Controller/Storwize cluster. The reason of this limitation is that with V8.1, the maximum number of ports are recognized by the SAN Volume controller/Storwize per each WWNN is sixteen. The more ports are assigned, the more throughput is obtained.

In a situation where the back-end storage has hosts direct attached, do not mix the host ports with the SAN Volume Controller and Storwize ports. The back-end storage ports must be dedicated to the SAN Volume Controller and Storwize. Therefore, sharing storage ports are only functional during migration and for a limited time. However, if you intend to have some hosts that are permanently directly attached to the back-end storage, you must segregate the SAN Volume Controller ports from the host ports.

XIV storage subsystem

IBM XIV storage is modular storage and is available as fully or partially populated configurations. XIV hardware configuration can include between 6 and 15 modules. Each additional module added to the configuration increases the XIV capacity, CPU, memory, and connectivity.

From a connectivity standpoint, four Fibre Channel ports are available in each interface module for a total of 24 Fibre Channel ports in a fully configured XIV system. The XIV modules with FC interfaces are present on modules 4 through module 9. Partial rack configurations do not use all ports, even though they might be physically present.

Table 1-2 shows the XIV port connectivity according to the number of installed modules.

Table 1-2 XIV connectivity ports as capacity grows

XIV Modules	Total Ports	Port interfaces	Active port modules
6	8	2	4 and 5
9	16	4	4, 5, 7, and 8
10	16	4	4, 5, 7, and 8
11	20	5	4, 5, 7, 8, and 9
12	20	5	4, 5, 7, 8, and 9
13	24	6	4, 5, 6,7, 8, and 9
14	24	6	4, 5, 6,7, 8, and 9
15	24	6	4, 5, 6,7, 8, and 9

Note: If the XIV has the capacity on demand (CoD) feature, all active Fibre Channel interface ports are usable at the time of installation, regardless of how much usable capacity you purchased. For example, if a 9-module system is delivered with six modules active, you can use the interface ports in modules 4, 5, 7, and 8 even though, effectively, three of the nine modules are not yet activated through CoD.

To use the combined capabilities of SAN Volume Controller/Storwize and XIV, you must connect two ports (one per fabric) from each interface module with the SAN Volume Controller/Storwize ports.

For redundancy and resiliency purposes, select one port from each HBA present on the interface modules. Use port 1 and 3 because both ports are on different HBAs. By default, port 4 is set as a SCSI initiator and is dedicated to XIV replication.

Therefore, if you decide to use port 4 to connect to a SAN Volume controller/Storwize, you must change its configuration from initiator to target. For more information, see *IBM XIV Storage System Architecture and Implementation*, SG24-7659. Figure 1-21 shows how to connect an XIV frame to a SAN Volume Controller storage controller.

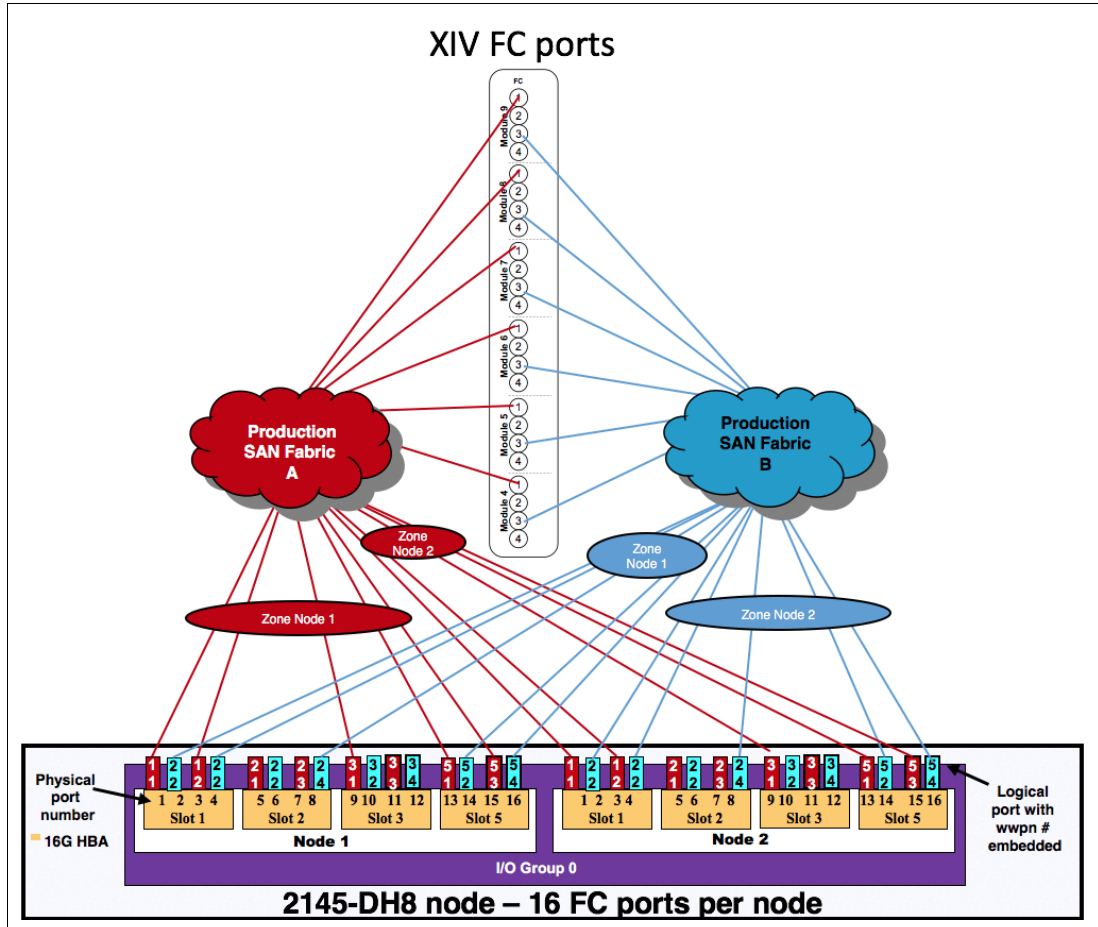


Figure 1-21 Connecting an XIV system as a back-end controller

The preferred practice for zoning is to create a single zoning to each SAN Volume Controller node on each SAN fabric. This zone must contain all ports from a single XIV and the SVC/Storwize V7000 node ports that are destined to connect host and back-end storage. All nodes in an SVC/Storwize V7000 cluster must be able to see the same set of XIV host ports.

Notice that on Figure 1-21, that there is a single zone to each XIV to SAN Volume controller node. So for this example, there are the following different zones:

- ▶ Fabric A, XIV → SVC Node 1: All XIV fabric A ports to SVC node 1
- ▶ Fabric A, XIV → SVC Node 2: All XIV fabric A ports to SVC node 2
- ▶ Fabric B, XIV → SVC Node 1: All XIV fabric B ports to SVC node 1
- ▶ Fabric B, XIV → SVC Node 2: All XIV fabric B ports to SVC node 2

Extra preferred practices and XIV consideration are detailed in Chapter 2, “Back-end storage” on page 49.

FlashSystem A9000 and A9000R storage systems

An IBM FlashSystem® A9000 system has a fixed configuration with three grid elements, with a total of 12 Fibre Channel (FC) ports. A preferred practice is to restrict ports 2 and 4 of each grid controller for replication/migration use, and use ports 1 and 3 for host access.

However, considering that any replication or migration is done through the IBM Spectrum Virtualize/Storwize, it would be possible to use also ports 2 and 4 for IBM Spectrum Virtualize connectivity. Port 4 must be set to target mode for this to work. Assuming a dual fabric configuration for redundancy and resiliency purposes, select one port from each HBA present on the grid controller. So, a total of 6 ports, 3 per fabric, will be used.

Figure 1-22 shows a possible connectivity scheme for SVC 2145-DH8 nodes and A9000 systems.

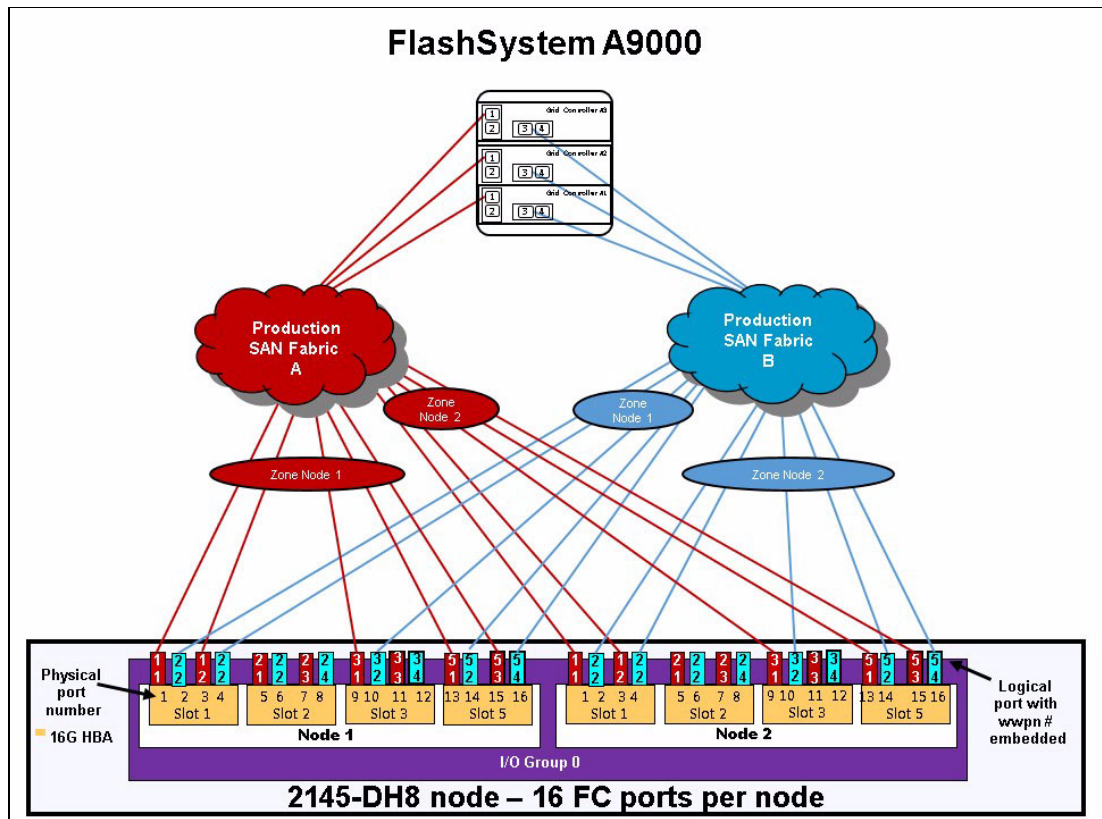


Figure 1-22 Connecting an A9000 system as a back-end controller

The IBM FlashSystem A9000R system has more choices because there are multiple configurations, as shown in Table 1-3.

Table 1-3 Number of host ports in an IBM FlashSystem A9000R system

Grid elements	Total host ports available
2	8
3	12
4	16
5	20
6	24

However, IBM Spectrum Virtualize can support only 16 WWPN from any single WWNN. The IBM FlashSystem A9000 or IBM FlashSystem A9000R system has only one WWNN, so you are limited to 16 ports to any IBM FlashSystem A9000R system.

Next is the same table (Table 1-4), but with columns added to show how many and which ports can be used for connectivity. The assumption is a dual fabric, with ports 1 in one fabric, and ports 3 in the other.

For the 4-grid element system, it is possible to attach 16 ports because that is the maximum that Spectrum Virtualize allows. For the 5- and 6-grid element systems, it is possible to use more ports up to the 16 maximum, but that is not recommended because it might create unbalanced work loads to the grid controllers with two ports attached.

Table 1-4 Host connections to SAN Volume Controller

Grid elements	Total host ports available	Total ports that are connected to Spectrum Virtualize	Total ports that are connected to Spectrum Virtualize
2	8	8	All controllers, ports 1 and 3
3	12	12	All controllers, ports 1 and 3
4	16	8	Odd controllers, port 1 Even controllers, port 3
5	20	10	Odd controllers, port 1 Even controllers, port 3
6	24	12	Odd controllers, port 1 Even controllers, port 3

Figure 1-23 shows a possible connectivity scheme for SVC 2145-DH8 nodes and A9000R systems with up to three grid elements.

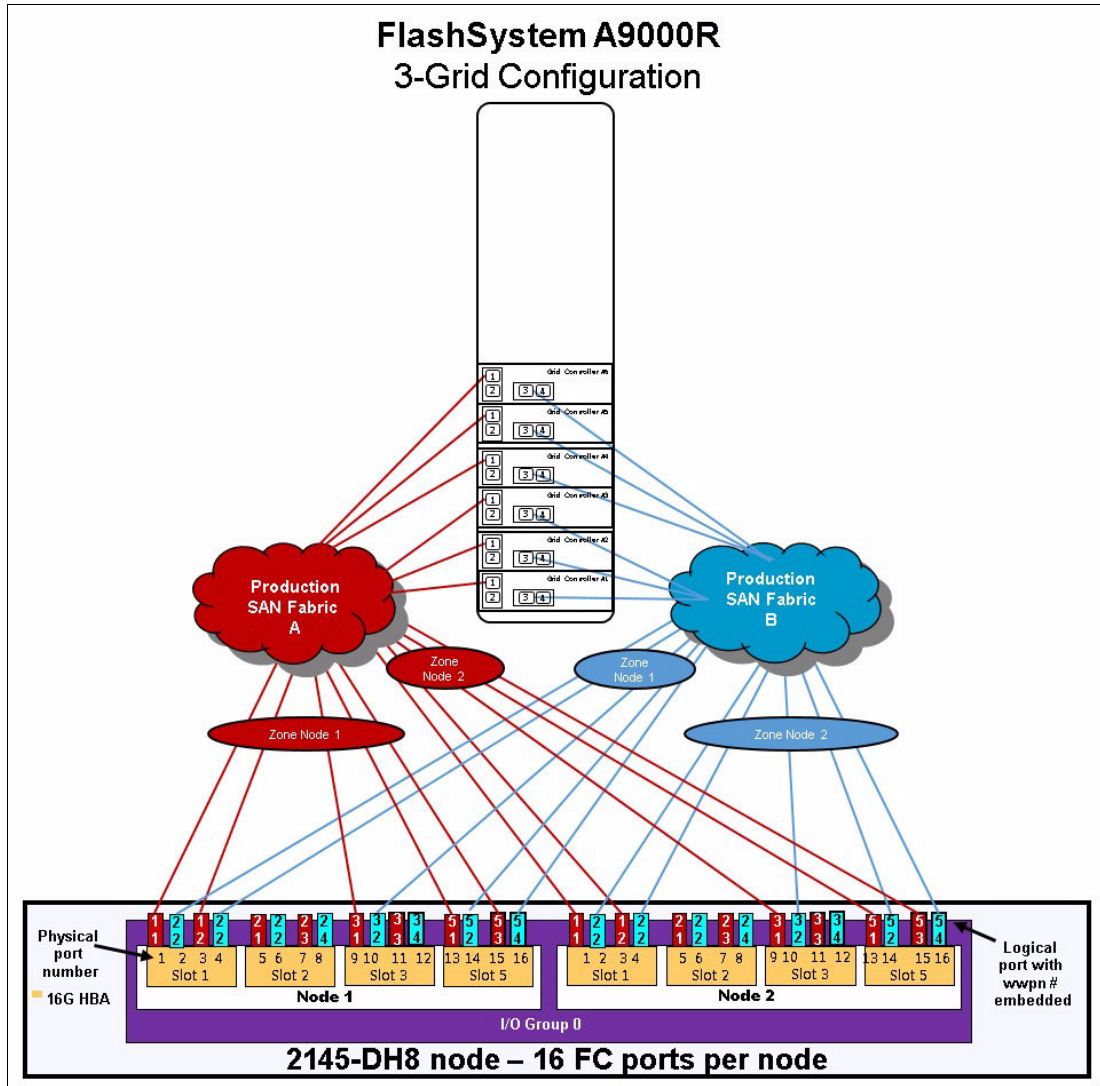


Figure 1-23 Connecting a 3-grid A9000R system as a back-end controller

Finally, Figure 1-24 shows a possible connectivity schema for SVC 2145-DH8 nodes and A9000R systems fully configured.

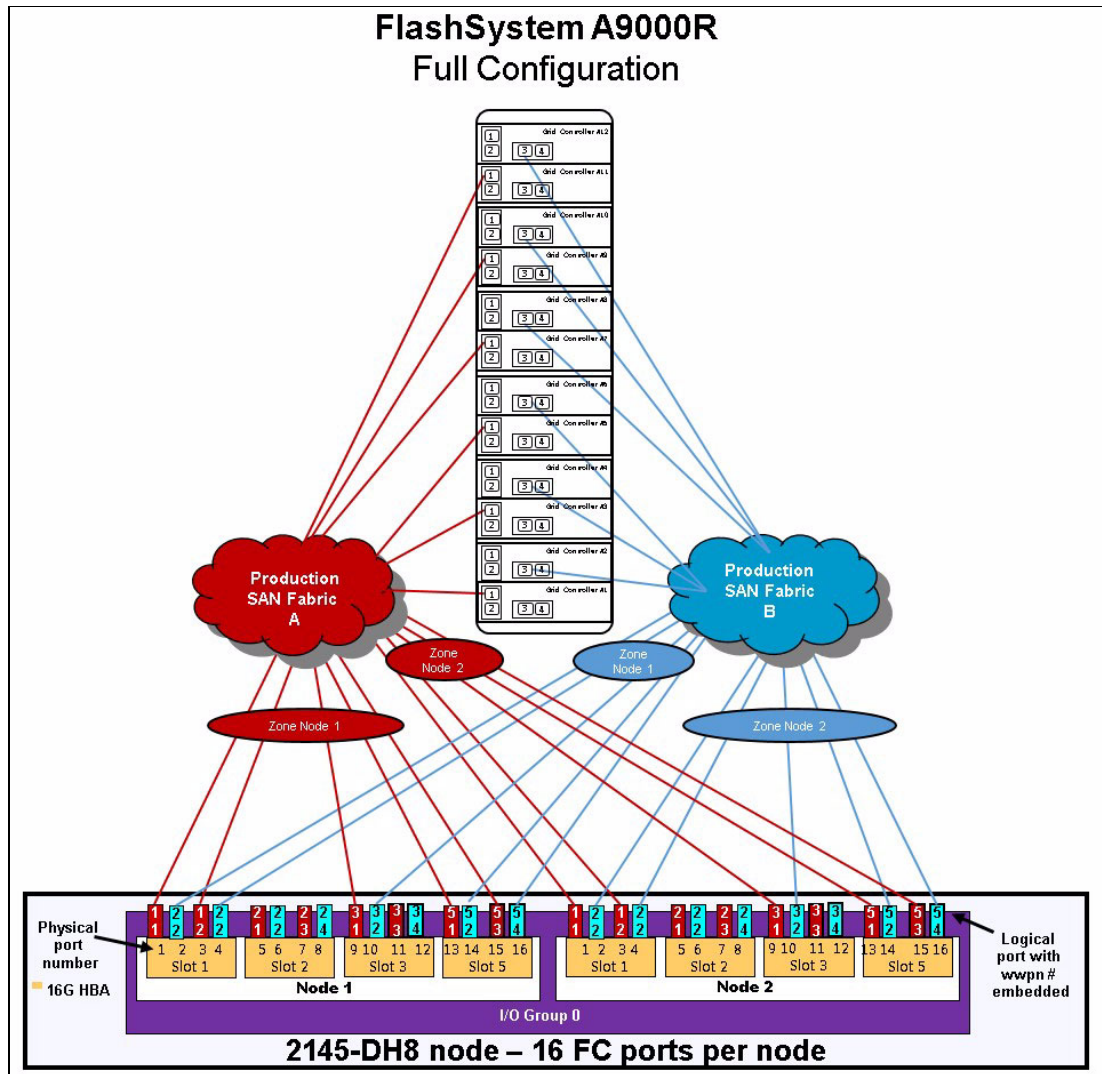


Figure 1-24 Connecting a fully configured A9000R system as a back-end controller

For more information about FlashSystem A9000 and A9000R implementation, see *IBM FlashSystem A9000 and IBM FlashSystem A9000R Architecture and Implementation*, SG24-8345.

Storwize V7000 storage subsystem

Storwize external storage systems can present volumes to a SAN Volume Controller or to another Storwize system. If you want to virtualize one Storwize by using another Storwize, change the *layer* of the Storwize to be used as virtualizer. By default, SAN Volume Controller includes the layer of *replication* and Storwize includes the layer of *storage*.

Volumes forming the storage layer can be presented to the replication layer and are seen on the replication layer as MDisks, but not vice versa. That is, the storage layer cannot see a replication layer's MDisks.

The SAN Volume Controller layer of replication cannot be changed, so you cannot virtualize SAN Volume Controller behind Storwize. However, Storwize can be changed from storage to replication and from replication to storage layer. If you want to virtualize one Storwize behind another, the Storwize used as external storage must have a layer of storage while the Storwize that is performing virtualization must have a layer of replication.

The following are the differences between the storage layer and the replication layer:

- ▶ In the *storage layer*, a Storwize family system has the following characteristics and requirements:
 - The system can complete Metro Mirror and Global Mirror replication with other storage layer systems.
 - The system can provide external storage for replication layer systems or SAN Volume Controller.
 - The system cannot use another Storwize family system that is configured with the storage layer as external storage.
- ▶ In the *replication layer*, a Storwize family system has the following characteristics and requirements:
 - The system can complete Metro Mirror and Global Mirror replication with other replication layer systems or SAN Volume Controller.
 - The system cannot provide external storage for a replication layer system or SAN Volume Controller.
 - The system can use another Storwize family system that is configured with storage layer as external storage.

Note: To change the layer, you must disable the visibility of every other Storwize or SAN Volume Controller on all fabrics. This process involves deleting partnerships, remote copy relationships, and zoning between Storwize and other Storwize or SAN Volume Controller. Then, use the command `chsystem -layer` to set the layer of the system.

You can find additional information about the storage layer in IBM Knowledge Center:

<http://www.ibm.com/support/knowledgecenter/>

To zone the Storwize as a back-end storage controller of SAN Volume Controller, every SAN Volume Controller node must have access to the same Storwize ports, as a minimum requirement. Create one zone per SAN Volume Controller node per fabric to the same ports from a Storwize V7000 storage.

Figure 1-25 shows a zone between a 16-port Storwize V7000 and a SAN Volume Controller.

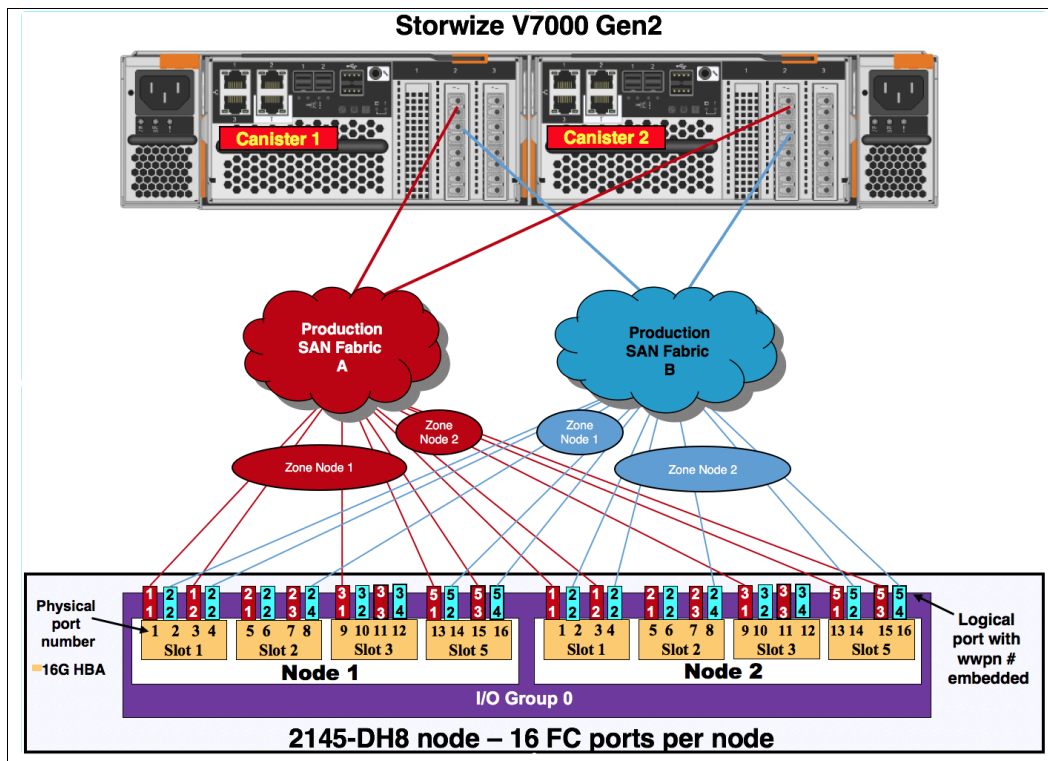


Figure 1-25 Storwize V7000 as a back-end controller zone

Notice that the ports from Storwize V7000 in Figure 1-25 are split between both fabrics. The odd ports are connected to Fabric A and the even ports are connected to Fabric B. You can also spread the traffic across the Storwize V7000 FC adapters on the same canister. However, it will not significantly increase the availability of the solution, because the mean time between failures (MTBF) of the adapters is not significantly less than that of the non-redundant canister components.

Note: If you are using an NPIV enabled Storwize system as backend storage, both the NPIV and physical ports on the Storwize system must be used for the storage backend zoning.

Connect as many ports as necessary to service your workload to the SAN Volume controller. For information about back-end port limitations and preferred practices, see “Back-end storage port count” on page 23.

FlashSystem 900

IBM FlashSystem 900 is an all-flash storage array that provides extreme performance and can sustain highly demanding throughput and low latency across its FC interfaces. It includes up to 16 ports of 8 Gbps or eight ports of 16 Gbps FC. It also provides enterprise-class reliability, large capacity, and green data center power and cooling requirements.

The main advantage of integrating FlashSystem 900 with SAN Volume Controller is to combine the extreme performance of IBM FlashSystem with the SAN Volume Controller enterprise-class solution such as tiering, mirroring, IBM FlashCopy, thin provisioning, IBM Real-time Compression™ and Copy Services.

Before starting, work closely with your IBM Sales, pre-sales, and IT architect to properly size the solution by defining the proper number of SAN Volume Controller I/O groups/cluster and FC ports that are necessary according to your servers and application workload demands.

To maximize the performance that you can achieve when deploying the FlashSystem 900 with SAN Volume Controller, carefully consider the assignment and usage of the FC HBA ports on SAN Volume Controller as described in 1.3.2, “Port naming and distribution” on page 12. The FlashSystem 900 ports must be dedicated to the SAN Volume Controller workload, so do not mix direct attached hosts on FlashSystem 900 with SAN Volume Controller ports.

Connect the FlashSystem 900 to the SAN network in the following manner:

1. Connect FlashSystem 900 odd ports to odd SAN fabric (or SAN Fabric A) and the even ports from to even SAN fabric (or SAN fabric B).
2. Create one zone for each SAN Volume Controller/Storwize node with all FlashSystem 900 ports on each fabric.

Figure 1-26 shows a 16-port FlashSystem 900 zoning to a SAN Volume Controller.

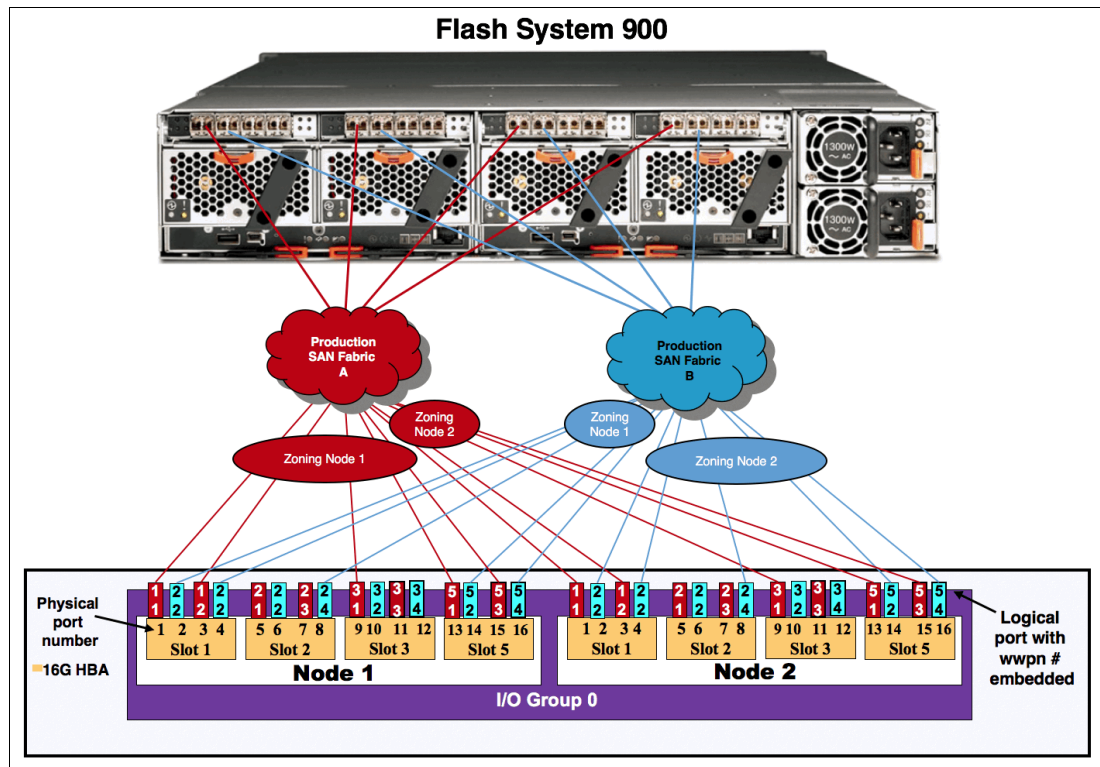


Figure 1-26 FlashSystem 900 to SAN Volume Controller zone

Notice that after the FlashSystem 900 is zoned to two SAN Volume Controller nodes. There are a total of four zones, with one zone per node and two zones per fabric.

You can decide to share or not the SAN Volume Controller and Storwize ports with other back-end storage. however it is important to monitor the buffer credit utilization on SAN Volume Controller switch ports and, if necessary, modify the buffer credit parameters to properly accommodate the traffic to avoid congestion issues.

For additional FlashSystem 900 best practices refer to Chapter 2, “Back-end storage” on page 49.

IBM DS88xx

The IBM DS8000 family is a high-performance, high capacity, highly secure, and resilient series of disk storage systems. The DS888x family is the latest and most advanced of the DS8000 series offerings to date. The high availability, multiplatform support, including IBM z® Systems, and simplified management tools help provide a cost-effective path to an on-demand world.

From a connectivity stand point the DS888x family is scalable. The DS8888 and DS8886 configurations support a maximum of 16 Host Adapters (HA) in the base frame, and an additional 16 Host adapters (HA) in the first expansion frame. The DS8884 configuration supports a maximum of 8 HAs in the base frame and an additional 8 HAs in the first expansion frame. With sixteen 8-port HAs, the maximum number is 128 HA ports. With sixteen 4-port HAs, the maximum number is 64 HA ports.

The 8 Gbps FC Host adapters are available as 4-port and 8-port cards. The 16 Gbps HAs are available as 4-port cards only. The intermixture of both adapters is supported and leads to a different maximum number of ports, as shown in Table 1-5.

Table 1-5 DS8880 port configurations

16 Gbps FC adapters	8 Gbps FC adapters	16 Gbps FC ports	8 Gbps FC ports (4-port/8-port)	Maximum ports
0	16	0	64 - 128	128
1	15	4	60 - 120	124
2	14	8	56 - 112	120
3	13	12	52 - 104	116
4	12	16	48 - 96	112
5	11	20	44 - 88	108
6	10	24	40 - 80	104
7	9	28	36 - 72	100
8	8	32	32 - 64	96
9	7	36	28 - 56	92
10	6	40	24 - 48	88
11	5	44	20 - 40	84
12	4	48	16 - 32	82
13	3	52	12 - 24	78
14	2	56	8 - 16	74
15	1	60	4 - 8	70
16	0	64	0	64

For additional information about DS888x hardware, port, and connectivity, see *IBM DS8880 Architecture and Implementation (Release 8.3)*, SG24-8323.

Despite the wide DS888x port availability, to attach a DS8880 series to a SAN Volume Controller, you must use two to 16 FC Ports, according to your workload. Spread the ports across different HAs for redundancy and resiliency proposes.

Note: To check the current code MAX limitation, search for the term “configuration limits and restrictions” for your current code level at the SAN Volume controller support website:
<http://www.ibm.com/storage/support/2145>

Figure 1-27 shows the connectivity between a SAN Volume Controller and a DS8886.

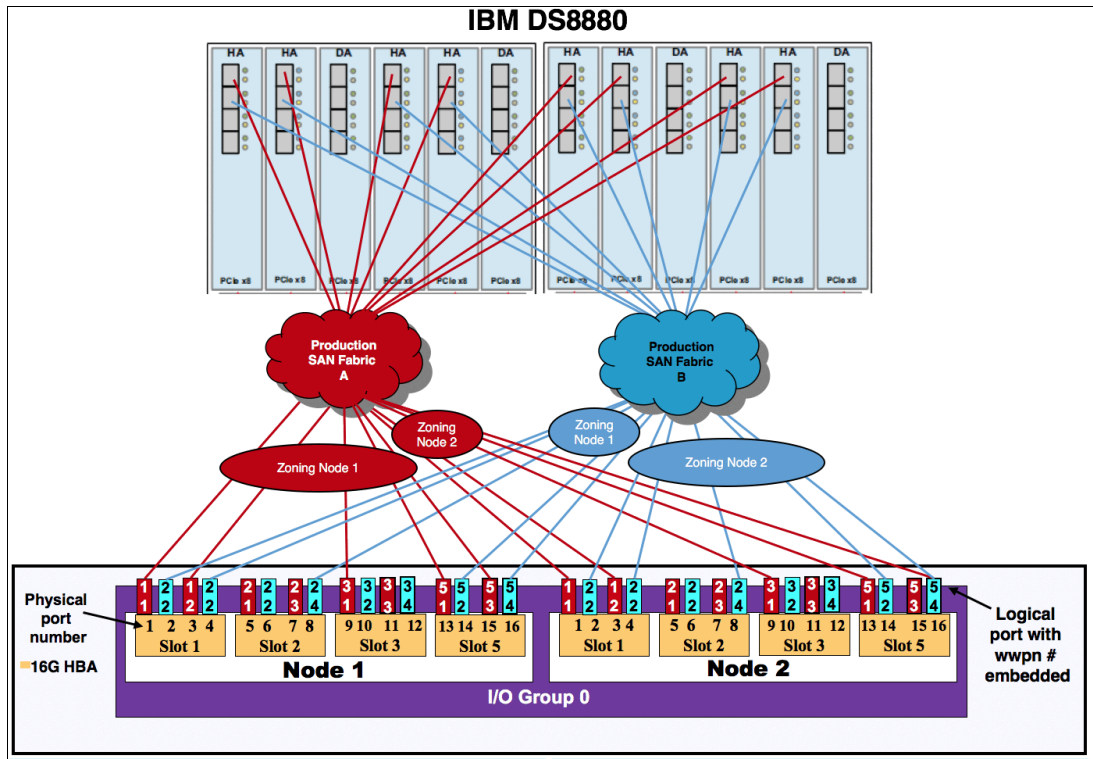


Figure 1-27 DS8886 to SAN volume controller connectivity

Notice that in Figure 1-27, there are 16 ports that are zoned to the SAN Volume Controller and the ports are spread across the different HAs available on the storage.

To maximize performance, the DS888x ports must be dedicated to the SAN Volume Controller connections. On the other hand, the SAN Volume Controller ports must be shared with hosts so you can obtain the maximum full duplex performance from these ports. For a list of port usage and assignment refer to 1.3.2, “Port naming and distribution” on page 12.

Create one zone per SAN Volume Controller node per fabric. The SAN Volume Controller must access the same storage ports on all nodes. Otherwise, the DS888x operation status is set to degraded on the SAN Volume Controller.

After the zoning steps, you must configure the *host connections* using the DS888x CLI (DSCLI) or GUI, to all SVC nodes WWPNs to create a single *Volume Group* adding all SVC cluster ports within this Volume Group. For more information about Volume Group, Host Connection, and DS8000 administration, see *IBM DS8880 Architecture and Implementation (Release 8.3)*, SG24-8323.

The specific preferred practices to present DS8880 LUNs as back-end storage to the SAN Volume Controller are detailed in Chapter 2, “Back-end storage” on page 49.

1.4.5 SAN Volume Controller/Storwize host zones

The preferred practice to connect a host into a SAN volume Controller/Storwize is creating a single zone to each host port. This zone must contain the host port and *one* port from each SAN Volume Controller/Storwize node that the host must access. Although two ports from each node per SAN fabric are in a usual dual-fabric configuration, ensure that the host accesses only one of them, as shown in Figure 1-28.

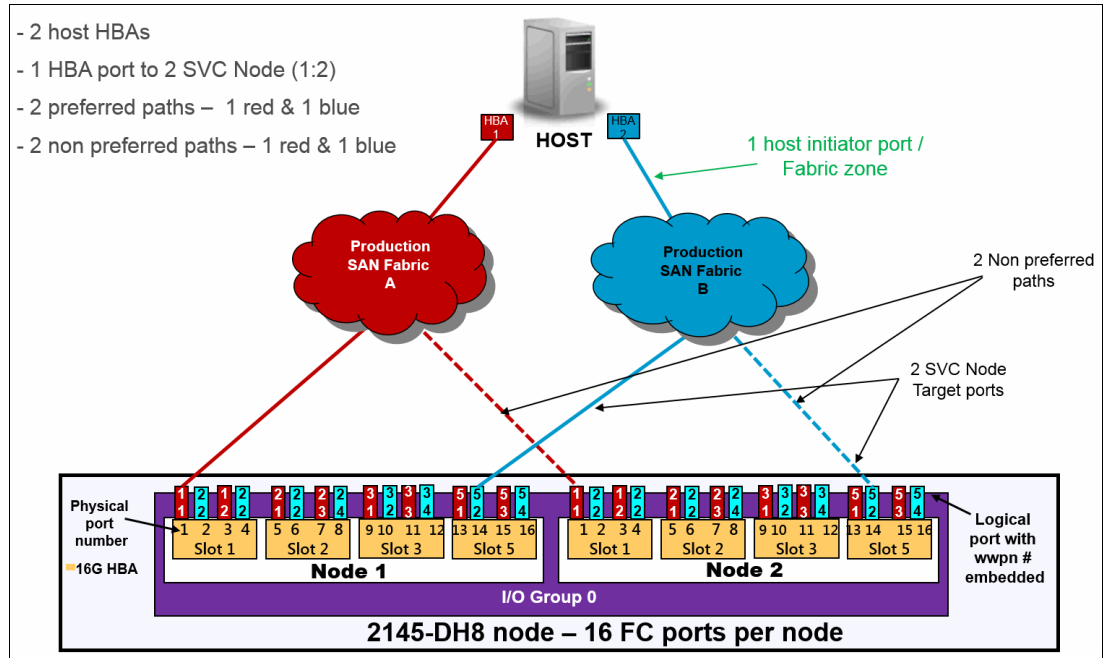


Figure 1-28 Typical host to SAN Volume Controller zoning

This configuration provides four paths to each volume, being two preferred paths (one per fabric) and two non-preferred paths. Four paths is the number of paths, per volume, for which multipathing software such as SDDPCM and SDDDSM, and the SAN Volume Controller/Storwize, are optimized to work with.

NPIV consideration: All the recommendations in this section also apply to NPIV-enabled configurations. For a list of the systems supported by the NPIV, see the following website:

<https://ibm.biz/BdjAHG>

When the recommended number of paths to a volume are exceeded, path failures sometimes are not recovered in the required amount of time. In some cases, too many paths to a volume can cause excessive I/O waits, resulting in application failures and, under certain circumstances, it can reduce performance.

Note: Eight paths by volume is also supported. However, this design provides no performance benefit and, in some circumstances, can reduce performance. Also, it does not significantly improve reliability nor availability. However, fewer than four paths does not satisfy the minimum redundancy, resiliency, and performance requirements.

To obtain the best overall performance of the system and to prevent overloading, the workload to each SAN Volume Controller/Storwize port must be equal. Having the same amount of workload typically involves zoning approximately the same number of host FC ports to each SAN Volume Controller/Storwize FC port.

Hosts with four or more host bus adapters

If you have four HBAs in your host instead of two HBAs, more planning is required. Because eight paths is not an optimum number, configure your SAN Volume Controller/Storwize host definitions (and zoning) as though the single host is two separate hosts. During volume assignment, you alternate which volume was assigned to one of the “pseudo hosts.”

The reason for not assigning one HBA to each path is because the SAN Volume Controller I/O group works as a cluster. When a volume is created, one node is assigned as preferred and the other node solely serves as a backup node for that specific volume. It means that using one HBA to each path will never balance the workload for that particular volume. Therefore, it is better to balance the load by I/O group instead so that the volume is assigned to nodes automatically.

Figure 1-29 shows an example of a four port host zoning.

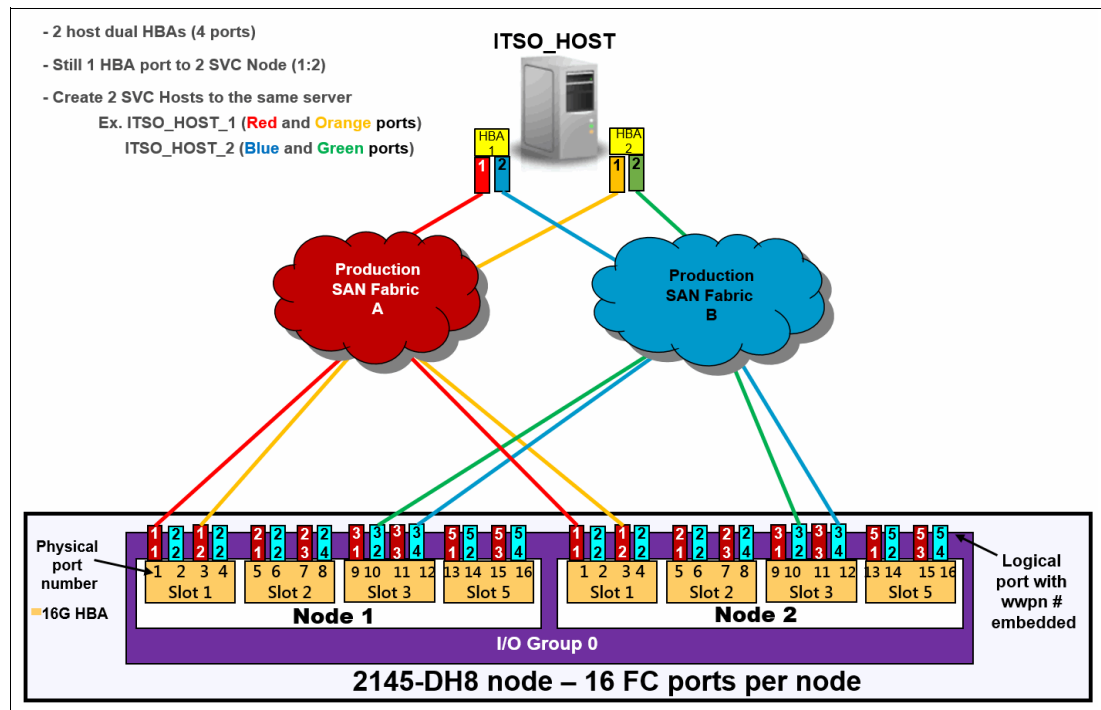


Figure 1-29 Four port host zoning

Because the optimal number of volume paths is four, you must create two or more hosts on SAN Volume Controller and Storwize. During volume assignment, alternate which volume is assigned to one of the “pseudo-hosts,” in a round-robin fashion.

Note: Pseudo-hosts is not a defined function or feature of SAN Volume Controller/Storwize. To create a pseudo-host, you simply need to add another host ID to the SAN Volume Controller and Storwize host configuration. Instead of creating one host ID with four WWPNs, you define two hosts with two WWPNs.

ESX Cluster zoning

For ESX Clusters, you must create separate zones for each host node in the ESX Cluster as shown in Figure 1-30.

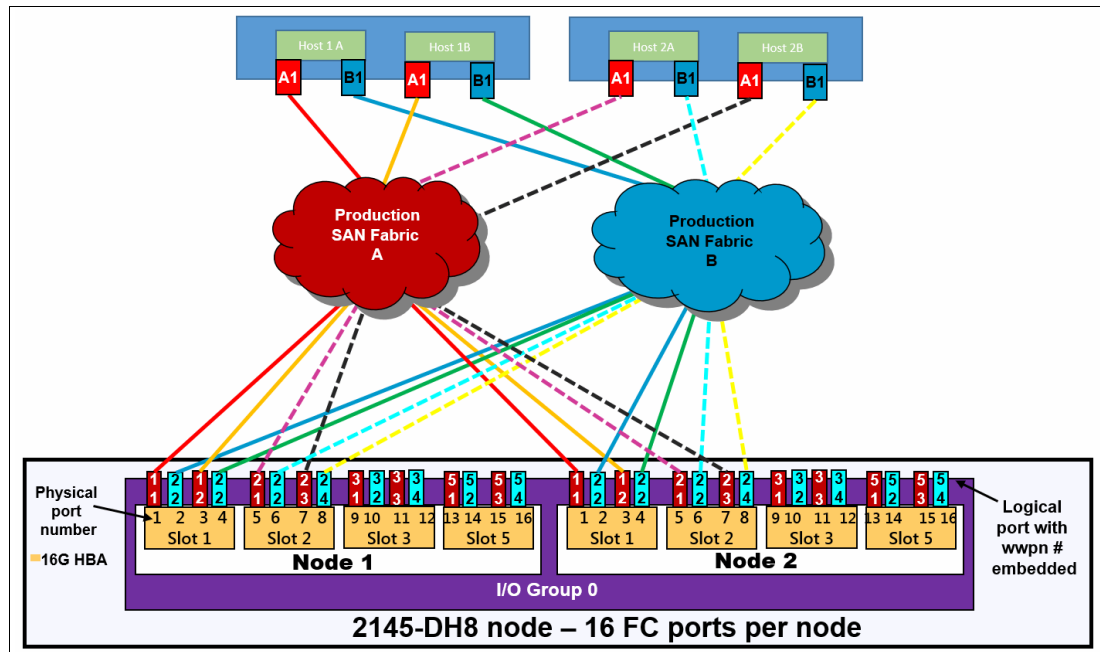


Figure 1-30 ESX Cluster zoning

Ensure that you apply the following preferred practices to your ESX VMware clustered hosts configuration:

- ▶ Zone a single ESX cluster in a manner that avoids ISL I/O traversing.
- ▶ Spread multiple host clusters evenly across the SAN Volume Controller/Storwize node ports and I/O Groups.
- ▶ Map LUNs and volume evenly across zoned ports, alternating the preferred node paths evenly for optimal I/O spread and balance.
- ▶ Create separate zones for each host node in SAN Volume Controller/Storwize and on the ESX cluster.

When allocating a LUN/volume to a clustered system, it is mandatory to manually specify the SCSI BUS ID (SCSI ID) on the SAN Volume Controller/Storwize.

The *SCSI ID* must be the same for every host where the LUN/volume is assigned, as shown in Figure 1-31.

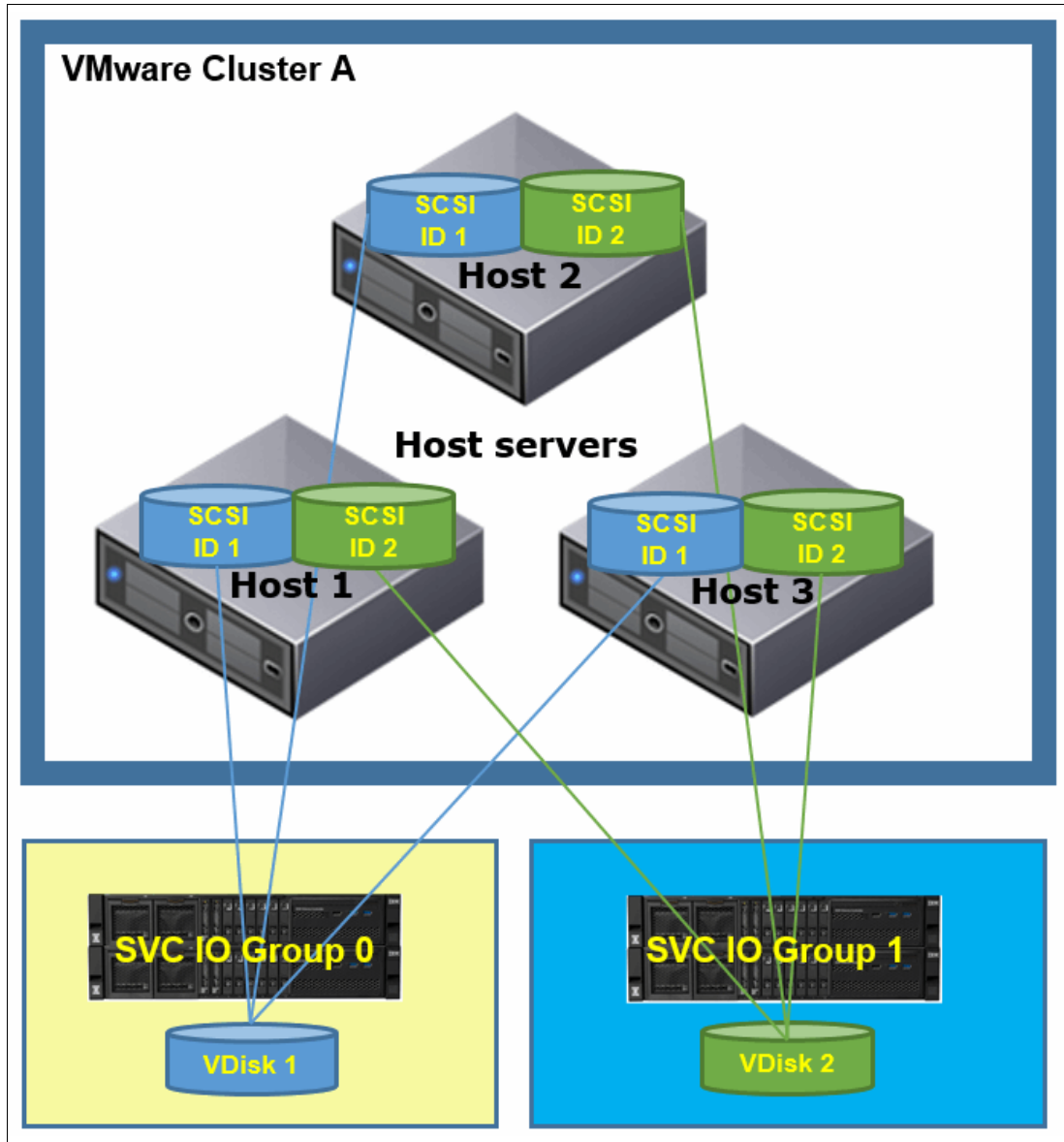


Figure 1-31 LUN/volume mapping to clustered ESX hosts

AIX VIOs: LPM zoning

When zoning IBM AIX VIOs to IBM Spectrum Virtualize, you must plan carefully. Because of its complexity, it is common to create more than four paths to each Volume and MDisk or not provide for proper redundancy. The following preferred practices can help you to have a non-degraded path error on IBM Spectrum Virtualize/Storwize with four paths per volume:

- ▶ Create two separate and isolated zones on each fabric for each LPAR.
- ▶ Do not put both the active and inactive LPAR WWPNs in either the same zone or same IBM Spectrum Virtualize/Storwize host definition.
- ▶ Map LUNs to the virtual host FC HBA port WWPNs, not the physical host FCA adapter WWPN.

- ▶ When using NPIV, generally make no more than a ratio of one physical adapter to eight Virtual ports. This configuration avoids I/O bandwidth oversubscription to the physical adapters.
- ▶ Create a pseudo host in IBM Spectrum Virtualize/Storage host definitions that contain only two virtual WWPNs, one from each fabric as shown in Figure 1-32.
- ▶ Map the LUNs/volumes to the pseudo LPARs (both the active and inactive) in a round-robin fashion.

Figure 1-32 shows a correct SAN connection and zoning for LPARs.

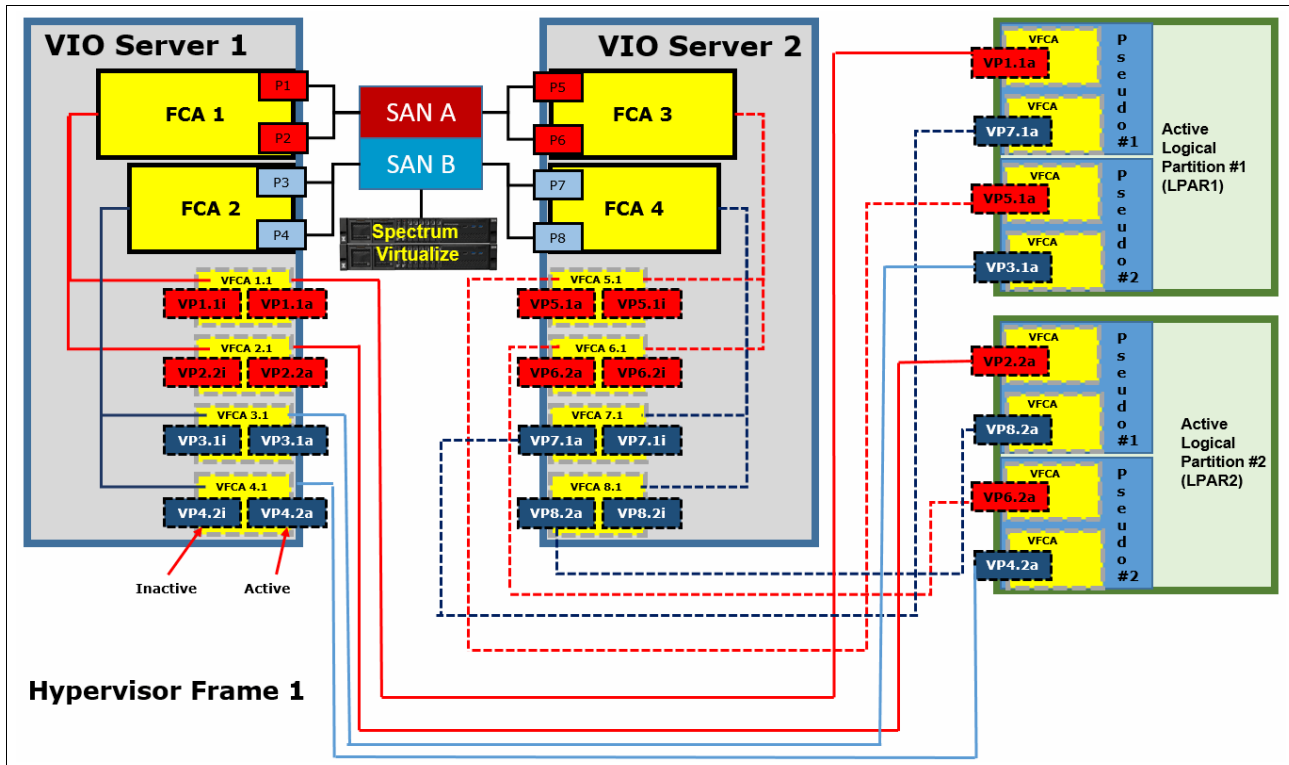


Figure 1-32 LPARs SAN connections

During Live Partition Migration (LPM), both inactive and active ports are active. When LPM is complete, the previously active ports show as inactive and the previously inactive ports show as active.

Figure 1-33 shows a Live partition migration from the hypervisor frame to another frame.

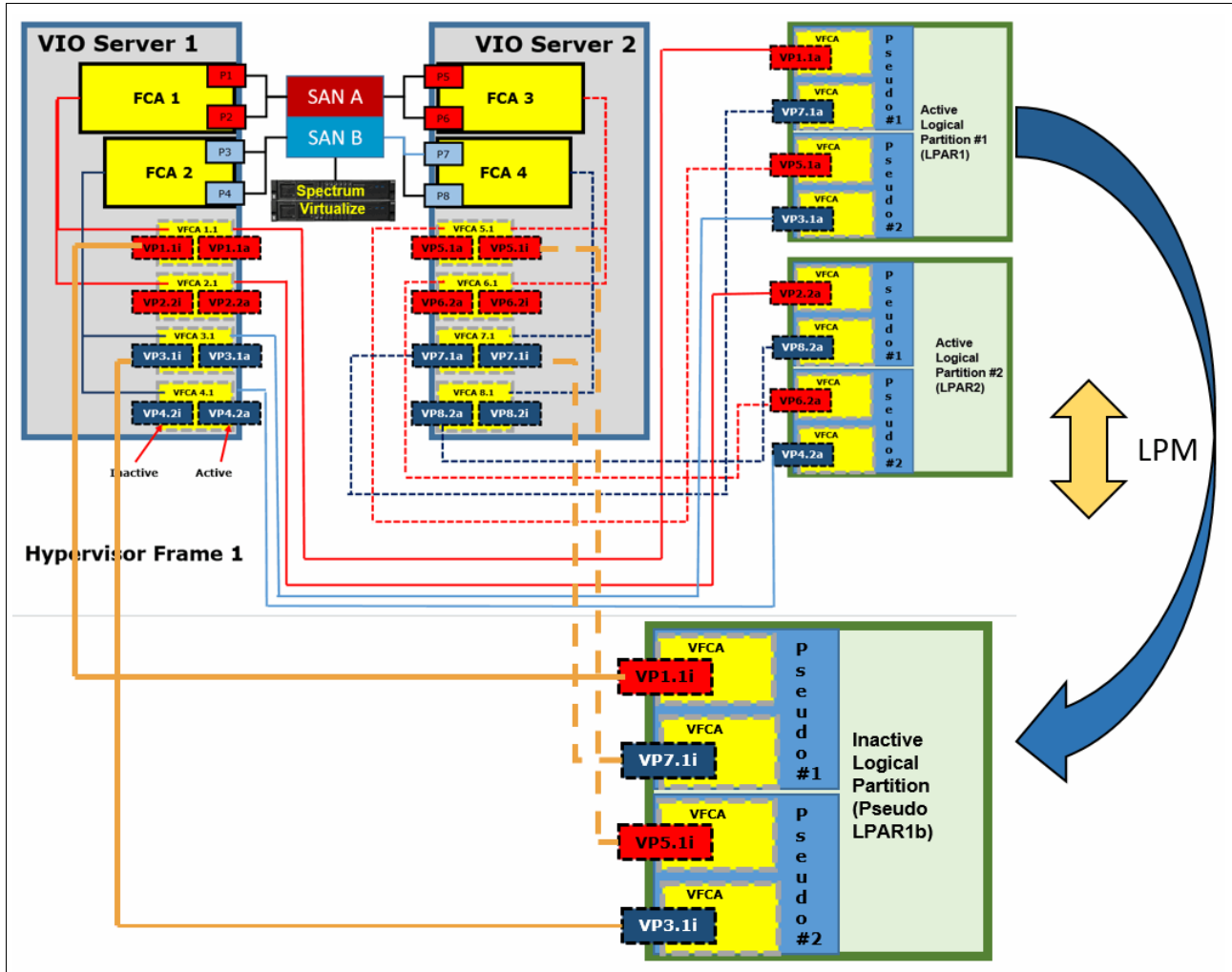


Figure 1-33 Live partition migration

Note: During LPM, the number of paths doubles from 4 to 8. Starting with eight paths per LUN/volume results in an unsupported 16 paths during LPM, which can lead to I/O interruption.

1.4.6 Hot Spare Node zoning considerations

IBM Spectrum Virtualize V8.1 introduced the Hot Spare Node (HSN) feature that provides a higher availability for SAN Volume Controller clusters by automatically swapping a spare node into the cluster if the cluster detects a failing node. Also the maintenance procedures, like code updates and hardware upgrades, benefit from this feature avoiding prolonged loss of redundancy during the node maintenance.

Hot Spare Node is fully described in the IBM Redpaper™ publication *IBM Spectrum Virtualize: Hot Spare Node and NPIV target ports*, REDP-5477, which is available at this website:

<https://www.redbooks.ibm.com/Redbooks.nsf/RedpieceAbstracts/redp5477.html>

For the Hot Spare Node feature to be fully effective requires the NPIV feature enabled. In an NPIV enabled cluster, each physical port is associated with two WWPNs. When the port initially logs into the SAN it uses the normal WWPN (*primary port*), which does not change from previous releases or from NPIV disabled mode. When the node has completed its startup and is ready to begin processing I/O, the *NPIV target ports* log on to the fabric with the second WWPN.

Special zoning requirements must be considered when implementing the HSN functionality.

Host zoning with HSN

Hosts should be zoned with NPIV target ports only. Spare nodes ports must not be included in the host zoning.

Intercluster and intracluster zoning with HSN

Communications between IBM Spectrum Virtualize nodes, including between different clusters, takes place over primary ports. Spare nodes ports must be included in the intracluster zoning likewise the other nodes.

Similarly, when a spare node comes online, its primary ports will be used for remote copy relationships and as such must be zoned with the remote cluster.

Backend controllers zoning with HSN

Backend controllers must be zoned to the primary ports on IBM Spectrum Virtualize nodes. Where a spare node is in use, that nodes ports must be included in the backend zoning, likewise the other nodes.

Note: Currently the zoning configuration for spare nodes is not policed while the spare is inactive and no errors will be logged if the zoning or backend configuration is incorrect.

Backend controller configuration with HSN

IBM Spectrum Virtualize uses the primary ports to communicate with the backend controller, including the spare. This means that all MDisk must be mapped to all IBM Spectrum Virtualize nodes, including spares. For IBM Spectrum Virtualize based backend controllers, such as Storwize V7000, it is recommended that the host clusters functionality is used, with each node forming one host within this cluster. This will ensure that each volume is mapped identically to each IBM Spectrum Virtualize node.

1.4.7 Zoning with multiple SAN Volume Controller/Storwize clustered systems

Unless two separate SAN Volume Controller/Storwize systems participate in a mirroring relationship, configure all zoning so that the two systems do not share a zone. If a single host requires access to two different clustered systems, create two zones with each zone to a separate system.

The back-end storage zones must also be separate, even if the two clustered systems share a storage subsystem. You also must zone separate I/O groups if you want to connect them in one clustered system. Up to four I/O groups can be connected to form one clustered system.

1.4.8 Split storage subsystem configurations

In some situations, a storage subsystem might be used for SAN Volume Controller/Storwize attachment and direct-attach hosts. In this case, pay attention during the LUN masking process on the storage subsystem. Assigning the same storage subsystem LUN to both a host and the SAN Volume Controller/Storwize can result in swift data corruption. If you perform a migration into or out of the SAN Volume Controller/Storwize, make sure that the LUN is removed from one place *before* it is added to another place.

1.5 Distance extension for remote copy services

To implement remote copy services over distance, the following choices are available:

- ▶ Optical multiplexors, such as Dense Wavelength Division Multiplexing (DWDM) or Coarse Wavelength Division Multiplexing (CWDM) devices
- ▶ Long-distance SFPs and XFPs
- ▶ FC-to-IP conversion boxes
- ▶ Native IP-based replication with SAN Volume Controller/Storwize code

Of these options, the optical varieties of distance extension are preferred. IP distance extension introduces more complexity, is less reliable, and has performance limitations. However, optical distance extension is impractical in many cases because of cost or unavailability.

1.5.1 Optical multiplexors

Optical multiplexors can extend your SAN up to hundreds of kilometers at high speeds. For this reason, they are the preferred method for long-distance expansion. When you are deploying optical multiplexing, make sure that the optical multiplexor is certified to work with your SAN switch model. The SAN Volume Controller/Storwize has no allegiance to a particular model of optical multiplexor.

If you use multiplexor-based distance extension, closely monitor your physical link error counts in your switches. Optical communication devices are high-precision units. When they shift out of calibration, you start to see errors in your frames.

1.5.2 Long-distance SFPs or XFPs

Long-distance optical transceivers have the advantage of extreme simplicity. Although no expensive equipment is required, a few configuration steps are necessary. Ensure that you use transceivers that are designed for your particular SAN switch *only*. Each switch vendor supports only a specific set of SFP or XFP transceivers, so it is unlikely that Cisco SFPs will work in an IBM b-type switch.

1.5.3 Fibre Channel over IP

Fibre Channel over IP (FCIP) conversion is by far the most common and least expensive form of distance extension. FCIP is a technology that allows FC routing to be implemented over long distances by using the TCP/IP protocol. In most cases, FCIP is implemented in Disaster Recovery scenarios with some kind of data replication between the primary and secondary site.

FCIP is a tunneling technology, which means FC frames are encapsulated in the TCP/IP packets. As such, it is not apparent to devices that are connected through the FCIP link. To use FCIP, you need some kind of tunneling device on both sides of the TCP/IP link that integrates FC and Ethernet connectivity. Most of the SAN vendors offer FCIP capability either through stand-alone devices (Multiprotocol routers) or using blades integrated in the director class product. Both SAN Volume Controller and Storwize family systems support FCIP connection.

An important aspect of the FCIP scenario is the IP link quality. With IP-based distance extension, you must dedicate bandwidth to your FC to IP traffic if the link is shared with other IP traffic. Because the link between two sites is low-traffic or used only for email, do not assume that this type of traffic is always the case. The design of FC is sensitive to congestion and you do not want a spyware problem or a DDOS attack on an IP network to disrupt your SAN Volume Controller/Storwize.

Also, when you are communicating with your organization's networking architects, distinguish between megabytes per second (MBps) and megabits per second (Mbps). In the storage world, bandwidth often is specified in MBps, but network engineers specify bandwidth in Mbps. If you fail to specify MB, you can end up with an impressive-sounding 155 Mbps OC-3 link, which supplies only 15 MBps or so to your SAN Volume Controller/Storwize. If you include the safety margins, this link is not as fast as you might hope, so ensure that the terminology is correct.

Consider the following steps when you are planning for your FCIP TCP/IP links:

- ▶ For redundancy purposes use as many TCP/IP links between sites as you have fabrics in each site that you want to connect. In most cases, there are two SAN FC fabrics in each site, so you need two TCP/IP connections between sites.
- ▶ Try to dedicate TCP/IP links only for storage interconnection. Separate them from other LAN/WAN traffic.
- ▶ Make sure that you have a service level agreement (SLA) with your TCP/IP link vendor that meets your needs and expectations.
- ▶ If you do not use Global Mirror with Change Volumes (GMCV), make sure that you have sized your TCP/IP link to sustain peak workloads.
- ▶ The use of SAN Volume Controller/Storwize internal Global Mirror (GM) simulation options can help you test your applications before production implementation. You can simulate the GM environment within one SAN Volume Controller or one Storwize system without partnership with another. Use the **chsystem** command with the following parameters to perform GM testing:
 - **gminterdelaysimulation**
 - **gmintradelaysimulation**

Further details on GM planning are described in Chapter 5, "Copy services" on page 139.

- ▶ If you are not sure about your TCP/IP link security, enable Internet Protocol Security (IPSec) on the all FCIP devices. IPSec is enabled on the Fabric OS level, so you do not need any external IPSec appliances.

In addition to planning for your TCP/IP link, consider adhering to the following preferred practices:

- ▶ Set the link bandwidth and background copy rate of partnership between your replicating SAN Volume Controller/Storwize to a value *lower* than your TCP/IP link capacity. Failing to do that can cause an unstable TCP/IP tunnel, which can lead to stopping all your remote copy relations that use that tunnel.
- ▶ The best case is to use GMCV when replication is done over long distances.

- ▶ Use compression on corresponding FCIP devices.
- ▶ Use at least two ISLs from your local FC switch to local FCIP router.
- ▶ On an IBM b-type SAN, use the Integrated Routing feature to avoid merging fabrics from both sites.

For more information about FCIP, see the following publications:

- ▶ *IBM System Storage b-type Multiprotocol Routing: An Introduction and Implementation, SG24-7544*
- ▶ *IBM/Cisco Multiprotocol Routing: An Introduction and Implementation, SG24-7543*

1.5.4 SAN extension with Business Continuity configurations

Spectrum Virtualize Enhanced Stretched Cluster and HyperSwap technologies provide Business Continuity solutions over metropolitan areas with distances up to 300 km. Usually this is achieved using SAN extension over WDM technology. Furthermore, in order to avoid single points of failure, multiple WDMs and physical links are implemented. When implementing these solutions, particular attention must be paid in the intercluster connectivity set up.

Consider a typical implementation of an Enhanced Stretched Cluster using ISLs, as shown in Figure 1-34.

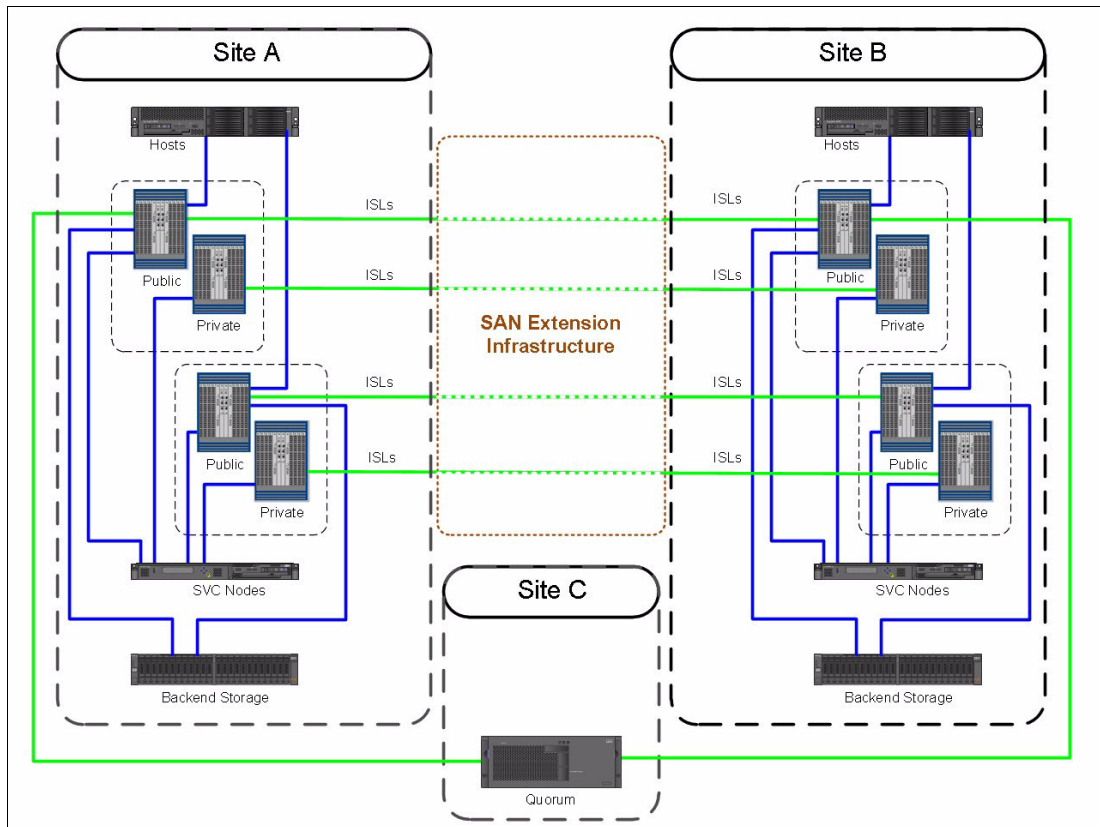


Figure 1-34 Typical Enhanced Stretched Cluster configuration

In this configuration, the intercluster communication is isolated in a Private SAN that interconnects Site A and Site B through a SAN extension infrastructure consisting of two DWDMs. Let's assume that, for redundancy reasons, two ISLs are used for each fabric for the Private SAN extension.

There are basically two possible configurations to interconnect the Private SANs. In the Configuration 1, shown in Figure 1-35, one ISL per fabric is attached to each DWDM. In this case, the physical paths Path A and Path B are used to extend both fabrics.

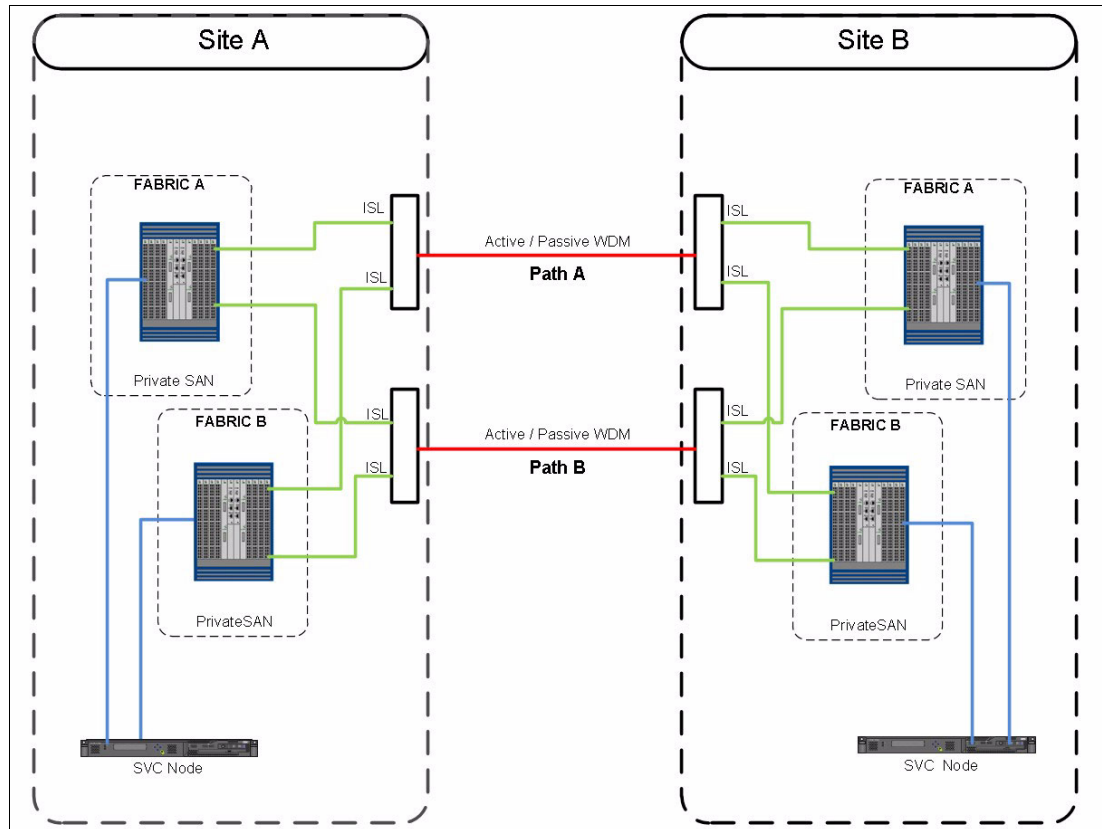


Figure 1-35 Configuration 1: physical paths shared among the fabrics

In Configuration 2, shown in Figure 1-36, ISLs of fabric A are attached only to Path A, while ISLs of fabric B are attached only to Path B. In this case the physical paths are not shared between the fabrics.

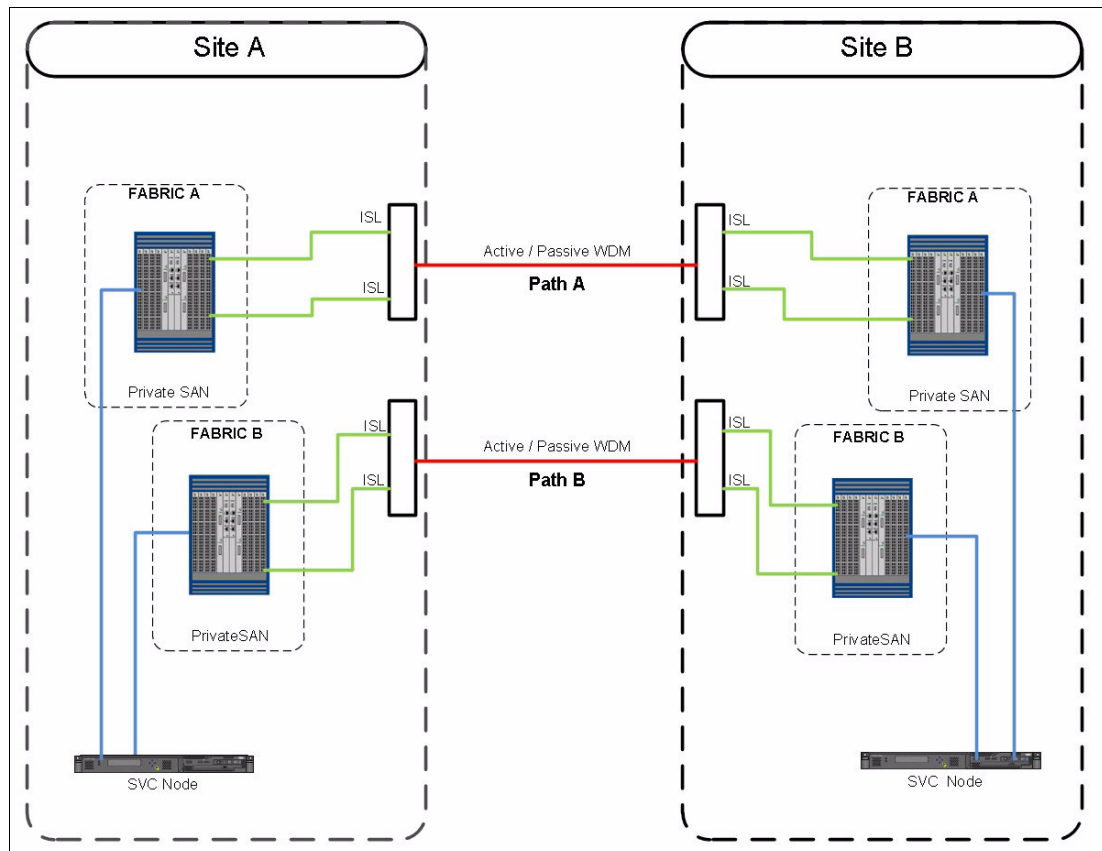


Figure 1-36 Configuration 2: physical paths not shared among the fabrics

With Configuration 1, in case of failure of one of the physical paths, both fabrics are simultaneously affected and a fabric reconfiguration occurs because of an ISL loss. This situation could lead to a temporary disruption of the intracluster communication and, in the worst case, to a split brain condition. To mitigate this situation, link aggregation features like IBM b-type ISL trunking can be implemented.

With Configuration 2, a physical path failure leads to a fabric segmentation of one of the two fabrics, leaving the other fabric unaffected. In this case the intracluster communication would be guaranteed through the unaffected fabric.

Summarizing, the recommendation is to fully understand the implication of a physical path or DWDM loss in the SAN extension infrastructure and implement the appropriate architecture in order to avoid a simultaneous impact.

1.5.5 Native IP replication

It is possible to implement native IP-based replication. *Native* means that SAN Volume Controller/Storwize does not need any FCIP routers to create a partnership. This partnership is based on the Internet Protocol network and not on the FC network. For more information about native IP replication, see Chapter 5, “Copy services” on page 139.

To enable native IP replication, SAN Volume Controller/Storwize implements the Bridgeworks SANSlide network optimization technology. For more information about this solution, see *IBM SAN Volume Controller and Storwize Family Native IP Replication*, REDP-5103.

1.6 Tape and disk traffic that share the SAN

If you have free ports on your core switch, you can place tape devices (and their associated backup servers) on the SAN Volume Controller/Storwize SAN. However, do not put tape and disk traffic on the same FC HBA.

Do not put tape ports and backup servers on different switches. Modern tape devices have high-bandwidth requirements. Placing tape ports and backup servers on different switches can quickly lead to SAN congestion over the ISL between the switches.

1.7 Switch interoperability

SAN Volume Controller/Storwize is flexible as far as switch vendors are concerned. All of the node connections on a particular SAN Volume Controller/Storwize clustered system must go to the switches of a single vendor. That is, you must not have several nodes or node ports plugged into vendor A and several nodes or node ports plugged into vendor B.

SAN Volume Controller/Storwize supports some combinations of SANs that are made up of switches from multiple vendors in the same SAN. However, this approach is not preferred in practice. Despite years of effort, interoperability among switch vendors is less than ideal because FC standards are not rigorously enforced. Interoperability problems between switch vendors are notoriously difficult and disruptive to isolate. Also, it can take a long time to obtain a fix. For these reasons, run only multiple switch vendors in the same SAN long enough to migrate from one vendor to another vendor, if this setup is possible with your hardware.

You can run a mixed-vendor SAN if you have agreement from both switch vendors that they fully support attachment with each other.

Interoperability between Cisco switches and IBM b-type switches is not recommended, except during fabric migrations, and then only if you have a back-out plan in place. Also, when connecting BladeCenter switches to a core switch, consider the use of the N-Port ID Virtualization (NPIV) technology.

When you have SAN fabrics with multiple vendors, pay special attention to any particular requirements. For example, observe from which switch in the fabric the zoning must be performed.



Back-end storage

This chapter describes aspects and characteristics to consider when you plan the attachment of a back-end storage device to be virtualized by an IBM System Storage SAN Volume Controller or Storwize.

This chapter includes the following sections:

- ▶ Storage controller path selection
- ▶ Considerations for DS8000 series
- ▶ Considerations for IBM XIV Storage System
- ▶ Considerations for IBM FlashSystem A9000/A9000R
- ▶ Considerations for IBM Storwize V7000/V5000/V3700
- ▶ Considerations for IBM FlashSystem 900
- ▶ Considerations for storage subsystem compression and deduplication capability
- ▶ Considerations for third-party storage with EMC VMAX and Hitachi Data Systems

2.1 Storage controller path selection

When a managed disk (MDisk) logical unit (LU) is accessible through multiple storage system ports, the system ensures that all nodes that access this LU coordinate their activity and access the LU through the same storage system port.

An MDisk path that is presented to the storage system for all system nodes must meet the following criteria.

- ▶ The system node is a member of a storage system
- ▶ The system node has Fibre Channel or iSCSI connections to the storage system port
- ▶ The system node has successfully discovered the LU
- ▶ The port selection process has not caused the system node to exclude access to the MDisk through the storage system port

When the IBM SAN Volume Controller nodes or Storwize canisters select a set of ports to access the storage system, the two types of path selection described in the next sections are supported to access the MDisks.

2.1.1 Round robin

Before V6.3 of SAN Volume Controller/Storwize code, all I/O to a particular MDisk was issued through only one back-end storage controller FC port. Even if there were 12 (XIV) or 16 (DS8000) FC ports that were zoned to SAN Volume Controller or Storwize, one MDisk was using only one port. If there was a port failure, another port on the backend storage controller was chosen.

This configuration changed in SAN Volume Controller/Storwize V6.3. Since V6.3, each MDisk used one path per target port per SAN Volume Controller/Storwize node. This change means that, in cases of storage systems without a preferred controller such as XIV or DS8000, each MDisk uses all of the available FC ports of that storage controller.

Note: With a round-robin compatible storage controller, there is no need to create as many volumes as there are storage FC ports anymore. Every volume, and therefore MDisk, on SAN Volume Controller/Storwize uses all available ports.

This configuration results in significant performance increase because MDisk is no longer bound to one backend FC port. Instead, it can issue IOs too many backend FC ports in parallel. Particularly, the sequential I/O within a single extent can benefit from this feature.

Additionally, round-robin path selection improves resilience to certain storage system failures. For example, if one of the backend storage system FC ports has some performance problems, the I/O to MDisks is sent through other ports. Moreover, because IOs to MDisks are sent through all backend storage FC ports, the port failure can be detected more quickly.

Preferred practice: If you have SAN Volume Controller/Storwize code V6.3 or later, zone as many FC ports from the backend storage controller to SAN Volume Controller/Storwize as possible. SAN Volume Controller/Storwize supports up to 16 FC ports per storage controller. See your storage system documentation for FC port connection and zoning guidelines.

Example 2-1 shows a storage controller that supports round-robin path selection.

Example 2-1 Round robin enabled storage controller

```
# lsmdisk 4
id 4
name mdisk4
status online
mode managed
mdisk_grp_id 0
mdisk_grp_name HPEe7200-3Pr-grp
capacity 250.0GB
quorum_index 0
block_size 512
controller_name controller2
ctrl_type 4
ctrl_WWNN 2FF70002AC00C202
controller_id 2
path_count 8
max_path_count 8
ctrl_LUN_# 0000000000000001
UID 60002ac00000000000000080000c202000000000000000000000000000000
preferred_WWPN
active_WWPN many ·Round Robin Enabled
fast_write_state empty
raid_status
raid_level
redundancy
strip_size
spare_goal
spare_protection_min
balanced
tier enterprise
slow_write_priority
fabric_type fc
site_id
site_name
easy_tier_load medium
encrypt no
```

2.1.2 MDisk group balanced and controller balanced

Although round-robin path selection provides optimized and balanced performance with minimum configuration required, there are storage systems that still require manual intervention to achieve the same goal.

For example, when virtualizing legacy storage subsystems, such as IBM DS4000® and DS3000, SAN Volume Controller/Storwize accesses an MDisk (LU) through one of the ports on the preferred controller. In order to best utilize the back-end storage, it is important to make sure that the number of LUs created is a multiple of the connected FC ports and aggregate all LUs to a single MDisk group.

Example 2-2 shows a storage controller that supports MDisk group balanced path selection.

Example 2-2 MDisk group balanced path selection (no round robin enabled) storage controller

```
# lsmdisk 4
id 4
name mdisk4
status online
mode managed
mdisk_grp_id 0
mdisk_grp_name HPEe7200-3Pr-grp
capacity 250.0GB
quorum_index 0
block_size 512
controller_name controller2
ctrl_type 4
ctrl_WWNN 2FF70002AC00C202
controller_id 2
path_count 2
max_path_count 2
ctrl_LUN_# 0000000000000001
UID60002ac00000000000000080000c202000000000000000000000000000000
preferred_WWPN
active_WWPN 20110002AC00C202 · Mdisk group balancing
fast_write_state empty
raid_status
raid_level
redundancy
strip_size
spare_goal
spare_protection_min
balanced
tier enterprise
slow_write_priority
fabric_type fc
site_id
site_name
easy_tier_load medium
encrypt no
```

For more information about the latest updates of this list, see *Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for SAN Volume Controller*, for your current firmware level which is available at:

<http://www.ibm.com/support/docview.wss?uid=ssg1S1003658>

2.2 Considerations for DS8000 series

Although all recommendations in this chapter are true for SAN Volume Controller and Storwize family storage systems, the DS8000 series might be virtualized behind one of the Storwize storage systems for a short period of time. Sometimes, it might be virtualized by Storwize only for data migration purposes.

2.2.1 Connectivity considerations

Configure a minimum of eight controller ports to the SAN Volume Controller per controller, regardless of the number of SAN Volume Controller nodes or Storwize canisters in the cluster. Configure up to 16 controller ports for large controller configurations where more than 40 ranks are presented to the SAN Volume Controller or Storwize cluster. Currently, 16 ports per storage subsystem are the maximum that is supported by the SAN Volume Controller and Storwize.

Generally, use ports from different host adapters and if possible from different I/O enclosures. This configuration is also important because during a DS8000 LIC update, a host adapter port might need to be taken offline. This configuration allows the SAN Volume Controller or Storwize I/O to survive a hardware failure on any component on the SAN path.

The number of ports to be used varies according to the number of ranks that are virtualized by SAN Volume Controller:

- ▶ Use eight ports to support up to 40 ranks.
- ▶ Use 16 ports (the maximum supported by SAN Volume Controller) for 40+ ranks.

For more information about SAN preferred practices and connectivity, see Chapter 1, “Storage area network” on page 1.

2.2.2 Defining storage

To optimize the DS8000 resource utilization, use the following guidelines:

- ▶ Distribute capacity and workload across device adapter pairs.
- ▶ Balance the ranks and extent pools between the two DS8000 internal servers to support the corresponding workloads on them.
- ▶ Spread the logical volume workload across the DS8000 internal servers by allocating the volumes equally on rank groups 0 and 1.
- ▶ Use as many disks as possible. Avoid idle disks, even if all storage capacity is not to be used initially.
- ▶ Use multi-rank extent pools.
- ▶ Stripe your logical volume across several ranks (the default for multi-rank extent pools).

Balancing workload across DS8000 series controllers

When you configure storage on the DS8000 series disk storage subsystem, ensure that ranks on a device adapter (DA) pair are evenly balanced between odd and even extent pools. If you do not ensure that the ranks are balanced, uneven device adapter loading can lead to a considerable performance degradation.

The DS8000 series controllers assign server (controller) affinity to ranks when they are added to an extent pool. Ranks that belong to an even-numbered extent pool have an affinity to server0, and ranks that belong to an odd-numbered extent pool have an affinity to server1.

Figure 2-1 shows an example of a configuration that results in a 50% reduction in available bandwidth. Notice how arrays on each of the DA pairs are accessed only by one of the adapters. In this case, all ranks on DA pair 0 are added to even-numbered extent pools, which means that they all have an affinity to server0. Therefore, the adapter in server1 is sitting idle. Because this condition is true for all four DA pairs, only half of the adapters are actively performing work. This condition can also occur on a subset of the configured DA pairs.

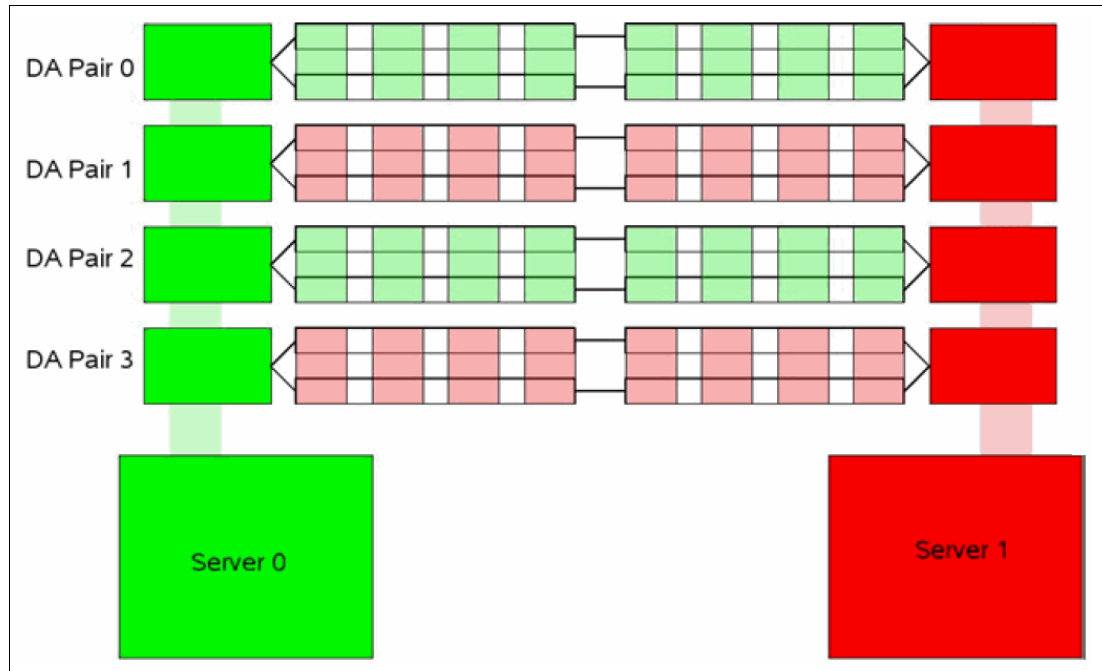


Figure 2-1 DA pair reduced bandwidth configuration

Example 2-3 shows what this invalid configuration looks like from the CLI output of the **lsarray** and **lsrank** commands. The arrays that are on the same DA pair contain the same group number (0 or 1), meaning that they have affinity to the same DS8000 series server. Here, server0 is represented by group0, and server1 is represented by group1.

As an example of this situation, consider arrays A0 and A4, which are attached to DA pair 0. In this example, both arrays are added to an even-numbered extent pool (P0 and P4) so that both ranks have affinity to server0 (represented by group0), which leaves the DA in server1 idle.

Example 2-3 Command output for the **lsarray** and **lsrank** commands

```

dscli> lsarray -l
Date/Time: Oct 20, 2016 12:20:23 AM CEST IBM DSCLI Version: 7.8.1.62 DS: IBM.2107-75L2321
Array State Data RAID type arsite Rank DA Pair DDMcap(10^9B) diskclass
-----
A0 Assign Normal 5 (6P+S) S1 R0 0 146.0 ENT
A1 Assign Normal 5 (6P+S) S9 R1 1 146.0 ENT
A2 Assign Normal 5 (6P+S) S17 R2 2 146.0 ENT
A3 Assign Normal 5 (6P+S) S25 R3 3 146.0 ENT
A4 Assign Normal 5 (6P+S) S2 R4 0 146.0 ENT
A5 Assign Normal 5 (6P+S) S10 R5 1 146.0 ENT
A6 Assign Normal 5 (6P+S) S18 R6 2 146.0 ENT
A7 Assign Normal 5 (6P+S) S26 R7 3 146.0 ENT

```

```

dscli> lsrank -l
Date/Time: Oct 20, 2016 12:22:05 AM CEST IBM DSCLI Version: 7.8.1.62 DS: IBM.2107-75L2321
ID  Group State  datastate Array RAIDtype  extpoolID  extpoolnam  stgtype  exts  usedexts
-----
R0   0 Normal  Normal   A0      5   P0      extpool0   fb    779    779
R1   1 Normal  Normal   A1      5   P1      extpool1   fb    779    779
R2   0 Normal  Normal   A2      5   P2      extpool2   fb    779    779
R3   1 Normal  Normal   A3      5   P3      extpool3   fb    779    779
R4   0 Normal  Normal   A4      5   P4      extpool4   fb    779    779
R5   1 Normal  Normal   A5      5   P5      extpool5   fb    779    779
R6   0 Normal  Normal   A6      5   P6      extpool6   fb    779    779
R7   1 Normal  Normal   A7      5   P7      extpool7   fb    779    779

```

Figure 2-2 shows a configuration that balances the workload across all four DA pairs.

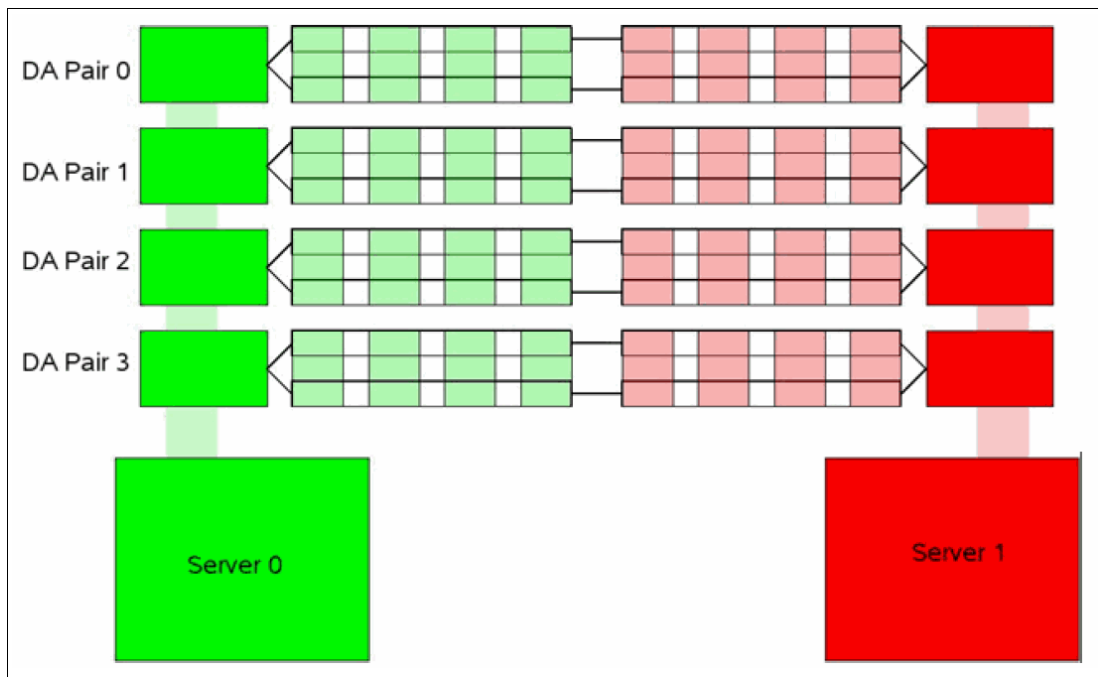


Figure 2-2 DA pair correct configuration

Figure 2-3 shows what a correct configuration looks like from the CLI output of the `lsarray` and `lsrank` commands. Notice that the output shows that this configuration balances the workload across all four DA pairs with an even balance between odd and even extent pools. The arrays that are on the same DA pair are split between groups 0 and 1.

```

dscli> lsarray -l
Date/Time: Oct 20, 2016 10:15:43 AM CEST IBM DSCLI Version: 7.8.1.62 DS: IBM.2107-75L2321
Array State Data RAID type arsite Rank DA Pair DDMcap(10^9B) diskclass
-----
A0 Assign Normal 5 (6+P+S) S1 R0 0 1200.0 ENT
A1 Assign Normal 5 (6+P+S) S2 R1 1 1200.0 ENT
A2 Assign Normal 5 (6+P+S) S3 R2 2 1200.0 ENT
A3 Assign Normal 5 (6+P+S) S4 R3 3 1200.0 ENT
A4 Assign Normal 5 (6+P+S) S5 R4 0 1200.0 ENT
A5 Assign Normal 5 (6+P+S) S6 R5 1 1200.0 ENT
A6 Assign Normal 5 (6+P+S) S7 R6 2 1200.0 ENT
A7 Assign Normal 5 (6+P+S) S8 R7 3 1200.0 ENT

dscli> lsrank -l
Date/Time: Oct 20, 2016 10:20:23 AM CEST IBM DSCLI Version: 7.8.1.62 DS: IBM.2107-75L2321
ID Group State datastate Array RAIDtype extpoolID extpoolnam stgtype exts usedexts encryptgrp marray
-----
R0 0 Normal Normal A0 5 P0 extpool0 fb 6348 6348 - MA1
R1 1 Normal Normal A1 5 P1 extpool1 fb 6348 6348 - MA2
R2 0 Normal Normal A2 5 P2 extpool2 fb 6348 6348 - MA3
R3 1 Normal Normal A3 5 P3 extpool3 fb 6348 6348 - MA4
R4 1 Normal Normal A4 5 P5 extpool5 fb 6348 6348 - MA5
R5 0 Normal Normal A5 5 P4 extpool4 fb 6348 6348 - MA6
R6 1 Normal Normal A6 5 P7 extpool7 fb 6348 6348 - MA7
R7 0 Normal Normal A7 5 P6 extpool6 fb 6348 6348 - MA8

```

Figure 2-3 The `lsarray` and `lsrank` command output

DS8000 series ranks to extent pools mapping

In the DS8000 architecture, extent pools are used to manage one or more ranks. An extent pool is visible to both processor complexes in the DS8000 storage system, but it is directly managed by only one of them. You must define a minimum of two extent pools with one extent pool that is created for each processor complex to fully use the resources. You can use the following approaches:

- **Classical approach:** One array per extent pool configuration.

For IBM Spectrum Virtualize or Storwize attachments, some clients formatted the DS8000 arrays in 1:1 assignments between arrays and extent pools. This configuration disabled any DS8000 storage pool striping or auto-rebalancing activity. Then, you can create one or two volumes in each extent pool exclusively on one rank only, and put all of those volumes into one IBM Spectrum Virtualize or Storwize storage pool.

IBM Spectrum Virtualize or Storwize controlled striping across all of these volumes, and balanced the load across the RAID ranks by that method. No more than two volumes per rank are needed with this approach. So, the rank size determines the volume size.

For example, if the rank is 3682 GiB, make two volumes of 1841 GiB each, and eventually put them in different storage pools to avoid double striping across one rank.

Often, clients worked with at least two storage pools: One (or two) containing MDisk of all the 6+P RAID 5 ranks of the DS8000 storage system, and the other one (or more) containing the slightly larger 7+P RAID 5 ranks. This approach maintains equal load balancing across all ranks when the IBM Spectrum Virtualize or Storwize striping occurs because each MDisk in a storage pool is the same size.

The IBM Spectrum Virtualize or Storwize extent size is the stripe size that is used to stripe across all these single-rank MDisk.

This approach delivered good performance and has its justifications. However, it also has a few minor drawbacks. There can be natural skew, such as a small file of a few hundred KiB that is heavily accessed. Even with a smaller IBM Spectrum Virtualize or Storwize extent size, such as 256 MiB, this classical setup led in a few cases to ranks that are more loaded than other ranks.

When you have more than two volumes from one rank, but not as many IBM Spectrum Virtualize or Storwize storage pools, IBM Spectrum Virtualize or Storwize might start striping across many entities that are effectively in the same rank, depending on the storage pool layout. Such striping should be avoided.

Clients tend to, in DS8000 installations, go to larger (multi-rank) extent pools to use modern features, such as auto-rebalancing or advanced tiering. An advantage of this classical approach is that it delivers more options for fault isolation and control over where a certain volume and extent are located.

► **Modern approach:** Multi-rank extent pool configuration

A more modern approach is to create a few DS8000 extent pools, for example, two DS8000 extent pools. Use either DS8000 storage pool striping or automated Easy Tier rebalancing to help prevent from overloading individual ranks.

Create at least two extent pools for each tier to balance the extent pools by Tier and Controller affinity. Mixing different tiers on the same extent pool is only effective when Easy Tier is activated on the DS8000 pools. However, when virtualized, tier management has more advantages when handled by the SAN Volume Controller.

You need only one volume size with this multi-rank approach because plenty of space is available in each large DS8000 extent pool. As mentioned previously, the maximum number of back-end storage ports to be presented to the SAN Volume Controller is 16. Each port represents a path to the SAN Volume Controller.

Therefore, when sizing the number of LUN/MDisks to be presented to the SAN Volume Controller, the suggestion is to present, at least, 2 - 4 volumes per path. So using the maximum of 16 paths, create 32, 48, or 64 DS8000 volumes, and for this configuration it maintains a good queue depth.

To maintain the highest flexibility and for easier management, large DS8000 extent pools are beneficial. However, if the IBM Spectrum Virtualize or Storwize DS8000 installation is dedicated to shared-nothing environments, such as Oracle ASM, IBM DB2® warehouses, or General Parallel File System (GPFS), use the single-rank extent pools.

Preferred practice: Configure four to eight ranks per extent pool.

For more information and preferred practices for Easy Tier, see Chapter 7, “IBM Easy Tier function” on page 273.

LUN masking

For a storage controller, all SAN Volume Controller nodes or Storwize canisters must detect the same set of LUNs from all target ports that logged in to the SAN Volume Controller nodes. If target ports are visible to the nodes or canisters that do not have the same set of LUNs assigned, SAN Volume Controller treats this situation as an error condition and generates error code 1625.

You must validate the LUN masking from the storage controller and then confirm the correct path count from within the SAN Volume Controller.

The DS8000 series controllers perform LUN masking that is based on the volume group. Example 2-4 shows the output of the `showvolgrp` command for volume group (V0), which contains 16 LUNs that are being presented to a two-node SAN Volume Controller cluster.

Example 2-4 Output of the showvolgrp command

```
dsccli> showvolgrp V0
Date/Time: Oct 20, 2016 10:33:23 AM BRT IBM DSCCLI Version: 7.8.1.62 DS: IBM.2107-75FPX81
Name ITS0_SVC
ID V0
Type SCSI Mask
Vols 1001 1002 1003 1004 1005 1006 1007 1008 1101 1102 1103 1104 1105 1106 1107 1108
```

Example 2-5 shows output for the `lshostconnect` command from the DS8000 series. In this example, you can see that four ports of the two-node cluster are assigned to the same volume group (V0) and, therefore, are assigned to the same four LUNs.

Example 2-5 Output for the lshostconnect command

```
dsccli> lshostconnect -volgrp v0
Date/Time: Oct 22, 2016 10:45:23 AM BRT IBM DSCCLI Version: 7.8.1.62 DS: IBM.2107-75FPX81
Name ID WWPN HostType Profile portgrp volgrpID ESSIOport
=====
ITS0_SVC_N1C1P4 0001 500507680C145232 SVC San Volume Controller 1 V0 all
ITS0_SVC_N1C2P3 0002 500507680C235232 SVC San Volume Controller 1 V0 all
ITS0_SVC_N2C1P4 0003 500507680C145231 SVC San Volume Controller 1 V0 all
ITS0_SVC_N2C2P3 0004 500507680C235231 SVC San Volume Controller 1 V0 all
dsccli>
```

From Example 2-5 you can see that only the SAN Volume Controller WWPNS are assigned to V0.

Attention: Data corruption can occur if the same LUNs are assigned to SAN Volume Controller nodes and non-SAN Volume Controller nodes, which are direct-attached hosts.

Next, you see how the SAN Volume Controller or Storwize detects these LUNs if the zoning is properly configured. The Managed Disk Link Count (`mdisk_link_count`) represents the total number of MDisks that are presented to the SAN Volume Controller cluster by that specific controller.

Example 2-6 shows the general details of the output storage controller by using the SAN Volume Controller command-line interface (CLI).

Example 2-6 Output of the lscontroller command

```
IBM_2145:ITS0_SVC:admin>svcinfo lscontroller DS8K75FPX81
id 1
controller_name DS8K75FPX81
WWNN 5005076305FFC74C
mdisk_link_count 16
max_mdisk_link_count 16
degraded no
vendor_id IBM
product_id_low 2107900
product_id_high
product_revision 3.44
ctrl_s/n 75FPX81FFFF
allow_quorum yes
```

```
fabric_type fc
site_id
site_name
WWPN 500507630500C74C
path_count 16
max_path_count 16
WWPN 500507630508C74C
path_count 16
max_path_count 16
iscsi_port_id
IBM_2145:ITS0_SVC:admin>
```

2.3 Considerations for IBM XIV Storage System

With IBM Spectrum Virtualize and Storwize V7.7.1.1 and later, Gen3 XIV volumes can be provisioned to SAN Volume Controller/Storwize via iSCSI. However, always implement FC over iSCSI for performance and stability considerations, unless a dedicated IP infrastructure for storage is available.

2.3.1 Connectivity considerations

The preferred practices related to cabling and zoning are detailed in Chapter 1, “Storage area network” on page 1.

2.3.2 Host options and settings for XIV systems

You must use specific settings to identify SAN Volume Controller/Storwize systems as hosts to XIV systems. An XIV node within an XIV system is a single WWPN. An XIV node is considered to be a single SCSI target. Each SAN Volume Controller/Storwize host object that is created within the XIV System must be associated with the same LUN map.

From a SAN Volume Controller/Storwize perspective, an XIV Type Number 281x controller can consist of more than one WWPN. However, all are placed under one worldwide node number (WWNN) that identifies the entire XIV system.

Creating a host object for SAN Volume Controller/Storwize for an IBM XIV

A single host object with all WWPNs for the SAN Volume Controller nodes or Storwize canisters can be created when implementing IBM XIV with SAN Volume Controller/Storwize. This technique makes the host configuration easier to configure. However, the ideal host definition for use with SAN Volume Controller/Storwize is to consider each node or canister of the SAN Volume Controller/Storwize as a host object, and create a cluster object to include all nodes or canisters.

By implementing the SAN Volume Controller/Storwize in this manner, host management is ultimately simplified. Also, statistical metrics are more effective because performance can be determined at the node level instead of the SAN Volume Controller/Storwize cluster level.

Consider an example where the SAN Volume Controller/Storwize is successfully configured with the XIV system. If an evaluation of the volume management at the I/O group level is needed to ensure efficient utilization among the nodes, you can compare the nodes by using the XIV statistics.

A detailed procedure to create a host on XIV is described in *IBM XIV Gen3 with IBM System Storage SAN Volume Controller and Storwize V7000*, REDP-5063.

2.3.3 Volume considerations

As modular storage, XIV storage can be presented from six modules and up to 15 modules in a configuration. Each additional module added to the configuration increases the XIV capacity, CPU, memory, and connectivity. The XIV system currently supports the following configurations:

- ▶ 28 - 81 TB when using 1 TB drives
- ▶ 55 - 161 TB when using 2 TB disks
- ▶ 84 - 243 TB when using 3 TB disks
- ▶ 112 - 325 TB when using 4 TB disks
- ▶ 169 - 489 TB when using 6 TB disks

Figure 2-4 details how XIV configuration varies according to the number of modules present on the system.

Rack Configuration								
Total number of modules (Configuration type)	6 partial	9 partial	10 partial	11 partial	12 partial	13 partial	14 partial	15 full
Total number of data modules	3	3	4	5	6	7	8	9
Total number of interface modules	3	6	6	6	6	6	6	6
Number of <i>active</i> interface modules	2	4	4	5	5	6	6	6
Interface module 9 state		Disabled	Disabled	Enabled	Enabled	Enabled	Enabled	Enabled
Interface module 8 state		Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Interface module 7 state		Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Interface module 6 state	Disabled	Disabled	Disabled	Disabled	Disabled	Enabled	Enabled	Enabled
Interface module 5 state	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
Interface module 4 state	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled	Enabled
FC ports	8	16	16	20	20	24	24	24
iSCSI ports (1 Gbps – mod 114)	6	14	14	18	18	22	22	22
iSCSI ports (10 Gbps – mod 214)	4	8	8	10	10	12	12	12
Number of disks	72	108	120	132	144	156	168	180
Usable capacity (1 / 2 / 3 / 4 / 6 TB)	28 TB	44 TB	51 TB	56 TB	63 TB	67 TB	75 TB	81 TB
	55 TB	88 TB	102 TB	111 TB	125 TB	134 TB	149 TB	161 TB
	84 TB	132 TB	154 TB	168 TB	190 TB	203 TB	225 TB	243 TB
	112 TB	177 TB	207 TB	225 TB	254 TB	272 TB	301 TB	325 TB
	169 TB	267 TB	311 TB	338 TB	382 TB	409 TB	453 TB	489 TB
# of CPUs (one per Module)	6	9	10	11	12	13	14	15
Memory (24 GB per module w 1/2/3 TB)	144 GB	216 GB	240 GB	264 GB	288 GB	312 GB	336 GB	360 GB
Memory (48 GB per module w 4/6 TB)	288 GB	432 GB	480 GB	528 GB	576 GB	624 GB	672 GB	720 GB
{Optional for 1, 2, 3, 4, 6 TB XIVs} 400 GB Flash Cache	2.4 TB	3.6 TB	4.0 TB	4.4 TB	4.8 TB	5.2 TB	5.6 TB	6.0 TB
{Optional for 4, 6 TB XIVs} 800 GB Flash Cache	4.8 TB	7.2 TB	8.0 TB	8.8 TB	9.2 TB	10.4 TB	11.2 TB	12.0 TB
Power (kVA) - Model 281x-214 / with SSD	2.5 / 2.6	3.6 / 3.9	4.0 / 4.3	4.3 / 4.6	4.7 / 5.09	5.0 / 5.4	5.5 / 5.8	5.8 / 6.2

Figure 2-4 XIV rack configuration: 281x-214

Although XIV has its own queue depth characteristics for direct host attachment, the following practice is preferred when you virtualize XIV with IBM Spectrum Virtualize or Storwize V7000:

- ▶ The maximum queue depth per MDisk is 60
- ▶ The maximum queue depth per target host port on an XIV is 1000

Based on this knowledge, you can determine an ideal number of XIV volumes to map to IBM Spectrum Virtualize or Storwize V7000 for use as MDisks by using the following algorithm:

$$Q = ((P \times C) / N) / M$$

The algorithm has the following components:

- Q** Calculated queue depth for each MDisk
- P** Number of XIV host ports (unique WWPNs) that are visible to the IBM Spectrum Virtualize or Storwize V7000 cluster (use 4, 8, 10, or 12, depending on the number of modules in the XIV)
- N** Number of nodes in the IBM Spectrum Virtualize or Storwize V7000 cluster (2, 4, 6, or 8)
- M** Number of volumes that are presented by the XIV to the IBM Spectrum Virtualize or Storwize V7000 cluster (detected as MDisks)
- C** 1000 (the maximum SCSI queue depth that an IBM Spectrum Virtualize or Storwize V7000 uses for each XIV host port)

If a 2-node IBM Spectrum Virtualize or Storwize V7000 cluster is being used with four ports on an IBM XIV System and 17 MDisks, this configuration yields the following queue depth:

$$Q = ((4 \text{ ports} \times 1000) / 2 \text{ nodes}) / 17 \text{ MDisks} = 117.6$$

Because 117.6 is greater than 60, IBM Spectrum Virtualize or Storwize V7000 uses a queue depth of 60 per MDisk.

If a 4-node IBM Spectrum Virtualize or Storwize V7000 cluster is being used with 12 host ports on the IBM XIV System and 50 MDisks, this configuration yields the following queue depth:

$$Q = ((12 \text{ ports} \times 1000) / 4 \text{ nodes}) / 50 \text{ MDisks} = 60$$

Because 60 is the maximum queue depth, IBM Spectrum Virtualize or Storwize V7000 uses a queue depth of 60 per MDisk. A 4-node IBM Spectrum Virtualize or Storwize V7000 is a good reference configuration for all other node configurations.

Starting with V6.4, IBM Spectrum Virtualize and Storwize support MDisks greater than 2 TB from the XIV system. If you use earlier versions of the IBM Spectrum Virtualize or Storwize code, smaller volume sizes for 2 TB, 3 TB, and 4 TB drives are necessary.

This consideration leads to the suggested volume sizes and quantities for IBM Spectrum Virtualize or a Storwize V7000 system on the XIV with different drive capacities, as shown in Table 2-1.

Table 2-1 XIV minimum volume size and quantity recommendations

Modules	XIV host ports	Volume size (GB) 1 TB drives	Volume size (GB) 2 TB drives	Volume size (GB) 3 TB drives	Volume size (GB) 4 TB drives	Volume size (GB) 6 TB drives	Volume quantity	Ratio of volumes to XIV host ports
6	4	1600	3201	4852	6401	9791	17	4.3
9	8	1600	3201	4852	6401	9791	27	3.4
10	8	1600	3201	4852	6401	9791	31	3.9
11	10	1600	3201	4852	6401	9791	34	3.4
12	10	1600	3201	4852	6401	9791	39	3.9
13	12	1600	3201	4852	6401	9791	41	3.4
14	12	1600	3201	4852	6401	9791	46	3.8
15	12	1600	3201	4852	6401	9791	50	4.2

2.3.4 Additional considerations

This section highlights restrictions for using the XIV system as back-end storage for the SAN Volume Controller/Storwize.

Volume mapping

When mapping a volume to the SAN Volume Controller/Storwize, you must use the same LUN ID to all SAN Volume Controller nodes or Storwize canisters. Therefore, map the volumes to the cluster, not to individual nodes of the cluster.

XIV Storage pools

When creating an XIV storage pool, define the Snapshot Size as zero (0). There is no need to reserve snapshot space, because it is not recommended to use XIV snapshots on LUNs mapped as IBM Spectrum Virtualize/Storwize MDisks. The snapshot functions should be used on SAN Volume controller/Storwize at the volume level.

Because all LUNs on a single XIV system share performance and capacity characteristics, use a single storage pool for a single XIV system.

Thin Provisioning

XIV thin provisioning pools are not supported by SAN Volume Controller/Storwize. Instead, you must use a regular pool.

Copy functions for XIV models

You cannot use advanced copy functions for XIV models, such as taking a snapshot and remote mirroring, with disks that are managed by the SAN Volume Controller/Storwize.

For more information about configuration of XIV behind IBM Spectrum Virtualize or Storwize, see the following resources:

- ▶ *IBM XIV Gen3 with IBM System Storage SAN Volume Controller and Storwize V7000*, REDP-5063
- ▶ “Can you use Spectrum Virtualize with XIV as storage?” is available at this website:
<https://ibm.biz/Bdr7U3>

2.4 Considerations for IBM FlashSystem A9000/A9000R

IBM FlashSystem A9000 and IBM FlashSystem A9000R use industry-leading data reduction technology that combines inline, real-time pattern matching and removal, data deduplication, and compression. Compression also uses hardware cards inside each grid controller. Compression can easily provide a 2:1 data reduction saving rate on its own, effectively doubling the system storage capacity. Combined with pattern removal and data deduplication services, IBM FlashSystem A9000/A9000R can easily yield an effective data capacity of five times the original usable physical capacity.

Deduplication can be implemented on the IBM Spectrum Virtualize/Storwize by attaching an IBM FlashSystem A9000/A9000R as external storage. While the IBM Spectrum Virtualize/Storwize does not currently provide deduplication natively, by creating a storage pool with managed disks from the IBM FlashSystem A9000/A9000R, deduplication is easily provided.

There are several other considerations when you are attaching an IBM FlashSystem A9000/A9000R system to a SAN Volume Controller or Storwize.

The IBM Spectrum Virtualize or Storwize cluster must be at one of the following code levels. Ensure IBM Spectrum Virtualize or Storwize is upgraded before connecting an IBM FlashSystem A9000/A9000R:

- ▶ Minimum of V7.4.0.10 for systems running V7.4
- ▶ Minimum of V7.5.0.8 for systems running V7.5
- ▶ Minimum of V7.6.1.4 for systems running V7.6
- ▶ For V7.7 and above, there is no specific version requirement

2.4.1 Connectivity considerations

The preferred practices related to cabling and zoning are detailed in Chapter 1, “Storage area network” on page 1.

2.4.2 Volume considerations

IBM FlashSystem A9000/A9000R designates resources to data reduction, and as it is always on, it is strongly advised that data reduction be done only in the IBM FlashSystem A9000/A9000R and not in the SAN Volume Controller. Otherwise, as IBM FlashSystem A9000/A9000R tries to reduce the data, needless additional latency occurs.

Estimated data reduction is important because that helps determine volume size. Always try to use a conservative data reduction ratio when attaching A9000/A9000R with SAN Volume Controller/Storwize systems due to the fact that the storage pool will go offline if the backend storage runs out of capacity.

For example, if the data reduction estimation tool provides a ratio of 4:1, calculate the effective as physical capacity*3.5 and divide it by the number of connected FC ports times 2 (nominal capacity/path*2) to determine the volume size. The remaining usable capacity can be added to the storage pool once the system reaches a stable data reduction ratio.

See Table 2-2 and Table 2-3 for the recommended number of FC ports (paths) that should be used for SAN Volume Controller/Storwize attachment.

Table 2-2 Host connections to SAN Volume Controller for A9000

Number of controllers	Total FC ports available	Total ports that are connected to SAN Volume Controller	Actual ports that are connected
3	12	6	All controllers, ports 1 and 3

Table 2-3 Host connections to SAN Volume Controller for A9000R

Grid Element	Number of controllers	Total FC ports available	Total ports that are connected to SAN Volume Controller	Actual ports that are connected
2	4	16	8	All controllers, ports 1 and 3
3	6	24	12	All controllers, ports 1 and 3
4	8	32	8	Controllers 1 - 4, port 1 Controllers 5 - 8, port 3
5	10	40	10	Controllers 1 - 5, port 1 Controllers 6 - 10, port 3
6	12	48	12	Controllers 1 - 6, port 1 Controllers 7 - 12, port 3

It is important not to run out of hard capacity on the back-end storage, because that takes the storage pool offline. Close monitoring of the FlashSystem A9000/A9000R is very important. If you start to run out of space, you can use the migration functions of IBM Spectrum Virtualize or Storwize to move data to another storage system. The following cases are two examples:

- ▶ First, a FlashSystem A9000 where we have 57 TB of usable capacity, or 300 TB of effective capacity, at the standard 5.26:1 data efficiency ratio.
We were able to run the data reduction tool on a good representative sample of the volumes that we will be virtualizing. We know that we have a data reduction ratio of 4.2:1. 4.2 x 57 gives you 239.4 TB. Divide this by 12 (six paths x 2), and you get 19.9 TB per volume.
- ▶ A five grid element FlashSystem A9000R, using 29 TB Flash enclosures, has a total usable capacity of 145 TB.
We are using 10 paths and have not run any of the estimation tools on the data. However, we know that the host is not compressing the data. 2.5 x 145 gives 362, and divided by 20 gives 18.1 TB per volume. In this case, if we see that we are getting a much better data reduction ratio than we planned for, we can always create more volumes in the pool and make them available to IBM Spectrum Virtualize or Storwize.

The biggest concern with the number of volumes is ensuring there is adequate queue depth. Given that the maximum volume size on the FlashSystem A9000/A9000R is 1 PB and we are ensuring two volumes per path, we should be able to create a small number of larger volumes and still have good queue depth and not have numerous volumes to manage.

2.4.3 Additional considerations

With IBM Spectrum Virtualize/Storwize, using IBM FlashSystem A9000 deduplication technology is simple. Figure 2-5 shows that the Deduplication attribute of the managed disk is Active.



Figure 2-5 Spectrum Virtualize/Storwize managed disks from IBM FlashSystem A9000R

Deduplication status is important because it also allows IBM Spectrum Virtualize/Storwize to recognize and enforce the following restrictions:

- ▶ Storage pools with deduplication MDisks should only contain MDisks from the same IBM FlashSystem A9000 or IBM FlashSystem A9000R storage controller.
- ▶ Deduplication MDisks cannot be mixed in an Easy Tier enabled storage pool.

Note: Currently IBM Spectrum Virtualize/Storwize does allow you to create compressed volumes in a storage pool that contains deduplication-enabled MDisks. This capability provides no benefit because the IBM FlashSystem A9000 cannot deduplicate or compress data that is already compressed. A good practice is to allow the IBM FlashSystem A9000 to perform the deduplication and compression.

2.3.4, “Additional considerations” on page 62 for IBM XIV Storage Systems apply to IBM FlashSystem A9000/A9000R as well.

2.5 Considerations for IBM Storwize V7000/V5000/V3700

Storwize V7000 provides the same virtualization capabilities as the SAN Volume Controller, and can also use internal disks. Storwize V7000 can also virtualize external storage systems (as the SAN Volume Controller does) and in many cases Storwize V7000 can satisfy performance and capacity requirements. Storwize V7000 is used with the SAN Volume Controller for the following reasons:

- ▶ To consolidate more Storwize V7000 systems into single larger environments for scalability reasons.
- ▶ Where SAN Volume Controller is already virtualizing other storage systems and also there is more capacity available from the Storwize V7000.
- ▶ Before V6.2, remote replication was not possible between the SAN Volume Controller and Storwize V7000. Therefore, if the SAN Volume Controller was used on the primary data center and Storwize V7000 was used for the secondary data center, SAN Volume Controller was required to support replication compatibility.
- ▶ The SAN Volume Controller with current versions (at the time of writing) provides more cache (32 GB up to 256 GB per node versus 32 GB or 64 GB per Storwize V7000 canister). Therefore, adding the SAN Volume Controller can provide more caching capability, which is beneficial for cache-friendly workloads.
- ▶ Virtualizing Storwize V7000 provides better performance because of the additional available cache and more flexibility than using the internal capacity on SAN Volume Controller as it eliminates the need to manually match the volume and I/O group affinity.

Storwize V5000 has the same virtualization features as Storwize V7000. However, its hardware is more restricted in port count and cache than Storwize V7000.

IBM Storwize V3700 only supports external virtualization for the purposes of data migration. Permanently virtualizing external storage for use under the Storwize V3700 is not supported.

Note: Consider not forming multiple I/O group Storwize V7000/V5000 clusters, because it introduces the need of a quorum from external storage or an IP quorum application. Otherwise, a 3124: No active quorum device found error will be generated, and a single I/O group failure might bring down the entire Storwize V7000/V5000 cluster.

2.5.1 Connectivity considerations

If you want to virtualize Storwize behind SAN Volume Controller or another Storwize, connect all FC ports of backend Storwize to the same SAN switches as the SAN Volume Controller or “front-end” Storwize. It is not imperative to dedicate some ports to intranode communication because Storwize node canisters communicate with each other through the internal bus.

Moreover, there is no need to dedicate FC ports to remote copy services because the back-end Storwize system is probably not used for this function. All remote copy services functions should be used from the front-end SAN Volume Controller/Storwize system unless there is a good reason not to.

Note: You can use all functions on the backend Storwize, such as FlashCopy or remote copy, but it adds more complexity and is not recommended.

For additional SAN and zoning preferred practices, see Chapter 1, “Storage area network” on page 1.

2.5.2 Defining internal storage

When you plan to attach a Storwize V7000/V5000/V3700 to the SAN Volume Controller or another Storwize V7000/V5000 system, the first aspect to consider is which RAID level is the most suited and appropriate for your environment. The preferred RAID type varies according to the capacity, performance, and protection level required. For instance, RAID5 can perform faster writes than RAID6, but RAID6 (double parity) provides protection in case of single or double disk failures.

Disk arrays and RAID protection

As disk capacity increases, array rebuilding time also increases. The bigger the disks on an array are, the longer the rebuilding time is. Therefore, the disk array remains unprotected for a longer time (RAID 5 tolerates one disk failure at a time). Figure 2-6 shows the approximate rebuild time for disks.

Traditional RAID Rebuild Times – Estimated						
Drive Class	Drive Capacity in Decimal		Approximate Drive Capacity in Binary	Rebuild Rate* WITH I/O to Array (MiB)	Approximate Drive Capacity in MiB	Approximate Rebuild Time in Hours
NL_SAS 7.2K	8	TB	7.275957614	60	7,629,394.53	35.32
	6		5.456968211	60	5,722,045.90	26.49
	4		3.637978807	45	3,814,697.27	23.55
	3		2.728484105	45	2,861,022.95	17.66
	2		1.818989404	30	1,907,348.63	17.66
	1		0.909494702	30	953,674.32	8.83
SAS 10K	1800	GB	1637.090463	60	1,716,613.77	7.95
	1200		1091.393642	35	1,144,409.18	9.08
	900		838.1903172	35	858,306.88	6.81
SAS 10K/15K	600		558.7935448	35	572,204.59	4.54
SAS 10K/15K	300		279.3967724	35	286,102.29	2.27
SAS 15K	146		135.9730959	35	139,236.45	1.11
SSD	3200	GB	2980.232239	60	3,051,757.81	14.13
SSD	1600		1490.116119	60	1,525,878.91	7.06
SSD	800		745.0580597	60	762,939.45	3.53
SSD	400		372.5290298	60	381,469.73	1.77
SSD	200		186.2645149	60	190,734.86	0.88

* Note that rebuild rate can be significantly faster if no I/O is occurring to the Array/MDisk

Figure 2-6 Traditional RAID rebuild time

Distributed RAID (DRAID)

Distributed RAID was launched in V7.6.0. It enables a RAID5 or RAID6 array to be distributed over a larger set of drives. Previously, if you created a RAID5 array over eight drives, the data was striped across them. In this configuration, each stripe has a data stripe on seven of the drives and a parity strip on the eighth.

With distributed RAID, you specify the stripe width and the number of drives separately. You can still have seven data stripes protected by a parity stripe, but those eight drives are selected from 64. Additionally, DRAID adds distributed sparing.

This is the concept that instead of having a spare set on the side that is not being used, each drive in the array gives up some of its capacity to make a spare. Because there are no idle disks to be used as spares, all disks in a DRAID array contribute to the performance. Distributed RAID solutions can improve the array rebuild time up to 10x faster.

Distributed RAID has a minimum and maximum number of disks in the same array:

- ▶ Minimum drive count in one array/MDisk:
 - DRAID 5: 4 (2+P+S)
 - DRAID 6: 6 (3+P+Q+S)
- ▶ Maximum drive count in a single Distributed RAID for Storwize: 128 disks
- ▶ Maximum drive count in a single Distributed RAID for SAN Volume Controller: 48 disks

The spare options, by default, varies according to the array size:

- ▶ Up to 36 disk drives: One rebuild area
- ▶ 37 - 72 disk drives: Two rebuild areas
- ▶ 73 - 100 disk drives: Three rebuild areas
- ▶ 101 - 128 disk drives: Four rebuild areas

DRAID uses more resources from the storage controller than traditional RAID. Keep in mind that the number of arrays/MDisks per I/O group decreases. Therefore, depending on the storage disk count/capacity, creating small DRAID arrays is not recommended, as you might run out of resources to create arrays/MDisks.

Figure 2-7 shows the DRAID and traditional current array/MDisk limitations (V7.8).

Storage	Arrays per system	
	DRAID	Traditional RAID
San Volume Controller 2145-DH8 & 2145-SV1	32	128
Storwize V7000	32	128
Storwize V5000	20	128
Storwize V3700	10	128*

* Limit depends on the number of drives in the system

Figure 2-7 DRAID and TRAIID arrays per system limitation

Note: The limitations in Figure 2-7 are related to the V8.1 firmware version. More information and limitations for other versions can be found at these websites:

- ▶ SAN Volume Controller:
 - <https://www.ibm.com/support/docview.wss?uid=ssg1S1003658>
- ▶ Storwize V7000:
 - <http://www.ibm.com/support/docview.wss?uid=ssg1S1003741>
- ▶ Storwize V5000:
 - <http://www.ibm.com/support/docview.wss?uid=ssg1S1004971>
- ▶ Storwize V3700:
 - <http://www.ibm.com/support/docview.wss?uid=ssg1S1004388>

Rebuild performance

The main reason for distributed RAID is to improve rebuild performance. When a drive fails, the data from that drive must be rebuilt from the surviving drives and written to a spare. By having a larger set of drives in the array, those rebuild reads come from more drives. Distributed sparing means that the writes are going to a larger set of drives.

Reading from a small set of drives and writing to a single drive is what causes rebuilds to take a long time in traditional RAID, especially if the drive you are writing to is a 4 TB nearline drive. The rebuild time could be up to 24 hours.

Note: DRAID rebuild employs a variable rebuild rate that optimizes application performance while maximizing system health. It does this by adjusting the rebuild rate based on the number of drives that have failed.'

Number of drives per DRAID array

One of the key decisions that you must make is how many drives to put into an array. As you increase the number of drives, it helps to improve the rebuild performance. However, this increase is not linear, and it does not go on forever as you hit other limits in the system. The GUI recommends 40-80, assuming that you have at least 40 drives from the same drive class. Typically, go with what the GUI recommends. However, keep it around 60 to be ideal for spinning disks.

Rebuild areas

A rebuild area is the equivalent capacity to a single drive. Therefore, the more rebuild areas that you have, the more drives that can fail one after another.

The number of rebuild areas you want is a mix of how many drives you have, how important the data is and how quickly you want to replace a failed drive. After a drive has been replaced, the data gets copied back from all the spare spaces to the replaced drive.

The copy back time needs to be taken into account. Replacing a drive does not immediately give you back the redundancy. Generally, go with the default suggested by the GUI, consider adding an extra one if the data is critical.

Important: Use DRAID6 whenever possible.

2.5.3 Volume considerations

After deciding which RAID protection for your environment and creating the MDisk/Arrays on the Storwize storage, create and present volumes to the SAN Volume Controller/Storwize to be virtualized as external storage.

Volumes in Storwize can be created as *striped* or *sequential*. The general rule is to create striped volumes if the storage pool is made up of DRAID arrays, and create one sequential volume per MDisk if the storage pool is made up of traditional arrays. But as usual, this decision depends on a number of factors. The main target is to take the 60 MDisk queue depth into account, which typically applies to HDD MDisks, to aim for eight spindles per MDisk. This configuration, when coupled with the new cache algorithms, works much better (number of drives/eight spindles).

If you have 64 drives on the back-end in a pool, for example, then $64/8 = 8$ volumes are created on the back-end and presented to the SAN Volume Controller or another Storwize as 8 MDisks. This configuration means when you get the 60 MDisk queue, you get roughly a

queue depth of 8 per drive, which keeps the spinning disks moving. It also gives nice concurrency across the ports of the back-end controller.

Storwize systems can have a mixed disk drive type, such as solid-state drives (SSDs), serial-attached SCSI (SAS), and nearline SAS (multitier storage). Therefore, pay attention when you map the Storwize volume to the SAN Volume Controller or another Storwize storage pool (as MDisks). Always assign the volumes with the same array or MDisk characteristics to the SAN Volume Controller or another Storwize storage pool.

In cases of multitier Storwize, consider the advantages of using Easy Tier to manage the tiering. Easy Tier can be enabled at the SAN Volume Controller/Storwize system level and on Storwize when used as back-end storage. However, the use of Easy Tier on back-end storage at the same time is not recommended.

Because SAN Volume Controller/Storwize Easy Tier does not monitor Easy Tier on both back-end/external storage, SAN Volume Controller/Storwize and the backend storage independently rebalance the hot areas according to their own heat map. This process causes a rebalance over a rebalance. Such a situation can eliminate the performance benefits of extent reallocation.

Specific recommendations for Easy Tier on external storages can be found in Chapter 7, “IBM Easy Tier function” on page 273.

Important: Use the same extent size on Storwize V7000 and on the SAN Volume Controller/Storwize. To optimize capacity, use an extent size of 1 GB. Although you can use smaller extent sizes, the 1 GB extent size limits the amount of capacity that can be managed by the SAN Volume Controller cluster. There is no performance benefit gained by using smaller or larger extent sizes.

2.6 Considerations for IBM FlashSystem 900

The main advantage of integrating FlashSystem 900 with SAN Volume Controller is to combine the extreme performance of IBM FlashSystem with the SAN Volume Controller enterprise-class solution such as tiering, volume mirroring, thin provisioning, IBM Real-time Compression, and Copy Services.

When you configure the SAN Volume Controller/Storwize with IBM FlashSystem storage systems, you must remember the considerations that are described in this section.

2.6.1 Connectivity considerations

The physical FC port connections and zoning are described in detail on Chapter 1, “Storage area network” on page 1.

2.6.2 Defining storage

IBM FlashSystem 900 supports up to 12 IBM MicroLatency® modules. Each IBM MicroLatency module has a usable capacity of either 1.06 TiB (1.2 TB), 2.62 TiB (2.9 TB), or 5.24 TiB (5.7 TB) of flash storage. IBM MicroLatency modules without the child board are either half-populated with 1.06 TiB (1.2 TB) or fully populated with 2.62 TiB (2.9 TB). The optional child board adds another 2.62 TiB (2.9 TB) for a total of 5.24 TiB (5.7 TB).

IBM MicroLatency modules are installed in the IBM FlashSystem 900 based on the following configuration guidelines:

- ▶ A minimum of four MicroLatency modules must be installed in the system. RAID 5 is the only supported configuration of the IBM FlashSystem 900.
- ▶ The system supports configurations of 4, 6, 8, 10, and 12 MicroLatency modules in RAID 5.
- ▶ All MicroLatency modules that are installed in the enclosure must be identical in capacity and type.
- ▶ For optimal airflow and cooling, if fewer than 12 MicroLatency modules are installed in the enclosure, populate the module bays beginning in the center of the slots and adding on either side until all 12 slots are populated.

The array configuration is performed during system setup. The system automatically creates MDisk/arrays and defines the RAID settings based on the number of flash modules in the system. The default supported RAID level is RAID 5.

2.6.3 Volume considerations

To fully use all SAN Volume Controller/Storwize resources, create multiples of eight volumes per FlashSystem storage controller. This way, all CPU cores, nodes, and FC ports are fully used. The number of volumes often is not a problem because in real-world scenarios the number of volumes is much higher.

However, one important factor must be considered when volumes are created from a pure FlashSystem MDisks storage pool. FlashSystem can process I/Os much faster than traditional HDDs. In fact, they are even faster than cache operations because with cache, all I/Os to the volume must be mirrored to another node in I/O group.

This operation can take as much as 1 millisecond while I/Os that are issued directly (which means without cache) to the FlashSystem can take 100 - 200 microseconds. So for Flash System backend arrays, consider disabling total cache (both Read and Write) in cases where you are experiencing FlashSystem volume latency issues.

You must keep the cache *enabled* in the following situations:

- ▶ If FlashSystems volumes are compressed
- ▶ If FlashSystems volumes are in a Metro/Global Mirror relationship
- ▶ If FlashSystems volumes are in a FlashCopy relationship (copy on write)
- ▶ If FlashSystems volumes are in an Easy Tier pool

Note: Latency can be reduced and maximum IOPS optimized by turning off cache for external FlashSystems volumes. However, there might be a slight performance impact on compression and the cache-dependent copy services, such as FlashCopy, Volume Mirroring and replication.

Some environments require you to have mirrored volumes that for security reasons must be written in two separate storage systems. When one copy of this mirror comes from FlashSystem MDisks but the other copy comes from the spinning-type of MDisks, you can optimize the performance by prioritizing the Read operations requests to the flash disks.

Writes to mirrored volumes can be processed synchronously or asynchronously to both copies. This configuration depends on the **writemirrorpriority** volume parameter, which can have the value of **latency** (asynchronous) or **redundancy** (synchronous).

Reads are processed only by the primary copy of the mirrored volume, so setting the Flash Disk volume as the primary copy prioritizes the read from the Flash System volumes, as shown in Figure 2-8.

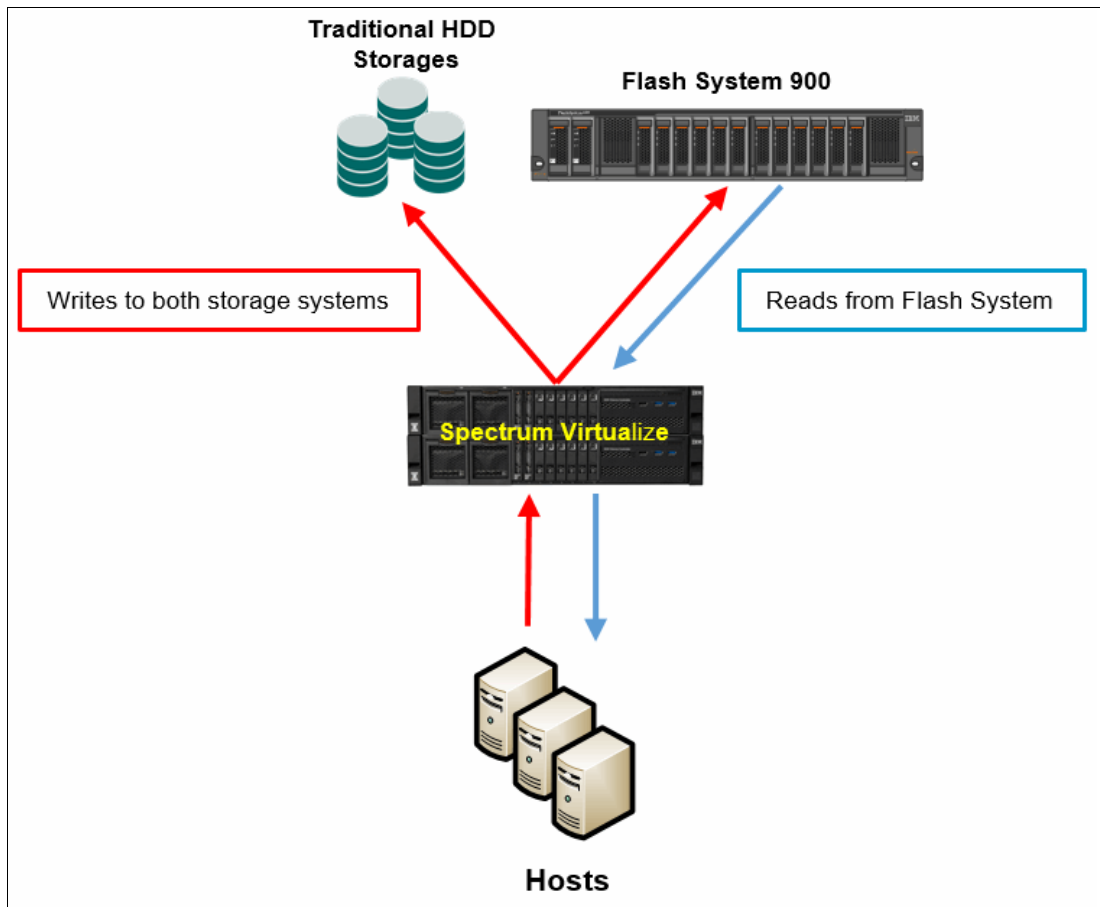


Figure 2-8 Flash System volume as primary copy - read priority.

Although FlashSystem copy might improve your write performance (depending on the **writemirrorpriority** setting of the volume), it can dramatically improve your read performance if you set the primary copy to FlashSystem MDisk copy.

To change a primary copy of a volume, use the command shown in Example 2-7.

Example 2-7 Command to change the primary copy for a mirrored volume

```
chvdisk -primary copy_id_of_mirrored_volume volume_name_or_volume_id
```

To change the mirroring type of a volume copy to synchronous or asynchronous, use the command shown in Example 2-8.

Example 2-8 Command to change the mirrorwritepriority option for a volume

```
chvdisk -mirrorwritepriority latency|redundancy volume_name_or_volume_id
```

Preferred practice: Always change the volume primary copy to the copy that was built out of FlashSystem MDisks and change the **mirrorwritepriority** setting to **latency**.

For more information, see *Implementing IBM FlashSystem 900*, SG24-8271.

2.7 Considerations for storage subsystem compression and deduplication capability

Using an appropriate compression and or data deduplication ratio is key to achieving a stable environment. If you are not sure about the real compression or data deduplication ratio, contact your IBM technical sales representative to obtain more information.

The nominal capacity from a compression and deduplication enabled storage system is not fixed and it varies based on the nature of the data. If we had assumed a compression or data deduplication ratio of ~5:1, that is a little optimistic in this case. Because of the nature of data, we have seen different ratios, such as 3:1. Always try to use a conservative data reduction ratio for the initial configuration.

There are multiple consequences: using the inappropriate ratio for capacity assignment to IBM Spectrum Virtualize or Storwize could cause an out of space situation if the IBM Spectrum Virtualize or Storwize Managed Disks do not provide enough capacity, and IBM Spectrum Virtualize or Storwize disables access to all the volumes in the storage pool. All volumes that are related to the pool go offline, so multiple IBM Spectrum Virtualize or Storwize pools might be affected.

For example:

- ▶ **Assumption 1:** Sizing with ~5:1 rate
- ▶ **Assumption 2:** Real rate is 3:1
 - Physical Capacity: 20 TB
 - Calculated capacity: $20 \text{ TB} \times 5 = 100 \text{ TB}$
 - Volume assigned from compression or deduplication enabled storage subsystem to SAN Volume Controller or Storwize is 100 TB
 - Real usable capacity: $20 \text{ TB} \times 3 = 60 \text{ TB}$

If the hosts try to write more than 60 TB data to the storage pool, the storage subsystem cannot provide any more capacity, and all volumes that are used as IBM Spectrum Virtualize or Storwize Managed Disks and all related pools go offline.

IBM Spectrum Virtualize or Storwize data removal or volume deletion are not reflected on the storage subsystem because the data is IBM Spectrum Virtualize or Storwize internally marked as removed, but the storage subsystem does not have this information. So the storage subsystem still provides capacity for removed data. The garbage collection process cannot free up the space because the area is not marked with zeros and is still in use by IBM Spectrum Virtualize or the Storwize cluster.

Therefore, removed data or removed volumes still count as used capacity in the storage subsystem. The only way to give the capacity back is to overwrite the volume to be deleted or the deleted data with zeros.

Note: A good practice is to allow the deduplication enabled back-end storage to perform the deduplication and compression. Currently IBM Spectrum Virtualize/Storwize does not support automatic deduplication attribute detection for MDisks from non-IBM storage, the restrictions on page 65 have to be enforced manually.

2.8 Considerations for third-party storage with EMC VMAX and Hitachi Data Systems

Although many third-party storage options are available and supported, this section highlights the pathing considerations for EMC VMAX and Hitachi Data Systems (HDS).

Most storage controllers, when presented to the SAN Volume Controller/Storwize, are recognized as a single WWNN per controller. However, for some EMC VMAX and HDS storage controller types, SAN Volume Controller/Storwize recognizes each port as a different WWNN. For this reason, each storage port, when zoned to the SAN Volume Controller/Storwize, appears as a different external storage controller.

SAN Volume Controller/Storwize V8.1 supports a maximum of 16 WWNNs per storage system, so it is preferred to connect up to 16 storage ports to a SAN Volume Controller/Storwize cluster, which results in 16 WWNNs and 16 WWPNS.



Storage pools and managed disks

This chapter highlights considerations when you are planning storage pools for an IBM Spectrum Virtualize and Storwize implementation. It explains various managed disk (MDisk) attributes and provides an overview of the process of adding and removing MDisks from existing storage pools.

This chapter includes the following sections:

- ▶ Availability considerations for storage pools
- ▶ Selecting storage subsystems
- ▶ Selecting the storage pool
- ▶ Quorum disk considerations
- ▶ Tiered storage pool
- ▶ Adding MDisks to existing storage pools
- ▶ Rebalancing extents across a storage pool
- ▶ Removing MDisks from existing storage pools
- ▶ Remapping managed MDisks
- ▶ Controlling extent allocation order for volume creation
- ▶ Considerations when using Encryption

3.1 Availability considerations for storage pools

Although the IBM Spectrum Virtualize and Storwize combination provides many advantages through consolidation of storage, you must understand the availability implications that storage subsystem failures can have on availability domains within the IBM Spectrum Virtualize or Storwize cluster. IBM Spectrum Virtualize and Storwize offers significant performance benefits through its ability to stripe across back-end storage volumes. However, consider the effects that various configurations have on availability.

When you select MDisks for a storage pool, performance is often the primary consideration. However, in many cases, the availability of the configuration is traded for little or no performance gain.

Performance: Increasing the performance potential of a storage pool does not necessarily cause an increase in application performance.

Remember that IBM Spectrum Virtualize and Storwize must take the entire storage pool offline if a single MDisk in that storage pool goes offline. Consider an example where you have 40 arrays of 1 TB each for a total capacity of 40 TB with all 40 arrays in the same storage pool.

In this case, you place the entire 40 TB of capacity at risk if one of the 40 arrays fails (which causes the storage pool to go offline). If you then spread the 40 arrays out over some of the storage pools, the effect of an array failure (an offline MDisk) affects less storage capacity, which limits the failure domain.

An exception exists with IBM XIV Storage System, because this system has unique characteristics.

If the solution you are going to implement must provide business continuity (BC) and high availability (HA) capabilities, all the preferred practices explained later in this chapter are still valid. However, it is strongly suggested that you review the following books for IBM Spectrum Virtualize and Storwize BC and HA solutions:

- ▶ *IBM Spectrum Virtualize and SAN Volume Controller Enhanced Stretched Cluster with VMware*, SG24-8211
- ▶ *IBM Storwize V7000, Spectrum Virtualize, HyperSwap, and VMware Implementation*, SG24-8317

To ensure optimum availability to well-designed storage pools, consider the following preferred practices:

- ▶ It is suggested that each storage pool must contain only MDisks from a single storage subsystem. An exception exists when you are working with IBM System Storage Easy Tier. For more information, see Chapter 7, “IBM Easy Tier function” on page 273.
- ▶ It is suggested that each storage pool contains only MDisks from a single storage tier (SSD or Flash, Enterprise, or NL_SAS). An exception exists when you are working with IBM System Storage Easy Tier. For more information, see Chapter 7, “IBM Easy Tier function” on page 273.
- ▶ Avoid allocating more than 250 TB per each storage pool to minimize the scope of impact in case of a single MDisk failure.

3.2 Selecting storage subsystems

When you are selecting storage subsystems, the decision often comes down to the ability of the storage subsystem to be more reliable and resilient, and meet application requirements. When IBM Spectrum Virtualize and Storwize does not provide any physical level data redundancy for virtualized external storages, the availability characteristics of the storage subsystems' controllers have the most impact on the overall availability of the data that is virtualized by IBM Spectrum Virtualize or Storwize.

When you use MDisk that were created from internal drives, each MDisk is a RAID array so it provides data redundancy according to the RAID level that is selected.

Performance is also a determining factor, where adding IBM Spectrum Virtualize or Storwize as a front-end results in considerable gains. Another factor is the ability of your storage subsystems to be scaled up or scaled out. For example, IBM System Storage DS8000 series is a scale-up architecture that delivers the best performance per unit, and the IBM System Storwize V7000 series can be scaled out with enough units to deliver the same performance.

A significant consideration when you compare native performance characteristics between storage subsystem types is the amount of scaling that is required to meet the performance objectives. Although lower-performing subsystems can typically be scaled to meet performance objectives, the additional hardware that is required lowers the availability characteristics of the IBM Spectrum Virtualize or Storwize cluster.

All storage subsystems possess an inherent failure rate. Therefore, the failure rate of a storage pool becomes the failure rate of the storage subsystem times the number of units.

3.3 Selecting the storage pool

Reducing hardware failure domain for back-end storage is only part of what you must consider. When you are determining the storage pool layout, you must also consider application boundaries and dependencies to identify any availability benefits that one configuration might have over another.

Sometimes, reducing the hardware failure domain, such as placing the volumes of an application into a single storage pool, is not always an advantage from the application perspective. Alternatively, splitting the volumes of an application across multiple storage pools increases the chances of having an application outage if one of the storage pools that is associated with that application goes offline.

The following actions are the starting preferred practices:

- ▶ Create a storage pool for each storage subsystem.
- ▶ Without any specific workload profile, use a 250 TB addressing space per each IOgrp that scales out to a 1 PB IBM Spectrum Virtualize or Storwize cluster with four IOgrps (eight SAN Volume Controller nodes or Storwize canisters).
- ▶ Create a dedicated storage pool if there is a specific performance application request, such as using FlashSystem dedicated for a particular application. One example is an SAP HANA application that means that you create a dedicated pool for this application. In this scenario, keep in mind that if you want to segregate a specific workload at an IBM Spectrum Virtualize or Storwize level that it makes sense only if you segregate the workload at a back-end level.

- ▶ In a Storwize V7000 clustered environment, create storage pools with IOgrp or Control Enclosure affinity. That means you have to use only arrays/MDisks supplied by the internal storage that is directly connected to one IOgrp SAS chain only. This configuration avoids unnecessary IOgrp-to-IOgrp communication traversing the SAN and consuming Fibre Channel bandwidth.
- ▶ In a SAN Volume Controller environment, when using a 12F, 24F, or 92F expansion enclosure, create storage pools with IOgrp and Expansion Enclosure affinity. That means you must create arrays/MDisk supplied by the internal storage that is directly connected to one IOgrp SAS chain only. This configuration avoids unnecessary IOgrp to IOgrp communication traversing the SAN and consuming Fibre Channel bandwidth.
- ▶ Try to limit the number of storage pools to seven or fewer when using 2145-DH8 or 2145-SV1 nodes. With the introduction of IBM Spectrum Virtualize or Storwize V7.3, IBM introduced a new two-level cache architecture. With the new cache architecture, potential performance problems caused by excessive cache partitioning have been mitigated.
- ▶ For storage pool extent size, for most clusters a 1 - 2 PB capacity is sufficient. In general, use 256 MB to address a 1 PB space, and for larger clusters use 512 MB as the standard extent size to address up to a 2 PB space. Alternatively, when you are working with the XIV system or DS8000 family system, use an extent size of 1 GB. Keep in mind that the smaller the extent size, the better the usage of a Tier0 disk like SSD or Flash is. This is because Easy Tier has a better chance to move around a larger number of hot extents simultaneously. In most cases, this configuration gives you better performance.
- ▶ Keep the same extent size for all pools. Volumes cannot be migrated between pools with different extent sizes.
- ▶ Consider implementing child pools when you need to have a logical division of your Volumes for each application set. There are often cases where you want to subdivide a storage pool but maintain a larger number of MDisks in that pool. Child pools are logically similar to storage pools, but allow you to specify one or more subdivided child pools. Thresholds and throttles can be set independently per child pool.

3.3.1 Capacity planning consideration

When you configure storage pools, consider leaving a small amount of MDisk capacity that can be used as *swing* (spare) capacity for image mode volume migrations. Generally, allow enough space that is equal to the capacity of your biggest configured volumes.

For those Easy Tier enabled storage pools that contain multiple tiers of storages, always allow some free capacity for Easy Tier to deliver better performance.

3.3.2 Selecting the number of arrays per storage pool

The capability to stripe across disk arrays is one of the most important performance advantages of IBM Spectrum Virtualize and Storwize. However, striping across more arrays is not necessarily better. The objective here is to add only as many arrays to a single storage pool as required to meet the performance objectives.

Because the number of arrays that are required in terms of performance must be defined in the pre-sales or solution design phase, when sizing the environment keep in mind that adding too many arrays to a single storage pool increases the failure domain (3.1, “Availability considerations for storage pools” on page 76). However, reducing the arrays’ number by using bigger disks (such as a 2 TB - 8 TB NL_SAS disk or a 1.2 TB - 1.8 TB disk) might affect performance because the bigger the disk, the bigger the I/O density. It is important to find the tradeoff between the performance, availability, and scalability cost of the solution.

Consider the effect of aggregate workload across multiple storage pools. Striping workload across multiple arrays has a positive effect on performance when you are dealing with dedicated resources. However, the performance gains diminish as the aggregate load increases across all available arrays. For example, if you have a total of eight arrays and are striping across all eight arrays, performance is much better than if you were striping across only four arrays.

However, consider a situation where the eight arrays are divided into two LUNs each and are included in another storage pool. In this case, the performance advantage drops as the load of storage pool 2 approaches the load of storage pool 1. When the workload is spread evenly across all storage pools, no difference in performance occurs.

3.3.3 Selecting LUN attributes

Configure LUNs to use the entire array, particularly for midrange storage subsystems where multiple LUNs that are configured to an array result in a significant performance degradation. The performance degradation is attributed mainly to smaller cache sizes and the inefficient use of available cache. This situation defeats the subsystem's ability to perform *full stripe writes* for RAID 5 arrays. Also, I/O queues for multiple LUNs directed at the same array can overdrive the array.

High-end storage controllers, such as the DS8000 series, make this situation much less of an issue by using large cache sizes. In addition, on high-end storage controllers, most workloads show the difference between a single LUN per array that is compared to multiple LUNs per array to be negligible. In version 7.x, the maximum supported MDisk size is 1 PB, so the maximum LUN size on the storage controller side is no longer an issue.

In cases where you have more than one LUN per array, include the LUNs in the same storage pool.

The selection of LUN attributes for storage pools requires the following primary considerations:

- ▶ Selecting an array size
- ▶ Selecting a LUN size
- ▶ Number of LUNs per array
- ▶ Number of physical disks per array

Important: Create one LUN per array so that you can use the entire capacity of the array.

All LUNs (MDisks) for a storage pool creation must have the same performance characteristics. If MDisks of varying performance levels are placed in the same storage pool, the performance of the storage pool can be reduced to the level of the poorest performing MDisk. Likewise, all LUNs must also possess the same availability characteristics.

If you are going to implement manual tiering over multiple storage pools, you need to include LUNs with the same performance characteristics on each pool. Varying the performance level means that you are not consistent with the type of data that you want to store on it (for example Gold, Silver, or Bronze data), or with different classifications (such as T1, T2, or T3).

Remember that IBM Spectrum Virtualize does not provide any RAID capabilities for virtualized external storage within a storage pool. The loss of access to any one of the MDisks within the storage pool affects the entire storage pool. However, with volume mirroring you can protect against the loss of a storage pool by mirroring a volume across multiple storage pools. For more information, see Chapter 4, "Volumes" on page 109.

For LUN selection within a storage pool, ensure that the LUNs have the following configuration:

- ▶ Same type
- ▶ Same RAID level
- ▶ Same RAID width (number of physical disks in array)
- ▶ Same availability and fault tolerance characteristics

This is not a technical limitation, but it is a preferred practice to optimize the performance, availability, and cost of the infrastructure.

You must place the MDisks that are created on LUNs with varying performance and availability characteristics in separate storage pools.

3.4 Quorum disk considerations

When back-end storage is initially added to an IBM Spectrum Virtualize or Storwize cluster as a storage pool, three quorum disks are automatically created by allocating space from the assigned MDisks. Only one of those disks is selected as the active quorum disk. As more back-end storage controllers (and therefore storage pools) are added to the IBM Spectrum Virtualize or Storwize cluster, the quorum disks are not reallocated to span multiple back-end storage subsystems.

For Storwize, the quorum by default is placed on the internal drives, not on the MDisks. You can change placement of all three quorums to external MDisks, or you can have some quorums on internal drives and some on the external MDisks. You should have quorums that are spread among storage controllers (for example, the active quorum on an internal drive) and the other two quorums on MDisks from the external storage systems.

To eliminate a situation where all quorum disks go offline because of a back-end storage subsystem failure, allocate quorum disks on multiple back-end storage subsystems. This design is possible only when multiple back-end storage subsystems (and therefore multiple storage pools) are available.

Even when only a single storage subsystem is available but multiple storage pools are created from it, the quorum disk must be allocated from several storage pools. This allocation avoids an array failure that causes a loss of the quorum. Reallocating quorum disks can be done from the GUI or from the CLI.

To list IBM Spectrum Virtualize or Storwize cluster quorum MDisks, and to view their number and status, run the **lsquorum** command as shown in Example 3-1.

Example 3-1 The lsquorum command

```
IBM_2145:ITS0_SVC_SPLIT:superuser>lsquorum
quorum_index status id name          controller_id controller_name
active object_type override
0             online 10 ITS0_V7K_SITEC_Q    5             ITS0_V7K_SITEC_Q_N2 yes
mdisk         yes
1             online 6  ITS0_V7K_SITEB_SAS0 0             ITS0_V7K_SITEB_N2  no
mdisk         yes
2             online 4  ITS0_V7K_SITEA_SAS3 1             ITS0_V7K_SITEA_N2  no
mdisk         yes
```

To move one quorum MDisk from one MDisk to another, or from one storage subsystem to another, use the **chquorum** command.

The cluster uses the quorum disk for the following purposes:

- ▶ As a tie breaker if a SAN fault occurs when exactly half of the nodes that were previously members of the cluster are present
- ▶ To hold a copy of important cluster configuration data

Only one active quorum disk is in a cluster. However, the cluster uses three MDisks as quorum disk candidates. The cluster automatically selects the actual active quorum disk from the pool of assigned quorum disk candidates.

If a tiebreaker condition occurs, the half of the cluster nodes that can reserve the quorum disk after the split occurs locks the disk and continues to operate. The other half stops its operation. This design prevents both sides from becoming inconsistent with each other.

Criteria for quorum disk eligibility: To be considered eligible as a quorum disk, the MDisk must meet the following criteria:

- ▶ An MDisk must be presented by a disk subsystem that is supported to provide IBM Spectrum Virtualize or Storwize quorum disks.
- ▶ To manually allow the controller to be a quorum disk candidate, you must enter the following command:

```
svctask chcontroller -allowquorum yes
```
- ▶ An MDisk must be in managed mode (no image mode disks).
- ▶ An MDisk must have sufficient free extents to hold the cluster state information and the stored configuration metadata.
- ▶ An MDisk must be visible to all of the nodes in the cluster.
- ▶ MDisk provisioned via iSCSI is not qualified as a quorum candidate.

For more information about special considerations for the placement of the active quorum disk for Stretched Cluster configurations, see *Guidance for Identifying and Changing Managed Disks Assigned as Quorum Disk Candidates*, S1003311, which is available at the following website:

<http://www.ibm.com/support/docview.wss?rs=591&uid=ssg1S1003311>

Attention: Running an IBM Spectrum Virtualize or Storwize cluster without a quorum disk can seriously affect your operation. A lack of available quorum disks for storing metadata prevents any migration operation (including a forced MDisk delete). Mirrored volumes can be taken offline if no quorum disk is available. This behavior occurs because synchronization status for mirrored volumes is recorded on the quorum disk.

During normal operation of the cluster, the nodes communicate with each other. If a node is idle for a few seconds, a heartbeat signal is sent to ensure connectivity with the cluster. If a node fails for any reason, the workload that is intended for it is taken over by another node until the failed node is restarted and admitted again to the cluster. This process happens automatically.

If the microcode on a node becomes corrupted (which results in a failure), the workload is transferred to another node. The code on the failed node is repaired and the node is admitted again to the cluster (all automatically).

The number of extents that are required depends on the extent size for the storage pool that contains the MDisk. Table 3-1 provides the number of extents that are reserved for quorum use by extent size.

Table 3-1 Number of extents that are reserved by extent size

Extent size (MB)	Number of extents that are reserved for quorum use
16	17
32	9
64	5
128	3
256	2
512	1
1024	1
2048	1
4096	1
8192	1

To discover what MDisks are acting as quorum in your environment, create three 1-GB LUNs, one on each back-end storage subsystem if possible. Then put these MDisks in their own dedicated storage pool. For Storwize V7000, quorum disks are HDD drives by default, so the SAN Volume Controller preferred practice does not apply. However, if your Storwize V7000 is virtualizing external storage, then the same preferred practice for the SAN Volume Controller does apply.

When these LUNs are in a *managed* state, they are eligible to act as a quorum disk and you can set them using the **chquorum** command or the GUI. In this way, you always know what your quorum disks are, and those three extra storage pools do not affect cache partitioning if volumes are not created in these storage pools.

When implementing an IBM Spectrum Virtualize ESC or IBM Spectrum Virtualize or Storwize HyperSwap solution, some clients like to have a preferred winning site in case of a split brain scenario. By using a quorum disk, this configuration is possible. Clients need to implement the active quorum disk in the site that is considered as preferred.

This storage subsystem needs to be defined as to be in Site 3 anyway, connecting to the remote site by using SAN extension connectivity. This is the same connectivity used to extend the Public and Private SAN. In this way, the winning site is always the one with two quorums implemented.

3.4.1 IP Quorum

With respect to quorum disk, IP quorum is a feature that was released in version 7.6.0. IP-based quorum support can enable the use of a low-cost IP network-attached host as a quorum disk for simplified implementation and operation.

SVC currently uses storage from MDisks or drives for quorum resolution. As stated earlier, normally three MDisks are either automatically or manually selected to be the quorum disk candidates. The storage system exclusively reserves areas in each of these disks to store quorum data.

During a quorum loss, in a split-brain scenario where half of the cluster cannot see the other half, the quorum disks are used to break the tie. The first half in a split-brain scenario to reach the quorum disks assumes ownership of the cluster and locks the disks. All the nodes in a cluster must have access to the quorum disks.

In standard implementation design, no extra hardware or networking is required beyond what is normally provisioned within a cluster, which is Fibre Channel (FC) or serial-attached SCSI (SAS) attached storage. But in an Enhanced Stretched Cluster or HyperSwap environment, the need for accessibility to the quorum device during a site failure necessitates the presence of a third independent domain for quorum resolution.

Prior to V7.6.0, the third site had to be connected using Fibre Channel, and maintaining this third site and storage controller over FC makes the system costly for site recovery implementation of IBM Spectrum Virtualize or Storwize V7000.

To overcome this limitation of maintaining a third site with FC connectivity along with a site 3 controller, you can implement Ethernet-attached quorum servers at the third site that can be run on hosts. Ethernet connectivity is generally easier and more economical to provide than FC connectivity, and hosts are typically less expensive than fully fledged network-attached storage controllers. This implementation of using a host application over an Ethernet connection can reduce the implementation and maintenance cost.

3.4.2 IP Quorum requirements

Connectivity from servers to the service IP addresses of all nodes has these requirements:

- ▶ Only the first Ethernet port can be used. There is no VLAN support yet.
- ▶ Port 1260 (through SSL/TLS) used for inbound connections from the app to the service IP address of each node.
- ▶ Maximum round-trip time = 80 ms, minimum bandwidth of 2 MBps.
- ▶ As a native OS or in a virtual machine (no need for dedicated server/VM):
 - Red Hat Enterprise Linux 6.5/7; SUSE Linux Enterprise Server 11m3/12; IBM Java 7.1/8.
 - Use the IBM SCORE process for others to see whether support is granted.
- ▶ Application must be able to create files (.LCK, .LOG) in its working directory.
- ▶ Cluster configuration changes (add/remove node, SSL certificate, IP addresses) require you to re-create the Java quorum application package.

IBM Spectrum Virtualize and Storwize V7000 can support up to five IP Quorums at the same time, but only one is active.

The active IP Quorum Application is either the first one started and detected by the IBM Spectrum Virtualize or Storwize cluster, or is the last one active (when there is more than one IP Quorum Application). You cannot make a specific IP Quorum Application the active one. However, by starting/restarting IP Quorum Applications, you might be able to select a specific one.

Generally, have at least two IP Quorum implemented in your environment to be sure that you will always have an IP Quorum active.

In an Enhanced Stretched Cluster or HyperSwap implementation, it is suggested to have three IP Quorums. Preferably, all of them are to be located in sites other than site A and site B.

Note that even with three IP Quorums, it is currently impossible to be certain which will be the winning site during a split brain scenario. This limitation is because FC connectivity interruption between your SAN Volume Controller nodes or Storwize canisters can happen in a rolling fashion or in a different time sequence versus IP connectivity.

Note 1: If you have a stretched cluster configuration with one or many IP Quorum Apps and a quorum disk at site 1 and site 2, and you lose all IP Quorum Apps, you will not have any tie-break protection. If you have no active Quorum device, then the node with the lowest node ID (as shown by `1snode`, and normally the node that was used when the system was first set up) is used to resolve the tie-break.

So, assuming the first node used to create the SVC cluster was `node_id 1`, then that is the node used as the tie breaker.

Note 2: This ID can change because any time that we remove it from the cluster and add it back in, for example, if we do a node upgrade procedure to go from 2145-CG8 to 2145-DH8 or 2145-SV1.

For detailed information about how to implement IP Quorum, see the following site:

https://www.ibm.com/support/knowledgecenter/en/STPVGU_8.1.0/com.ibm.storage.svc.ccsnsole.810.doc/svc_ipquorumconfig.html

When the IP Quorum App is installed in a Linux environment, make sure that it starts automatically at every reboot. The script is supplied to IBMers:

<https://ibm.biz/BdrWGT>

Non-IBMers: The link is available from an IBM internal-only network. Ask your IBM service representative whether they are able to provide you with a copy.

Alternatively, save the script in Example 3-2 as `quorum.sh` on a supported Linux host and change the permission of the script to 755. Run the script using `./quorum.sh &` as root or a user with pseudo privilege.

Example 3-2 Script to start IP Quorum App after a system reboot and restart IP Quorum App if it quits unexpectedly

```
#!/bin/bash

java_home="/usr/bin/java" #change the path if required
quorum_app_location=`find / -name ip_quorum.jar|head -1`
quorum_directory="/opt/IBM/ip_quorum"
quorum_file="/opt/IBM/ip_quorum/ip_quorum.jar"
quorum_script=`find / -name quorum.sh|head -1`
current_directory=`pwd`
if [[ ! -d "$quorum_directory" ]]
then
    mkdir -p /opt/IBM/ip_quorum
fi
if [[ ! -f "$quorum_file" ]]
then
    cp $quorum_app_location $quorum_directory
fi
if [[ ! -f "$quorum_directory/quorum.sh" ]]
then
    cp $quorum_script $quorum_directory
```



```

fi
if [[ $current_directory != $quorum_directory ]]
then
cd $quorum_directory
fi
if [[ -f "/etc/redhat-release" ]]
then
count=`grep '/opt/IBM/ip_quorum/quorum.sh > /dev/null 2>&1 &'
/etc/rc.d/rc.local|wc -l`
chmod +x /etc/rc.d/rc.local
if [[ $count == 0 ]]
then
echo '/opt/IBM/ip_quorum/quorum.sh > /dev/null 2>&1 &' >>
/etc/rc.d/rc.local
fi
fi
if [[ -f "/etc/SuSE-release" ]]
then
if [[ ! -f "/etc/init.d/after.local" ]]
then
echo '/opt/IBM/ip_quorum/quorum.sh > /dev/null 2>&1 &' >>
/etc/init.d/after.local
else
count=`grep '/opt/IBM/ip_quorum/quorum.sh > /dev/null 2>&1 &'
/etc/init.d/after.local|wc -l`
if [[ $count == 0 ]]
then
echo '/opt/IBM/ip_quorum/quorum.sh > /dev/null 2>&1 &' >>
/etc/init.d/after.local
fi
fi
fi
while true
do
quorum_process_count=`ps -ef|grep ip_quorum.jar|grep -v grep|wc -l`
if [[ $quorum_process_count == 0 ]]
then $java_home -jar $quorum_file > /dev/null 2>&1 &
fi
sleep 30
done

```

3.5 Tiered storage pool

IBM Spectrum Virtualize or Storwize makes it easy to configure multiple tiers of storage within the same IBM Spectrum Virtualize or Storwize cluster. You might have single-tiered pools, multitiered storage pools, or both.

In a *single-tiered storage pool*, the MDisks must have the following characteristics, even if they are not a technical limitation, to avoid inducing performance problems and other issues:

- ▶ They have the same hardware characteristics. For example, they need the same RAID type, RAID array size, disk type, and disk revolutions per minute (RPM). This is mostly true about “same disk with same RPMs”. Today you can have *Enterprise* disks with different sizes but the same RPMs, for example 10K RPM 600 GB SAS disk and 900 GB, 1.2 TB, and 1.8 TB.

The suggestion here is to try to keep size and speed consistency in a *single-tiered storage pool*. If this consistency is not possible because of storage pool space, upgrades at different times, and disks with the same size would no longer be available, use disks with sizes closer to the original one. For example, if the pool was configured with a 10K RPM 900 GB SAS disk, a mix with 10K RPM 1.2 TB SAS disk might be accepted.

This configuration would not have a serious side effect on the performance, because Easy Tier has introduced Intra-Tier balance that balances the workload on different MDisks (in this case, with different drive sizes) based on I/O density and response time.

- ▶ The MDisks that are configured must have the same size whenever possible. If this requirement is not feasible, IBM Spectrum Virtualize or Storwize Easy Tier with Intra-Tier balance can balance the workload on different MDisks (in this case with different drive size) based on I/O density and response time.

In a *multitiered storage pool*, you have a mix of MDisks with more than one type of disk tier attribute. For example, a storage pool contains a mix of drive with different technologies:

- ▶ **sas_ssd**

Specifies an SSD (or flash drive) hard disk drive or an external MDisk for the newly discovered or external volume. Starting from version 7.8, the naming convention has changed as follows:

- tier0_flash
- tier1_flash

- ▶ **sas_hdd**

Specifies an enterprise hard disk drive or an external MDisk for the newly discovered or external volume. Starting from version 7.8, the naming convention has changed as follows:

- tier2_hdd

- ▶ **sas_nearline_hdd**

Specifies a nearline hard disk drive or an external MDisk for the newly discovered or external volume. Starting from V7.8 the naming convention has changed as follows:

- tier3_nearline

Figure 3-1 shows changes to Tech Types introduced in V7.8.

Old Name	New Name
sas_ssd	tier0_flash
	tier1_flash
sas_hdd	tier2_hdd
sas_nearline_hdd	tier3_nearline

Figure 3-1 Tech Types

Therefore, a multitiered storage pool contains MDisks with various characteristics, as opposed to a single-tier storage pool. However, each tier must try to follow the same rules applied for the *single-tiered storage pool*. Multi-tiered storage pools are used to enable the automatic migration of extents between disk tiers by using the IBM Spectrum Virtualize or Storwize Easy Tier function. For more information about IBM System Storage Easy Tier, see Chapter 7, “IBM Easy Tier function” on page 273.

It is likely that the MDisks (LUNs) that are presented to the IBM Spectrum Virtualize or Storwize cluster have various performance attributes because of the type of disk or RAID array on which they are installed. The MDisks can be Flash, SDD, 10K/15K RPM SAS disk, or nearline SAS (NL_SAS).

Mdisk with the same HDD size and RPMs can be supplied by different storage controllers, such as Storwize V7000 or the DS8000 family, with different hardware characteristics and different performance, therefore it is not recommended to mix the MDisks from different storage controllers in the same storage pool.

Before IBM Spectrum Virtualize or Storwize V7.8, there were three types of storage tier: *ssd*, *enterprise*, and *nearline*.

Starting with V7.8, you have three tiers encompassing four Tech Types, as shown in Figure 3-2.

Tier	Tech Types
ET_Tier1	tier0_flash
ET_Tier2	tier1_flash, tier2_hdd
ET_Tier3	tier3_nearline

Figure 3-2 Tier and Tech Types

For Storwize and SAN Volume Controller direct-attached storage, when you create an array with tier0_flash drives, the Mdisk becomes ET_Tier1 by default. If you create an array with tier1_flash and tier2_hdd drives, the Mdisk becomes ET_Tier2 by default. If you create an array with tier3_nearline drives, the Mdisk becomes ET_Tier3 by default.

If you present an external controller to IBM Spectrum Virtualize or Storwize, a specific Easy Tier `easytierload` profile is assigned. It can be low, medium, high, or very_high. It specifies the Easy Tier load (amount) to place on a non-array Mdisk within its tier.

If you present an external Mdisk to IBM Spectrum Virtualize or Storwize, it becomes ET_Tier2 by default, even if that external Mdisk was built by using SSD drives or a flash memory system. You must change the Mdisk tier only for MDisks that are presented from external storage systems accordingly with the owning storage controller by using the `chmdisk` command.

When multiple storage tier pools are defined, precautions must be taken to ensure that storage is provisioned from the appropriate tiers. You can ensure that storage is provisioned from the appropriate tiers through storage pool and Mdisk naming conventions, with clearly defined storage requirements for all hosts within the installation.

Naming conventions: When multiple tiers are configured, clearly indicate the storage tier in the naming convention that is used for the storage pools and MDisks.

Effectively, you have four tiers within a 3-Tier Mapping. When you create a volume, the initial capacity is always allocated from ET_Tier2 (tier1_flash and then tier2_hdd) by default. Example 3-3 shows the default MDisk selection for a volume.

Example 3-3 Default MDisk selection

```

IBM_2145:ITSO_DH8_A:superuser>lsmdisk |while read -a b;do echo ${b[0]} ${b[1]}
${b[10]};done
id name tier
0 mdisk0 tier0_flash
1 mdisk1 tier1_flash
2 mdisk2 tier1_flash
3 mdisk3 tier_enterprise
4 mdisk4 tier_nearline
IBM_2145:ITSO_DH8_A:superuser>mkvdisk -iogrp 0 -mdiskgrp 0 -size 50 -unit gb
-name test_vol_08
IBM_2145:ITSO_DH8_A:superuser>lsvdiskextent test_vol_08
id number_extents
1 25
2 25

```

This default behavior can be manipulated by specifying an MDisk list with the **mkvdisk** command to avoid exhausting the capacity from tier1_flash or tier2_hdd. Example 3-4 shows how to manually select MDisks for a volume.

Example 3-4 Manually select mdisks for a volume

```

IBM_2145:ITSO_DH8_A:superuser>lsmdisk |while read -a b;do echo ${b[0]} ${b[1]}
${b[10]};done
id name tier
0 mdisk0 tier0_flash
1 mdisk1 tier1_flash
2 mdisk2 tier1_flash
3 mdisk3 tier_enterprise
4 mdisk4 tier_nearline
IBM_2145:ITSO_DH8_A:superuser>mkvdisk -iogrp 0 -mdiskgrp 0 -mdisk 3 -size 50
-unit gb -name test_vol_09
IBM_2145:ITSO_DH8_A:superuser>lsvdiskextent test_vol_09
id number_extents
3 50

```

Figure 3-3 shows where data lands based on a combination of tier and tech types.

User (VG) Tiers	EasyTier Tier (by configuration)													
	T0	T0+T1	T0+T1 +T2	T0+T1 +T2+T3	T0+T2	T0+T2 +T3	T0+T3	T1	T1+T2	T1+T2 +T3	T1+T3	T2	T2+T3	T3
T0 (Tier0 Flash)	1	1	1	1	1	1	1							
T1 (Tier1 Flash)		2	2	2				2	2	1	2			
T2 (Tier2 HDD)			3	2	2	2			3	2		2	2	
T3 (Tier3 NearLine)				3		3	2			3	3		3	3

Figure 3-3 Tier and tech types combination

3.6 Adding MDisks to existing storage pools

If MDisks are being added to the IBM Spectrum Virtualize or Storwize cluster, you probably do it because you want to provide more capacity. Adding MDisks to storage pools is a simple task, but it is suggested that you perform some checks in advance.

3.6.1 Checking access to new MDisks

Be careful when you add MDisks to existing storage pools to ensure that the availability of the storage pool is not compromised by adding a faulty MDisk. The reason is that loss of access to a single MDisk causes the entire storage pool to go offline.

In IBM Spectrum Virtualize or Storwize, a feature tests an MDisk automatically for reliable read/write access before it is added to a storage pool so that no user action is required. The test fails under the following conditions:

- ▶ One or more nodes cannot access the MDisk through the chosen controller port.
- ▶ I/O to the disk does not complete within a reasonable time.
- ▶ The SCSI inquiry data that is provided for the disk is incorrect or incomplete.
- ▶ The IBM Spectrum Virtualize or Storwize cluster suffers a software error during the MDisk test.

Image-mode MDisks are not tested before they are added to a storage pool because an offline image-mode MDisk does not take the storage pool offline. Therefore, the suggestion here is to use a dedicated storage pool for each Image.mode MDisk. This preferred practice makes it easier to discover what the MDisk is going to be virtualized as and reduce the chance of human error.

3.6.2 Persistent reserve

A common condition where MDisks can be configured by IBM Spectrum Virtualize or Storwize, but cannot perform read/write, is when a persistent reserve is left on a LUN from a previously attached host. Subsystems that are exposed to this condition were previously attached with Subsystem Device Driver (SDD) or Subsystem Device Driver Path Control Module (SDDPCM) because support for persistent reserve comes from these multipath drivers.

In this condition, rezone the back-end storage and map them back to the host that is holding the reserve. Alternatively, map them to another host that can remove the reserve by using a utility, such as **1querypr** (which is included with SDD and SDDPCM) or the Microsoft Windows SDD Persistent Reserve Tool.

3.6.3 Renaming MDisks

After you discover MDisks, rename them from their IBM Spectrum Virtualize or Storwize-assigned name. This will help during problem isolation and avoid confusion that can lead to an administrative error by using a naming convention for MDisks that associates the MDisk with the controller and array.

When multiple tiers of storage are on the same IBM Spectrum Virtualize or Storwize cluster, you might also want to indicate the storage tier in the name. For example, you can use R5 and R10 to differentiate RAID levels, or you can use T1, T2, and so on, to indicate the defined tiers.

Preferred practice: Use a naming convention for MDisks that associates the MDisk with its corresponding controller and array within the controller, such as DS8K_<extent pool name/id>_<volume id>.

3.7 Rebalancing extents across a storage pool

For V7.3 and onwards, IBM Spectrum Virtualize or Storwize Easy Tier has introduced Intra-Tier balancing that balances the extent and workload on different MDisks based on I/O density and response time. Before this feature, adding MDisks to existing storage pools (before IBM Spectrum Virtualize and Storwize) could result in reduced performance across the storage pool because of any extent imbalance that occurred and the potential to create hot spots within the storage pool. After adding MDisks to storage pools, rebalancing extents across all available MDisks was accomplished by using the CLI or alternatively by using a Perl script. This balancing is now automatically taken care of by the code.

3.8 Removing MDisks from existing storage pools

You might want to remove MDisks from a storage pool (for example, when you decommission a storage controller). When you remove MDisks from a storage pool, consider whether to manually migrate extents from the MDisks. It is also necessary to make sure that you remove the correct MDisks.

Sufficient space: The removal occurs only if sufficient space is available to migrate the volume data to other extents on other MDisks that remain in the storage pool. After you remove the MDisk from the storage pool, it takes time to change the mode from managed to unmanaged, depending on the size of the MDisk that you are removing.

When you remove the MDisk made of internal disk drives from the storage pool on Storwize family systems, this MDisk is deleted. This process also deletes the array on which this MDisk was built, and converts all drives that were included in this array to *candidate* state. You can now use those disk drives to create another array of different size and raid type, or you can use them as hot spares.

3.8.1 Migrating extents from the MDisk to be deleted

If an MDisk contains volume extents, you must move these extents to the remaining MDisks in the storage pool. Example 3-5 shows how to list the volumes that have extents on a MDisk by using the CLI.

Example 3-5 Listing of volumes that have extents on an MDisk to be deleted

```
IBM_2145:itsosvcc11:admin>svcinfolsmdiskextent mdisk14
id          number_of_extents  copy_id
5           16                 0
3           16                 0
6           16                 0
8           13                 1
9           23                 0
8           25                 0
```

Specify the **-force** flag on the `svctask rmmdisk` command, or select the corresponding option in the GUI. Both actions cause IBM Spectrum Virtualize or Storwize to automatically move all used extents on the MDisk to the remaining MDisks in the storage pool.

Alternatively, you might want to manually perform the extent migrations. Otherwise, the automatic migration randomly allocates extents to MDisks (and areas of MDisks). After all of the extents are manually migrated, the MDisk removal can proceed without the **-force** flag.

3.8.2 Verifying the identity of an MDisk before removal

MDisks must appear to the IBM Spectrum Virtualize or Storwize cluster as unmanaged before their controller LUN mapping is removed. Unmapping LUNs from IBM Spectrum Virtualize or Storwize that are still part of a storage pool results in the storage pool that goes offline and affects all hosts with mappings to volumes in that storage pool.

If the MDisk was named by using the preferred practices, the correct LUNs are easier to identify. However, ensure that the identification of LUNs that are being unmapped from the controller match the associated MDisk on IBM Spectrum Virtualize or Storwize by using the Controller LUN Number field and the unique identifier (UID) field.

The UID is unique across all MDisks on all controllers. However, the controller LUN is unique only within a specified controller and for a certain host. Therefore, when you use the controller LUN, check that you are managing the correct storage controller and that you are looking at the mappings for the correct IBM Spectrum Virtualize or Storwize host object.

Tip: Renaming your back-end storage controllers as recommended also helps you with MDisk identification.

For more information about how to correlate back-end volumes (LUNs) to MDisks, see 3.8.3, “Correlating the back-end volume with the MDisk” on page 91.

3.8.3 Correlating the back-end volume with the MDisk

The correct correlation between the back-end volume (LUN) with the IBM Spectrum Virtualize or Storwize MDisk is crucial to avoid mistakes and possible outages. You can correlate the back-end volume with MDisk for, DS8000 series, XIV, and V7000 storage controllers.

DS8000 LUN

The LUN ID only uniquely identifies LUNs within the same storage controller. If multiple storage devices are attached to the same IBM Spectrum Virtualize or Storwize cluster, the LUN ID must be combined with the worldwide node name (WWNN) attribute to uniquely identify LUNs within the IBM Spectrum Virtualize or Storwize cluster.

To get the WWNN of the DS8000 controller, take the first 16 digits of the MDisk UID and change the first digit from 6 to 5, such as 6005076305ffc74c to 5005076305ffc74c.

When detected as IBM Spectrum Virtualize or Storwize `ctrl_LUN_#`, the DS8000 LUN is decoded as 40XX40YY00000000, where XX is the logical subsystem (LSS) and YY is the LUN within the LSS. As detected by the DS8000, the LUN ID is the four digits starting from the 29th digit, as in the Example 3-6.

Example 3-6 DS8000 UID example

```
6005076305ffc74c00000000000010070000000000000000000000000000000000000000
```

In Example 3-6 on page 91, you can identify the MDisk supplied by the DS8000, which is LUN ID 1007.

XIV system volumes

Identify the XIV volumes by using the volume serial number and the LUN that is associated with the host mapping. The example in this section uses the following values:

- ▶ Serial number: 897
- ▶ LUN: 2

To identify the volume serial number, right-click a volume and select **Properties**. Figure 3-4 shows the Volume Properties dialog box that opens.

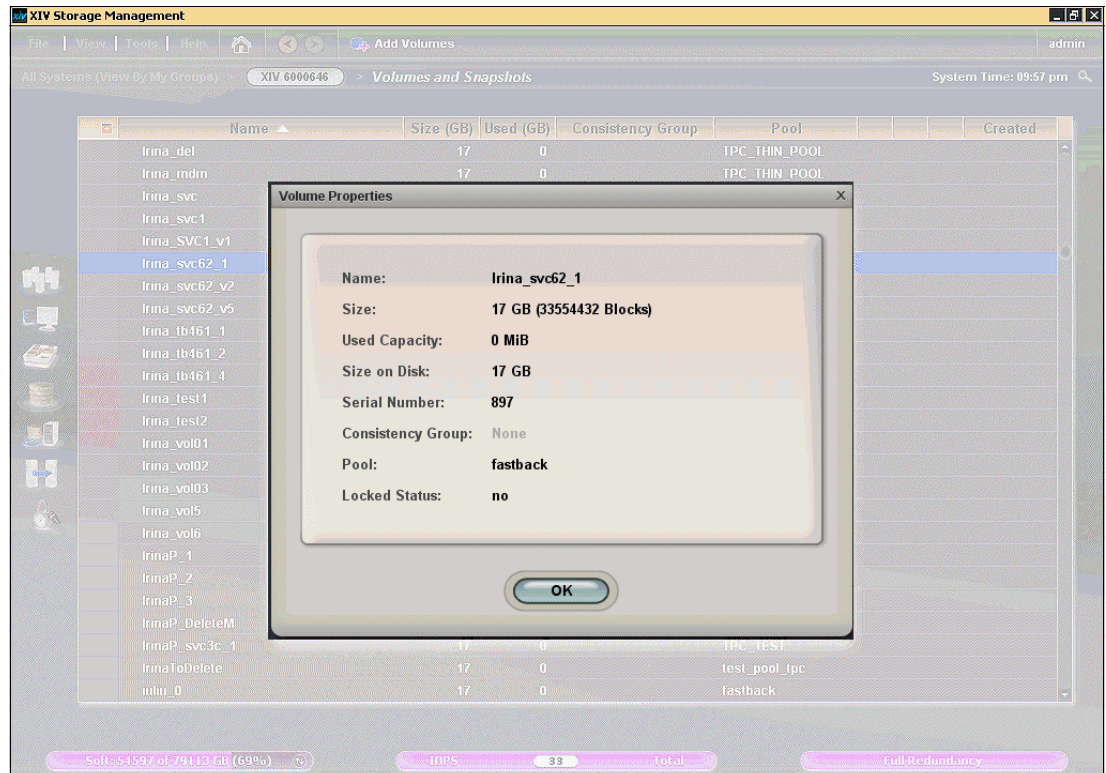


Figure 3-4 XIV Volume Properties dialog box

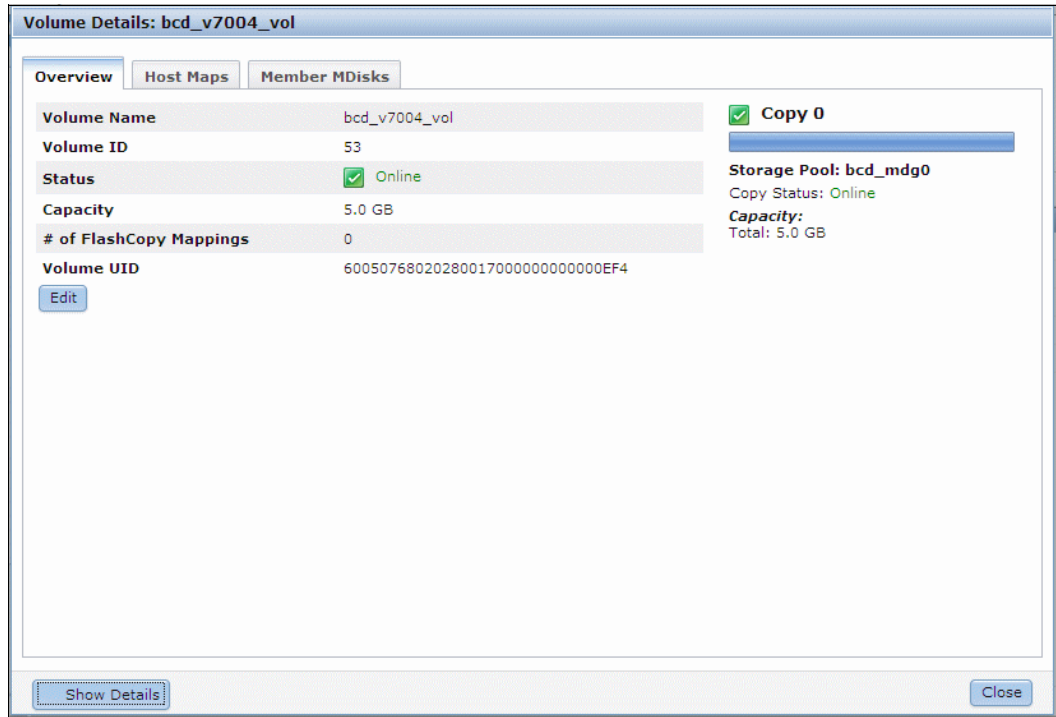


Figure 3-7 Storwize volume details

2. On the Host Maps tab (see Figure 3-8), check the SCSI ID number for the specific volume. This value is used to match the IBM Spectrum Virtualize ctrl_LUN_# (in hexadecimal format).

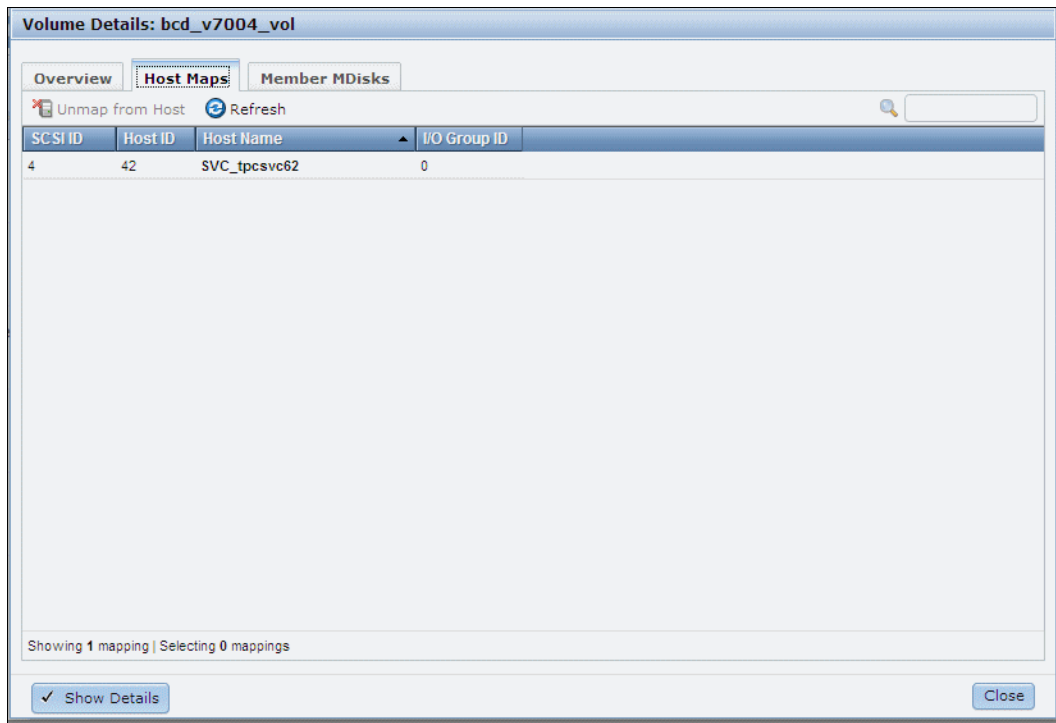


Figure 3-8 Storwize volume details for host maps

3.10 Controlling extent allocation order for volume creation

With early versions of IBM Spectrum Virtualize or Storwize, when you create a virtual disk, you might want to control the order in which extents are allocated across the MDisks in the storage pool to balance workload across controller resources. For example, you can alternate extent allocation across DA pairs and even and odd extent pools in the DS8000.

Today, when creating a new virtual disk, the first disk to allocate an extent from is chosen in a pseudo-random way rather than choosing the next disk in a round-robin fashion. The pseudo-random algorithm avoids the situation where the “striping effect” inherent in a round-robin algorithm places the first extent for many volumes on the same MDisk.

Placing the first extent of a number of volumes on the same MDisk might lead to poor performance for workloads that place a large I/O load on the first extent of each volume or that create multiple sequential streams. This effect caused poor performance with early code, and today has been completely fixed.

Since version 7.3, any kind of *extent congestion* is handled by Easy Tier itself that moves the extents around the MDisk to get the best performance balance, even in a case of a single Tier Pool, thanks to the Intra-Tier balance capability.

3.11 Considerations when using Encryption

SAN Volume Controller 2145-DH8 and 2145-SV1, or Storwize systems support optional encryption of data at rest. This support protects against the potential exposure of sensitive user data and user metadata that is stored on discarded, lost, or stolen storage devices. To use encryption on the system, an encryption license is required for each pair of SAN Volume Controller nodes or Storwize canisters that support encryption.

3.11.1 General considerations

USB encryption, key server encryption, or both can be enabled on the system. The system supports IBM Security Key Lifecycle Manager version 2.6.0 or later for enabling encryption with a key server. To encrypt data that is stored on drives, the SAN Volume Controller nodes or Storwize canisters that are capable of encryption must be licensed and configured to use encryption.

When encryption is activated and enabled on the system, valid encryption keys must be present on the system when the system unlocks the drives or the user generates a new key. If USB encryption is enabled on the system, the encryption key must be stored on USB flash drives that contain a copy of the key that was generated when encryption was enabled. If key server encryption is enabled on the system, the key is retrieved from the key server.

It is not possible to convert the existing data to an encrypted copy. You can use the volume migration function to migrate the data to an encrypted storage pool or encrypted child pool. Alternatively, you can also use the volume mirroring function to add a copy to an encrypted storage pool or encrypted child pool and delete the unencrypted copy after the migration.

Note: Hot Spare Nodes also need encryption licenses if they are to be used to replace the failed nodes that support encryption.

Before you activate and enable encryption, you must determine the method of accessing key information during times when the system requires an encryption key to be present. The system requires an encryption key to be present during the following operations:

- ▶ System power-on
- ▶ System restart
- ▶ User initiated rekey operations
- ▶ System recovery

Several factors must be considered when planning for encryption:

- ▶ Physical security of the system
- ▶ Need and benefit of manually accessing encryption keys when the system requires
- ▶ Availability of key data
- ▶ Encryption license is purchased, activated, and enabled on the system
- ▶ Using Security Key Lifecycle Manager clones

Note: It is suggested that IBM Security Key Lifecycle Manager version 2.7.0 or later is used for any new clone end points created on the system.

For configuration details about IBM Spectrum Virtualize and Storwize encryption, see the following publications:

- ▶ *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.1*, SG24-7933
- ▶ *Implementing the IBM Storwize V7000 with IBM Spectrum Virtualize V8.1*, SG24-7938

3.11.2 Hardware and software encryption

There are two ways to perform encryption on devices running IBM Spectrum Virtualize: hardware encryption and software encryption. Both methods of encryption protect against the potential exposure of sensitive user data that are stored on discarded, lost, or stolen media. Both can also facilitate the warranty return or disposal of hardware. Which method is used for encryption is chosen automatically by the system based on the placement of the data.

Figure 3-10 shows the encryption placement in the IBM Spectrum Virtualize and Storwize software stack.

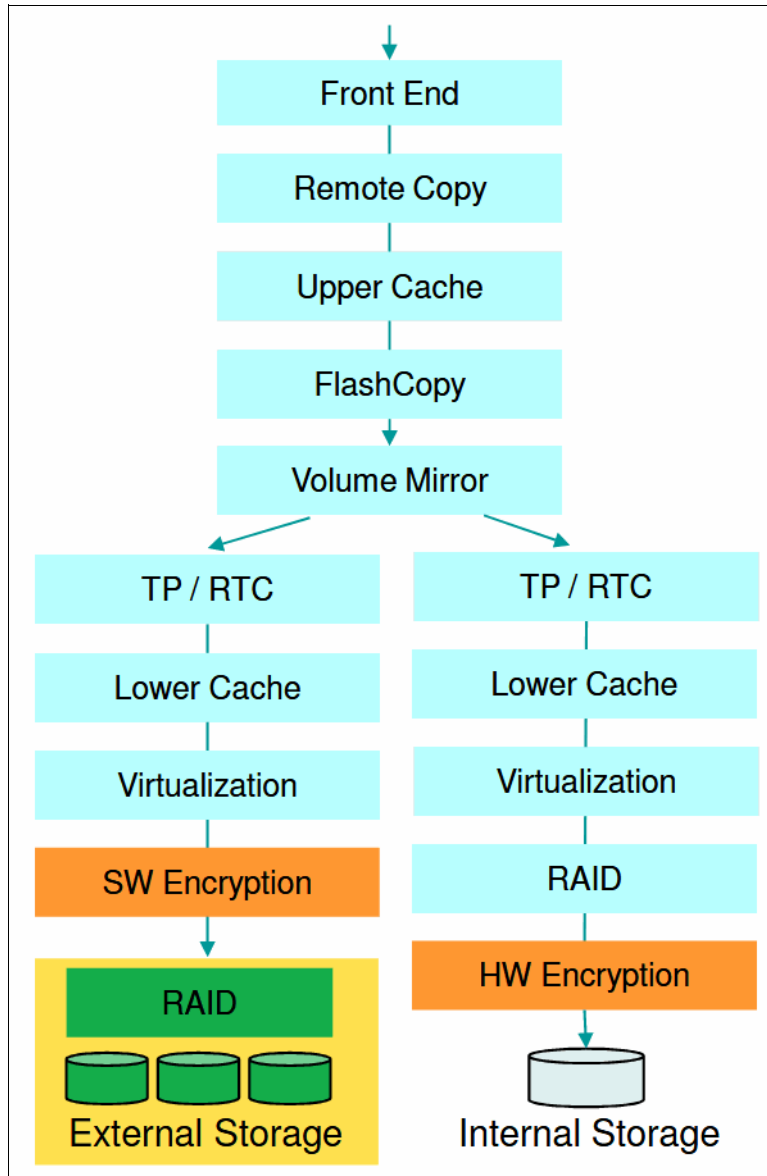


Figure 3-10 Encryption placement in the IBM Spectrum Virtualize and Storwize software stack

Hardware encryption only storage pool

Hardware encryption has the following characteristics:

- ▶ Algorithm is built in SAS chip
- ▶ No system overhead
- ▶ Only available to direct attached SAS disks
- ▶ Can only be enabled when you create internal arrays
- ▶ Child pools can not be encrypted if the parent storage pool is not encrypted
- ▶ Child pools are automatically encrypted if the parent storage pool is encrypted

Software encryption only storage pool

Software encryption has the following characteristics:

- ▶ Algorithm is running on SAN Volume Controller nodes or Storwize canisters
- ▶ Uses special CPU instruction set and engines (AES_NI)
- ▶ Allows encryption for virtualized external storages, which are not capable of self-encryption
- ▶ Potential system overhead
- ▶ Only available to virtualized external storages
- ▶ Can only be enabled when you create storage pools and child pools made up of virtualized external storages
- ▶ Child pools can be encrypted even if the parent storage pool is not encrypted

Mixed encryption in a storage pool

It is possible to mix hardware and software encryption in a storage pool as shown in Figure 3-11.

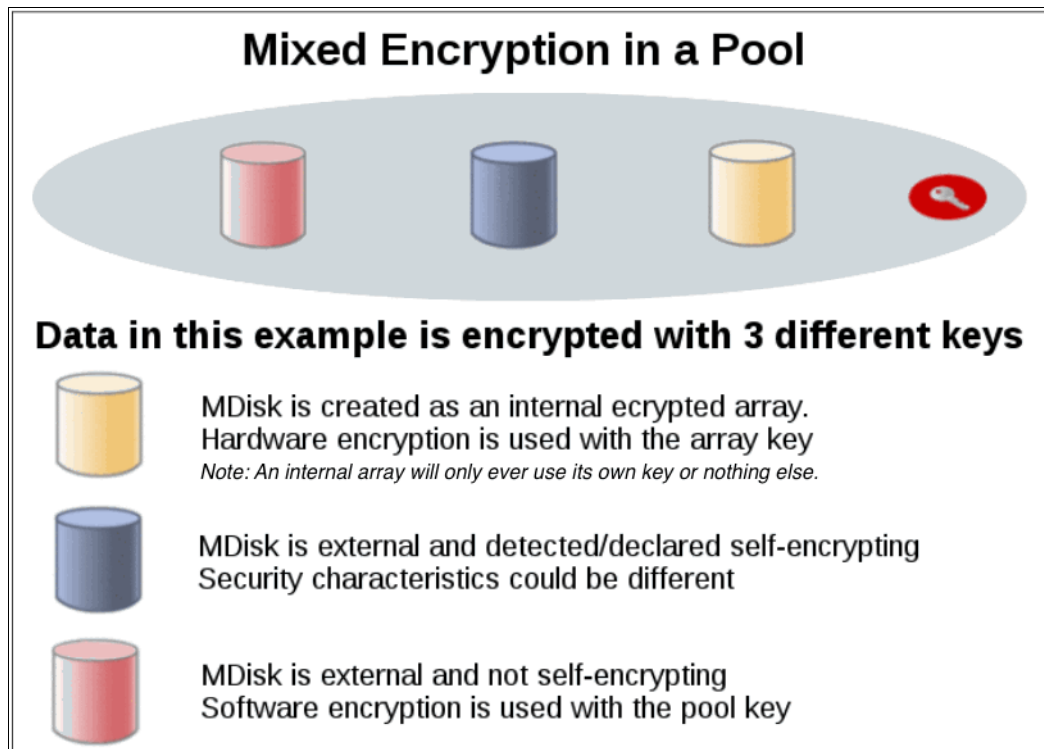


Figure 3-11 Mixed encryption in a storage pool

However, if you want to create encrypted child pools from an unencrypted storage pool containing a mix of internal arrays and external MDisks, the following restrictions apply:

- ▶ The parent pool must not contain any unencrypted internal arrays
- ▶ All SAN Volume Controller nodes or Storwize canisters in the system must support software encryption and have encryption license activated

Note: An encrypted child pool created from an unencrypted parent storage pool reports as unencrypted if the parent pool contains any unencrypted internal arrays. Remove these arrays to ensure that the child pool is fully encrypted.

The general rule is not to mix different types of MDisks in a storage pool, unless it is intended to use the Easy Tier tiering function. In this scenario, the internal arrays must be encrypted if you want to create encrypted child pools from an unencrypted parent storage pool. All the methods of encryption use the same encryption algorithm, the same key management infrastructure, and the same license.

Note: Always implement encryption on the self-encryption capable backend storage such as IBM Storwize V7000/V5000, IBM XIV, IBM FlashSystem 900/A9000/A9000R and IBM DS8000 to avoid potential system overhead.

Declare/identify the self-encrypted virtualized external MDisks as encrypted on IBM Spectrum Virtualize or Storwize by specifying the **-encrypt** option to **yes** with the **chmdisk** command as shown in Example 3-8. This configuration is important to avoid IBM Spectrum Virtualize or Storwize trying to encrypt them again.

Example 3-8 Command to declare/identify a self-encrypted MDisk from a virtualized external storage

```
IBM_2145:ITS0_DH8_A:superuser>chmdisk -encrypt yes mdisk0
```

Note: It is important to declare/identify the self-encrypted MDisks from a virtualized external storage before create a encrypted storage pool or child pool on IBM Spectrum Virtualize or Storwize.

3.11.3 Encryption at rest with USB keys

The following section describes the characteristics of using USB flash drives for encryption and the available options to access the key information.

USB flash drives have the following characteristics:

- ▶ Physical access to the system is required to process a rekeying operation
- ▶ No mechanical components to maintain with almost no read operations or write operations to the USB flash drive
- ▶ Inexpensive to maintain and use
- ▶ Convenient and easy to have multiple identical USB flash drives available as backups

Two options are available for accessing key information on USB flash drives:

- ▶ USB flash drives are left inserted in the system at all times

If you want the system to restart automatically, a USB flash drive must be left inserted in all the nodes on the system. When you power on, all nodes then have access to the encryption key. This method requires that the physical environment where the system is located is secure. If the location is secure, it prevents an unauthorized person from making copies of the encryption keys, stealing the system, or accessing data that is stored on the system.

- ▶ USB flash drives are not left inserted into the system except as required

For the most secure operation, do not keep the USB flash drives inserted into the nodes on the system. However, this method requires that you manually insert the USB flash drives that contain copies of the encryption key in the nodes during operations that the system requires an encryption key to be present. USB flash drives that contain the keys must be stored securely to prevent theft or loss. During operations that the system requires an encryption key to be present, the USB flash drives must be inserted manually into each node so data can be accessed. After the system completes unlocking the drives, the USB flash drives must be removed and stored securely to prevent theft or loss.

3.11.4 Encryption at rest with key servers

The following section describes the characteristics of using key servers for encryption and essential recommendations for key server configuration with IBM Spectrum Virtualize and Storwize.

Key servers

Key servers have the following characteristics:

- ▶ Physical access to the system is not required to process a rekeying operation
- ▶ Support for businesses that have security requirements not to use USB ports
- ▶ Strong key generation
- ▶ Key self-replication and automatic backups
- ▶ Implementations follow an open standard that aids in interoperability
- ▶ Audit detail
- ▶ Ability to administer access to data separately from storage devices

Encryption key servers create and manage encryption keys that are used by the system. In environments with a large number of systems, key servers distribute keys remotely without requiring physical access to the systems.

A key server is a centralized system that generates, stores, and sends encryption keys to the system. If the key server provider supports replication of keys among multiple key servers, you can specify up to 4 key servers (one master and three clones) that connect to the system over both a public network or a separate private network.

The system supports enabling encryption using an IBM Security Key Lifecycle Manager key server. All key servers must be configured on the IBM Security Key Lifecycle Manager before defining the key servers in the management GUI. IBM Security Key Lifecycle Manager supports Key Management Interoperability Protocol (KMIP), which is a standard for encryption of stored data and management of cryptographic keys.

IBM Security Key Lifecycle Manager can be used to create managed keys for the system and provide access to these keys through a certificate. If you are configuring multiple key servers, use IBM Security Key Lifecycle Manager 2.6.0.2 or later. The additional key servers (clones) support more paths when delivering keys to the system; however, during rekeying only the path to the primary key server is used. When the system is rekeyed, secondary key servers are unavailable until the primary has replicated the new keys to these secondary key servers.

Replication must complete before keys can be used on the system. You can either schedule automatic replication or complete it manually with IBM Security Key Lifecycle Manager. During replication, key servers are not available to distribute keys or accept new keys.

The time a replication completes on the IBM Security Key Lifecycle Manager depends on the number of key servers that are configured as clones and the amount of key and certificate information that is being replicated. The IBM Security Key Lifecycle Manager issues a completion message when the replication completes. Verify that all key servers contain replicated key and certificate information before keys are used on the system.

Recommendations for key server configuration

The following section provides some essential recommendations for key server configuration with IBM Spectrum Virtualize and Storwize.

Transport Layer Security

Define the IBM Security Key Lifecycle Manager to use Transport Layer Security version 2 (TLSv2). The default setting on IBM Security Key Lifecycle Manager is TLSv1, but the IBM Spectrum Virtualize and Storwize only support version 2. On the IBM Security Key Lifecycle Manager set the value to SSL_TLSv2, which is a set of protocols that includes TLSv1.2.

Example 3-9 shows the example of a SKLMConfig.properties configuration file. The default path on a Linux based server is

`/opt/IBM/WebSphere/AppServer/products/sklm/config/SKLMConfig.properties.`

Example 3-9 Example of a SKLMConfig.properties configuration file

```
#Mon Nov 20 18:37:01 EST 2017
KMIPListener.ssl.port=5696
Audit.isSyslog=false
Audit.syslog.server.host=
TransportListener.ssl.timeout=10
Audit.handler.file.size=10000
user.gui.init.config=true
config.keystore.name=defaultKeyStore
tklm.encryption.password=D1181E14054B1E1526491F152A4A1F3B16491E3B160520151206
Audit.event.types=runtime,authorization,authentication,authorization_terminate,resource_management,key_management
tklm.lockout.enable=true
enableKeyRelease=false
TransportListener.tcp.port=3801
Audit.handler.file.name=logs/audit/sklm_audit.log
config.keystore.batchUpdateTimer=60000
Audit.eventQueue.max=0
enableClientCertPush=true
debug=none
tklm.encryption.keysize=256
TransportListener.tcp.timeout=10
backup.keycert.before.serving=false
TransportListener.ssl.protocols=SSL_TLSv2
Audit.syslog.isSSL=false
cert.validate=false
config.keystore.batchUpdateSize=10000
useSKIDefaultLabels=false
maximum.keycert.expiration.period.in.years=50
config.keystore.ssl.certalias=sklm
TransportListener.ssl.port=441
Transport.ssl.vulnerableciphers.patterns=_RC4_,RSA_EXPORT,_DES_
Audit.syslog.server.port=
tklm.lockout.attempts=3
fips=off
Audit.event.outcome=failure
```

Self-signed certificate type and validity period

The default certificate type on IBM Security Key Lifecycle Manager server and IBM Spectrum Virtualize or Storwize is RSA. If it is intended to use different certificate type, make sure you match the certificate type on both end. The default certificate validity period is 1095 days on IBM Security Key Lifecycle Manager server and 5475 days on IBM Spectrum Virtualize or Storwize.

You can adjust the validity period to comply with specific security policies and always match the certificate validity period on IBM Spectrum Virtualize or Storwize and IBM Security Key Lifecycle Manager server. A mismatch will cause certificate authorization error and lead to unnecessary certificate exchange. Figure 3-12 shows the default certificate type and validity period on IBM Spectrum Virtualize and Storwize.

Update Certificate

Certificate type: Self-signed certificate
 Signed certificate

Key type: 2048-bit RSA

Validity days: 5,475

Figure 3-12 Update certificate on IBM Spectrum Virtualize and Storwize

Figure 3-13 shows the default certificate type and validity period on IBM Security Key Lifecycle Manager server.

Self-signed Certificate

*Certificate label in keystore:

*Certificate description (common name):

*Validity period of new certificate (in days; for example, 3 years is 365 x 3 = 1095 days):
1095 The interval in days ranges from 1 to 9000

*Algorithm:
RSA

Figure 3-13 Create self-signed certificate on IBM Security Key Lifecycle Manager server

Device group configuration

The SPECTRUM_VIRT device group is not pre-defined on IBM Security Key Lifecycle Manager, it must be created based on a GPFS device family as shown in Figure 3-14.

Create Device Group

*Device family:

Many asymmetric keys to many devices (3592)

Many devices to many keys with access via certificate (GPFS)

Many symmetric keys to many devices (LTO)

Symmetric Keys directly tied to a single device (DS5000) Enable machine affinity

*Device group name:

SPECTRUM_VIRT

Create Cancel

Figure 3-14 Create device group for IBM Spectrum Virtualize or Storwize

By default, IBM Spectrum Virtualize and Storwize has the SPECTRUM_VIRT pre-defined in the encryption configuration wizard, and SPECTRUM_VIRT contains all of the keys for the managed IBM Spectrum Virtualize and Storwize systems. However, It is possible to use different device groups as long as they are GPFS device family based. For example, one device group for each environment (Production or DR). Each device group maintains its own key database, and this approach allows more granular key management.

Clone servers configuration management

The minimum replication interval on IBM Security Key Lifecycle Manager is one hour, as shown in Figure 3-15 on page 106. It is more practical to perform backup and restore or manual replication for the initial configuration to speed up the configuration synchronization.

Also, the rekey process creates a new configuration on the IBM Security Key Lifecycle Manager server, and it is important not to wait for the next replication window but to manually synchronize the configuration to the additional key servers (clones), otherwise, an error message will be generated by the IBM Spectrum Virtualize or Storwize system indicating the key is missing on the clones.

Figure 3-15 shows the replication interval.

The screenshot displays the 'Advance Properties' configuration window for the SKLM Replication Schedule. It includes the following settings:

- Replication backup destination directory:** /opt/IBM/WebSphere/AppServ
- Maximum number of replication files to keep before rollover:** 2
- Replication Scheduler Section:**
 - Replication frequency (in hours):** 24 (selected)
 - Daily replication time (in HH:MM format):** 00:00
- Replication Log Section:**
 - Replication log file name:** replicationMaster.log
 - Maximum log file size (in KB):** 1000
 - Maximum number of log files to keep:** 30

Figure 3-15 SKLM Replication Schedule

Example 3-10 shows an example of manually triggered replication.

Example 3-10 Manually triggered replication

```
/opt/IBM/WebSphere/AppServer/bin/wsadmin.sh -username SKLMAdmin -password  
<password> -lang jython -c "print AdminTask.tklmReplicationNow()"
```

Encryption key management

There is always only one active key for each encryption enabled IBM Spectrum Virtualize or Storwize system. The previously used key is deactivated after the rekey process. It is possible to delete the deactivated keys to keep the key database tidy and up to date.

Figure 3-16 on page 107 shows the keys associated with a device group. In this example, the SG247933_REDB00K device group contains one encryption enabled Storwize V7000 system, and it has three associated keys. Only one of the keys is activated, and the other two were deactivated after the rekey process.

SG247933_REDBOOK

The screen below allows you to add or delete certificate and their associated node name. As well as, modify the node name associated with a certificate. New keys can be added and a name associated with that key(s).

Home Refresh Add Modify Delete

No filter applied		
Certificate UUID	Name	Endpoint Count
CERTIFICATE-8a89d57-70cfd4f7-adda-4b29-9b1c-89c200fd1745	sg247933_redbook	2

Total: 1 Selected: 0

No filter applied	
Key UUID	Name
KEY-8a89d57-15bf8f41-cea6-4df3-8f4e-be0c36318615	mmm008a89d57000000870
KEY-8a89d57-74edaef9-b6d9-4766-9b39-7e21d9911011	mmm008a89d5700000086e
KEY-8a89d57-ebe5d5a1-8987-4aff-ab58-5f808a078269	mmm008a89d5700000086f

Total: 3 Selected: 0

Figure 3-16 Keys associated to a device group

Example 3-11 shows an example to check the state of the keys.

Example 3-11 Verify key state

```
/opt/IBM/WebSphere/AppServer/bin/wsadmin.sh -username SKLMAdmin -password
<password> -lang jython
wsadmin>print AdminTask.tklmKeyList('[-uuid
KEY-8a89d57-15bf8f41-cea6-4df3-8f4e-be0c36318615]')
CTGKM0001I Command succeeded.
```

```
uuid = KEY-8a89d57-15bf8f41-cea6-4df3-8f4e-be0c36318615
alias = mmm008a89d57000000870
key algorithm = AES
key store name = defaultKeyStore
key state = ACTIVE
creation date = 18/11/2017, 01:43:27 Greenwich Mean Time
expiration date = null
```

```
wsadmin>print AdminTask.tklmKeyList('[-uuid
KEY-8a89d57-74edaef9-b6d9-4766-9b39-7e21d9911011]')
CTGKM0001I Command succeeded.
```

```
uuid = KEY-8a89d57-74edaef9-b6d9-4766-9b39-7e21d9911011
alias = mmm008a89d5700000086e
key algorithm = AES
key store name = defaultKeyStore
key state = DEACTIVATED
creation date = 17/11/2017, 20:07:19 Greenwich Mean Time
expiration date = 17/11/2017, 23:18:37 Greenwich Mean Time
```

```
wsadmin>print AdminTask.tklmKeyList('[-uuid
KEY-8a89d57-ebe5d5a1-8987-4aff-ab58-5f808a078269]')
```

CTGKM0001I Command succeeded.

```
uuid = KEY-8a89d57-ebe5d5a1-8987-4aff-ab58-5f808a078269
alias = mmm008a89d5700000086f
key algorithm = AES
key store name = defaultKeyStore
key state = DEACTIVATED
creation date = 17/11/2017, 23:18:34 Greenwich Mean Time
expiration date = 18/11/2017, 01:43:32 Greenwich Mean Time
```

Note: The initial configuration, such as certificate exchange and Transport Layer Security configuration, is only required on the master IBM Security Key Lifecycle Manager server. The restore or replication process will duplicate all required configurations to the clone servers.

If encryption was enabled on a pre-V7.8.0 code level system and the system is updated to V7.8.x or above, you must run a USB rekey operation to enable key server encryption. Run the **chencryption** command before you enable key server encryption. To perform a rekey operation, run the commands shown in Example 3-12.

Example 3-12 Commands to enable key server encryption option on a system upgraded from pre-7.8.0

```
chencryption -usb newkey -key prepare
chencryption -usb newkey -key commit
```

For the most up-to-date information about Encryption with Key Server, check the IBM Spectrum Virtualize or Storwize Knowledge Center at the following link:

<https://ibm.biz/Bdjvhm>

Also see the SKLM IBM Knowledge Center at the following link:

<https://ibm.biz/BdsvEP>



Volumes

This chapter shows recommendations about how to work with volumes (formerly known as VDisks), such as creating using different types, with different configurations, and migrating across pools and I/O groups.

This chapter includes the following sections:

- ▶ Overview of volumes
- ▶ Guidance for creating volumes
- ▶ Striped versus sequential volumes
- ▶ Thin-provisioned volumes
- ▶ Volume migration
- ▶ Preferred paths to a volume
- ▶ Changing the preferred node within or across I/O groups
- ▶ Volume throttling
- ▶ Volume cache mode
- ▶ VMware Virtual Volumes
- ▶ Additional considerations

4.1 Overview of volumes

There are different types of volumes:

- ▶ Non-high availability volumes:
 - Striped
 - Sequential
 - Image
- ▶ High availability volumes
- ▶ VMware vSphere Virtual volumes
- ▶ Cloud volumes

High availability volumes are designated for IBM HyperSwap and Stretched Cluster systems. Usage of this type of volumes is described in Appendix B, “Business continuity” on page 401.

The three types of non-high availability volumes are determined by how the extents are allocated from the storage pool. The *striped-mode volume* has extents that are allocated from each managed disk (MDisk) in the storage pool in a round-robin fashion. With a *sequential-mode volume*, extents are allocated sequentially from an MDisk. And an *image-mode volume* is a one-to-one mapped extent mode volume. This last one is used for importing (preserving) data on a managed disk from another storage system.

VMware vSphere Virtual Volumes, sometimes referred to as VVols, are volumes that allow VMware vCenter to automate the management of system objects like volumes and pools. Details on implementing VVols can be found in *Configuring VMware Virtual Volumes for Systems Powered by IBM Spectrum Virtualize*, SG24-8328.

A cloud volume is any volume that is enabled for transparent cloud tiering. When transparent cloud tiering is enabled on a volume, point-in-time copies, or snapshots, can be created and copied to cloud storage that is provided by a cloud service provider. Information to configure transparent cloud tiering on the system and manage cloud volumes can be found in IBM Knowledge Center:

<https://ibm.biz/BdjAWS>

4.2 Guidance for creating volumes

To create volumes, follow the procedure that is described in *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.1*, SG24-7933.

When you are creating volumes, adhere to the following guidelines. More information about each of these items can be found in the next sections of this chapter:

- ▶ Decide on your naming convention before you begin. It is much easier to assign the correct names when the volume is created than to modify them afterwards.
- ▶ Plan the type of volume that you will create. If it will be thick (fully allocated) or thin-provisioned, striped or sequential, and if you decide to create a thin-provisioned volume if it will have compression enabled or not. To help you on the decision to use compression or not, you can refer to 4.4.1, “Compressed volumes” on page 115.
 - If you decide to create a fully allocated volume, you have to take into consideration that starting with V7.5 of IBM Spectrum Virtualize, volumes are automatically formatted through the quick initialization process right after their creation.

This quick initialization is a background process that fills a fully allocated volume with zeros. This does not affect the usage of the volume. It is available for use immediately. But, actions like moving, expanding, shrinking, or adding a volume copy are disabled when the specified volume is initializing. If you don't want the volume to be automatically formatted after its creation, you can disable this function on the Custom tab of the volume creation window in the GUI, or by using the CLI parameter **-nofmtdisk** in the command line.

Therefore, remember this when creating a volume, because depending on the volume size, it takes a long time to complete, and those actions won't be available for the volume until the initialization finishes. For reference, the initialization of a 1 TB volume can take around 6 days to complete with the default syncrate value 50, or around 4.5 hours if you manually set the syncrate to 100. Example 4-1 shows the command to create a volume with the **-nofmtdisk** parameter.

Example 4-1 Volume creation with no formatting option

```
IBM_2145:ITSO_DH8_A:superuser>mkvdisk -name TESTVOL60 -mdiskgrp Pool0 -size
1 -unit gb -vtype striped -iogrp io_grp0 -nofmtdisk
Virtual Disk, id [38], successfully created
IBM_2145:ITSO_DH8_A:superuser>lsvdisk TESTVOL60
id 38
name TESTVOL60
IO_group_id 0
IO_group_name io_grp0
status online
mdisk_grp_id 0
mdisk_grp_name Pool0
capacity 1.00GB
type striped
formatted no
formatting no
.
lines removed for brevity
```

- If you decide to create a thin-provisioned volume, with compression enabled or not, you have to be careful if you run out of space in the volume and pool where the volume is created. You should set the warning threshold in the pools containing thin-provisioned volumes, and you can set warning threshold in the volume too.

However, if you configure the thin volume with the **autoexpand** option enabled, which is highly recommended, you don't need to worry about monitoring volume capacity and you can set the warning in the volume level to 0 (disabled). Also, when you create a thin volume, you must specify the space which is initially allocated to it (**-rsize** option in the CLI) and the grain size.

By default, **rsize** (or real capacity) is set to 2% of the volume virtual capacity, and grain size is 256 KiB. These default values, with **autoexpand** enabled and warning disabled options will work in most scenarios. There are some cases that you might consider using different values, and they are shown in the next sections of this chapter with more details about these parameters.

Example 4-2 shows the command to create a volume with the parameters mentioned previously.

Example 4-2 Thin-provisioned volume creation

```
IBM_2145:ITSO_DH8_A:superuser>mkvdisk -name TESTVOL70 -mdiskgrp Pool0 -size
100 -unit gb -vtype striped -iogrp io_grp0 -rsize 2% -autoexpand -warning 0
-grainsize 256
Virtual Disk, id [40], successfully created
IBM_2145:ITSO_DH8_A:superuser>lsvdisk TESTVOL70
id 40
name TESTVOL70
.
lines removed for brevity
.
capacity 100.00GB
.
lines removed for brevity
.
used_capacity 0.75MB
real_capacity 2.02GB
free_capacity 2.01GB
overallocation 4961
autoexpand on
warning 0
grainsize 256
se_copy yes
.
lines removed for brevity
```

- ▶ Each volume has an I/O group and preferred node that balances the load between nodes in the I/O group. Therefore, balance the volumes across the I/O groups in the cluster to balance the load across the cluster.

In configurations with many attached hosts where it is not possible to zone a host to multiple I/O groups, you might not be able to choose to which I/O group to attach the volumes. The volume must be created in the I/O group to which its host belongs.

Tip: Migrating volumes across I/O groups can be a disruptive action. Therefore, specify the correct I/O group at the time the volume is created.

Also, when a volume is being created, it is possible to define a list of I/O groups in which a volume can be accessible to hosts. By default, only the caching I/O group, which is the I/O group in which the volume is created, is in this list. It is recommended that a volume is accessible to hosts only by the caching I/O group. You can have more than one I/O group in the access list of a volume only in some scenarios with specific requirements, like when a volume is being migrated to another I/O group.

- ▶ By default, the *preferred node*, which owns a volume within an I/O group, is selected on a *load balancing* basis. At the time that the volume is created, the workload to be placed on the volume might be unknown. However, you must distribute the workload evenly on each node within an I/O group. If you must change the preferred node, see 4.3, “Striped versus sequential volumes” on page 114.
- ▶ In Stretched Cluster environments, it is best to configure the *preferred node* based on site awareness.

- ▶ With the exception of a few cases, cache mode of a volume should be set to readwrite. More details can be found in section 4.9, “Volume cache mode” on page 130.
- ▶ The maximum number of volumes per I/O group and system is described in Table 4-1.

Table 4-1 Maximum number of volumes in IBM SVC and IBM Storwize V7000

Volume type	Maximum number	Comments
Basic volumes (VDisks) per system	10,000	Each basic volume uses 1 VDisk, each with one copy.
Stretched volumes per system ^a	5,000	Each stretched volume uses 1 VDisk, each with two copies.
HyperSwap volumes per system	1,250	Each HyperSwap volume uses 4 VDisks, each with one copy, 1 active-active remote copy relationship and 4 FlashCopy mappings.
Volumes per I/O group (volumes per caching I/O group)	10,000	
Volumes accessible per I/O group	10,000	
Thin-provisioned (space-efficient) volume copies per system	-	No limit is imposed here beyond the volume copies per system limit.
Volume copies per system	10,000	
Compressed volume copies per I/O group	512	IBM SVC: with 64GB memory option only. IBM Storwize V7000: with 32GB cache upgrade and 2nd compression accelerator card installed
Compressed volume copies per system	2,048	IBM SVC: requires an 8-node cluster. IBM Storwize V7000: requires 4 control enclosures.

a. This does not apply for IBM Storwize V7000 systems.

- ▶ The pool extent size does not affect performance, and it is set during Storage Pool creation. A volume occupies an integer number of extents, but its length does not need to be an integer multiple of the extent size. The length does need to be an integer multiple of the block size. Any space left over between the last logical block in the volume and the end of the last extent in the volume is unused.

A small extent size is used to minimize this unused space, and also to have a finer granularity of the volume space that is occupied on the underlying storage controller. On the other hand, you have to consider the best extent size for your storage pools considering the back-end storage. More details on extent size can be found in Chapter 3, “Storage pools and managed disks” on page 75.

Important: You can migrate volumes by using the `migratevdisk` command only between storage pools that have the same extent size. If you need to migrate volumes between storage pools with different extent sizes, you have to use volume mirroring.

- ▶ As mentioned previously, a volume can be created as thin-provisioned or fully allocated, in one mode (striped, sequential, or image), and with one or two copies (volume mirroring). With a few rare exceptions, you must always configure volumes by using striped mode.

4.3 Striped versus sequential volumes

With a few exceptions, you should always use striped volumes.

The main target is to take the 60 MDisk queue depth into account, which typically says for HDD MDisks, to aim for eight spindles per MDisk. Previous cache algorithms meant that for sequential workloads and to avoid stripe on stripe issues, you had to keep things more logical and doing the single pool or single volume per MDisk on the underlying storage. This is no longer valid now with the cache algorithms of today (for reference it was added with V7.3).

So here, the rule would be, for the number of drives you have on the backend in a pool, say 64 drives in the pool, then $64/8 = 8$ volumes created on the backend and presented to IBM Spectrum Virtualize or Storwize as 8 MDisks. This rule means that when you get the 60 MDisk queue depth, you get roughly a queue depth of 8 per drive. This setting keeps a spinning disk well used. It also gives better concurrency across the ports of the back-end controller.

4.3.1 Use cases of sequential volumes

One exception for not using striped volumes, and using sequential volumes, is in an environment in which you have a 100% sequential workload, and disk loading across all volumes is guaranteed to be balanced by the nature of the application. An example of this exception is specialized video streaming applications.

Another exception for not using striped volume is an environment with a high dependency on many FlashCopies. In this case, FlashCopy loads the volumes evenly, and the sequential I/O, which is generated by the flash copies, has a higher throughput than what is possible with striping. This situation is rare, considering that you almost won't need to optimize for FlashCopy as opposed to an online workload.

There could be other scenarios, such as when IBM Spectrum Virtualize or Storwize is acting as back-end storage for IBM TS7650G ProtecTIER Gateway. In this scenario, generally it is better to use sequential volumes mostly when using disk drives with very large sizes, such as 2 TB or 3 TB for the user data repository. The reason is that those large disk drives end up having very large arrays/MDisks/LUNs. If ProtecTIER handles this large LUN by itself, it is able to optimize its file system structure and workload without overcommitting or congesting a single array, rather than striping the LUNs over an entire multi-array Storage Pool.

4.4 Thin-provisioned volumes

Volumes can be configured as fully allocated or thin-provisioned. Fully allocated volumes are created with the same amount of real capacity and virtual capacity. A thin-provisioned volume is created to save capacity, so it has different virtual capacity and real capacity. You can still create volumes by using a striped, sequential, or image mode virtualization policy as you can with any other volume.

Real capacity defines how much disk space from a pool is allocated to a volume. *Virtual capacity* is the capacity of the volume that is reported to other IBM Spectrum Virtualize or Storwize components (such as FlashCopy or remote copy) and to the hosts.

A directory maps the virtual address space to the real address space. The directory and the user data share the real capacity.

Thin-provisioned volumes are available in two operating modes: *Autoexpand* and *nonautoexpand*. You can switch the mode at any time. If you select the autoexpand feature, IBM Spectrum Virtualize or Storwize automatically adds a fixed amount of extra real capacity to the thin volume as required. Therefore, the autoexpand feature attempts to maintain a fixed amount of unused real capacity for the volume.

This amount is known as the *contingency capacity*. The contingency capacity is initially set to the real capacity that is assigned when the volume is created. If the user modifies the real capacity, the contingency capacity is reset to be the difference between the used capacity and real capacity.

A volume that is created *without* the **autoexpand** feature, and therefore has a zero contingency capacity, goes offline when the real capacity is used. In this case, it must be expanded.

Warning threshold: When you are working with thin-provisioned volumes, enable the warning threshold (by using email or an SNMP trap) in the storage pool. If you are not using the **autoexpand** feature, you must enable the warning threshold on the volume side too. These settings help to monitor the volume capacity, because if the pool or volume runs out of space, the thin volume goes offline.

Autoexpand mode does not cause real capacity to grow much beyond the virtual capacity. The real capacity can be manually expanded to more than the maximum that is required by the current virtual capacity, and the contingency capacity is recalculated.

A thin-provisioned volume can be converted nondisruptively to a fully allocated volume, or vice versa, by using the volume mirroring function. For example, you can add a thin-provisioned copy to a fully allocated primary volume and then remove the fully allocated copy from the volume after they are synchronized.

The fully allocated to thin-provisioned migration procedure uses a zero-detection algorithm so that grains that contain all zeros do not cause any real capacity to be used.

Tip: Consider the use of thin-provisioned volumes as targets in FlashCopy relationships.

4.4.1 Compressed volumes

A compressed volume is, first of all, a thin-provisioned volume. The compression technology is implemented into the IBM Spectrum Virtualize or Storwize Thin Provisioning layer and is an organic part of the stack.

You can create, delete, migrate, mirror, map (assign), and unmap (unassign) a compressed volume as though it were a fully allocated volume. This compression method provides nondisruptive conversion between compressed and uncompressed volumes. This conversion provides a uniform user experience and eliminates the need for special procedures to deal with compressed volumes.

For more information about compression technology, see *IBM Real-time Compression in IBM SAN Volume Controller and IBM Storwize V7000*, REDP-4859.

When using Real-time Compression (RtC), always use IBM Spectrum Virtualize nodes or Storwize hardware with dedicated RtC CPU and RtC accelerator cards installed where available.

Refer to your IBM SSR or representative before implementing RtC in production, so that person can perform a space and performance assessment first.

To identify the best volume candidates to be compressed, use the RtC estimator tool that is available on your IBM Spectrum Virtualize and Storwize CLI starting with V7.6, and with the GUI starting from V7.7.

When using the CLI, use the commands shown in Example 4-3 to run volume analysis on a single volume.

Example 4-3 An analyzevdisk command example

```
IBM_2145:ITS0_DH8_B:superuser>svctask analyzevdisk -h
```

```
analyzevdisk
```

Syntax

```
>>- analyzevdisk -- --+-----+-- --+ vdisk_id ---+-----><
                        '- -cancel-'      '- vdisk_name -'
```

For more details type 'help analyzevdisk'.

```
IBM_2145:ITS0_DH8_B:superuser>svctask analyzevdisk volume0
```

```
IBM_2145:ITS0_DH8_B:superuser>
```

When using the CLI, use the commands shown in Example 4-4 to run Volume analysis for an entire subsystem.

Example 4-4 An analyzevdiskbysystem command example

```
IBM_2145:ITS0_DH8_B:superuser>svctask analyzevdiskbysystem -h
```

```
analyzevdiskbysystem
```

Syntax

```
>>- analyzevdiskbysystem -- --+-----+-- -----><
                        '- -cancel-'
```

For more details type 'help analyzevdiskbysystem'.

```
IBM_2145:ITS0_DH8_B:superuser>svctask analyzevdiskbysystem
```

```
IBM_2145:ITS0_DH8_B:superuser>
```

Note: The **analyzevdisk** and **analyzevdiskbysystem** commands return to the prompt.

To see the result of the analysis and its progress, run the CLI commands shown in Example 4-5.

Example 4-5 An example of lsvdiskanalysis and lsvdiskanalysis progress commands

```
IBM_2145:ITS0_DH8_B:superuser>svcinfo lsvdiskanalysis
```

```
id name          state  analysis_time capacity thin_size thin_savings
thin_savings_ratio compressed_size compression_savings compression_savings_ratio
total_savings total_savings_ratio margin_of_error
```



```

0 volume0 sparse 171109225031 10.00GB 0.00MB 0.00MB 0
0.00MB 0.00MB 0 0.00MB 0
0
.
lines omitted for brevity
.
5 volume3tp sparse 171109225041 10.00GB 0.00MB 0.00MB 0
0.00MB 0.00MB 0 0.00MB 0
0
IBM_2145:ITSO_DH8_B:superuser>svcinfolsvdiskanalysisprogress
vdisk_count pending_analysis estimated_completion_time
6 0
IBM_2145:ITSO_DH8_B:superuser>

```

When using the GUI, go to the menu shown in Figure 4-1 to run volume analysis by single volume or by multiple volumes. Select all of the volumes that you need to be analyzed.

From the same menu shown in Figure 4-1, you can download the report in csv format.

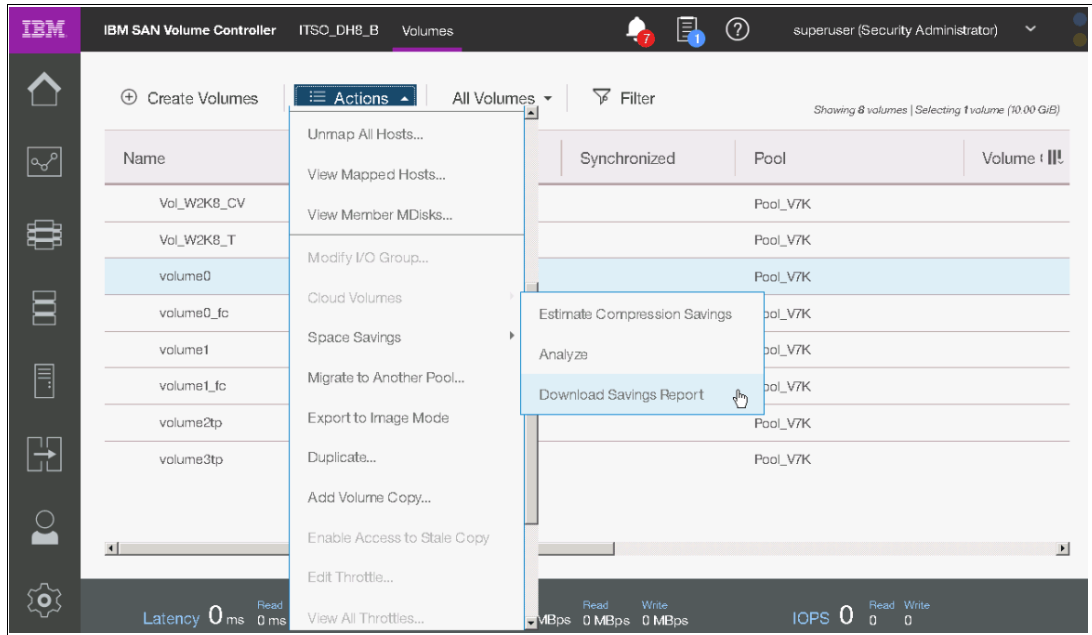


Figure 4-1 Use of Estimate Compression Saving with GUI

If you are planning to virtualize volumes that are connected to your hosts directly from any storage subsystems, and you want to know what the space saving you will achieve using RfC on those volumes, run the Comprestimator Utility:

<http://www-304.ibm.com/webapp/set2/sas/f/comprestimator/home.html>

Comprestimator is a command-line, host-based utility that can be used to estimate an expected compression rate for block devices. The previous link provides all of the instructions needed.

The following actions are preferred practices:

- ▶ After you run Comprestimator, consider applying RtC only on those volumes that show a capacity saving of not less than 40%. For other volumes, the tradeoff between space saving and hardware resource consumption to compress your data might not make sense.
- ▶ After you compress your selected volumes, look at which volumes have the most space saving benefits from Thin Provisioning rather than RtC. Consider moving these volumes to Thin Provisioning only. This configuration requires some effort, but saves hardware resources that are then available to give better performance to those volumes, which achieves more benefit from RtC than Thin Provisioning.

The GUI can help you by going to the Volumes menu and selecting the fields shown in Figure 4-2. Customize the Volume view to get all the metrics you might need to help make your decision.

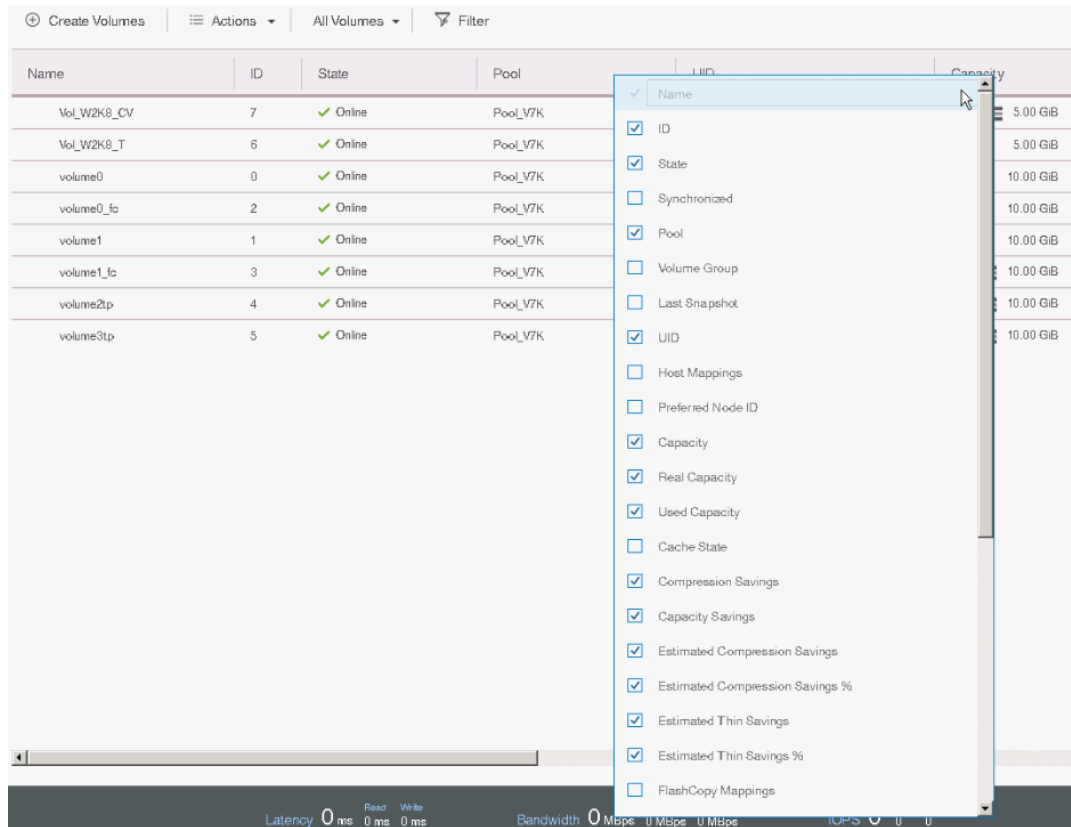


Figure 4-2 Customized view

4.4.2 Space allocation

When a thin-provisioned volume is created, a small amount of the real capacity is used for initial metadata. Write I/Os to the grains of the thin volume (that were not previously written to) cause grains of the real capacity to be used to store metadata and user data. Write I/Os to the grains (that were previously written to) update the grain where data was previously written.

Grain definition: The grain value is defined when the volume is created and can be 32 KB, 64 KB, 128 KB, or 256 KB (default). The grain size cannot be changed after the thin-provisioned volume is created.

Smaller granularities can save more space, but they have larger directories. If you select 32 KB for the grain size, the volume size cannot exceed 260,000 GB. Therefore, if you are not going to use the thin-provisioned volume as a FlashCopy source or target volume, use 256 KB to maximize performance.

Now, if you are planning to use Thin Provisioning with FlashCopy, you must remember that grain size for thin-provisioned FlashCopy volumes can be only 64 KB or 256 KB. In addition, to achieve best performance, the grain size for the thin-provisioned volume and FlashCopy mapping must match.

4.4.3 Thin Provisioning considerations

Thin Provisioning is a well-understood technology in the storage industry and it saves capacity only if the host server does not write to whole volumes. Whether the thin-provisioned volume works well, partly depends on how the file system has allocated the space.

A volume that is thin-provisioned by IBM SVC or Storwize is a volume where the large chunk of binary zeros are not stored in the storage pool. So, if you have not written to the volume yet, you do not need to use valuable resources storing data that does not exist yet in the form of zeros.

It is important to note that there are some file systems that are more Thin Provisioning friendly than others. Figure 4-3 shows some examples. This is not an official reference, but it is information that is based on experience and observation.

File System	OS	Thin-Friendly?
VMFS	VMware	yes
NTFS	Windows	Yes (quick format)
Ext 3	Linux	no
Ext 2	Linux	no
ReiserFS	Linux	yes
XFS	SGI	yes
UFS	Solaris	no
Veritas File System (VxFS)	HP, Sun, Linux, AIX	yes
IBM GPFS	AIX/Linux	no
ZFS	Solaris	yes
Oracle OCFS2	Linux	no
HFS	HP/UX	no
ASM*	Oracle	yes
JFS2	AIX/Linux	yes
JFS (VxFS)	HP/UX	yes

Figure 4-3 Friendly file systems

There are a number of different properties of thin-provisioned volumes that are useful to understand for the rest of the chapter:

- ▶ The size of the volume presented to the host. This does not really have a name, but we refer to this concept as *volume capacity*.
- ▶ The amount of user data that has been written to the storage pool. This is called the *used capacity*. In a compressed volume, this is the data written to storage pool after user data is compressed.

- ▶ The capacity that has been removed from the storage pool and has been dedicated to this volume. This is called the *real capacity*. The real capacity must always be greater than the used capacity.
- ▶ When the used capacity first exceeds the volume *warning threshold*, an event is raised, indicating additional real capacity is required. The default warning threshold value is 80% of the volume capacity. To disable warnings, specify 0 or 0%.
- ▶ For a Compressed Volume only (because Compressed Volumes are based on thin-provisioned volumes), there is the amount of uncompressed user data that has been written onto the volume. This is called the *uncompressed used capacity*. In other words, the uncompressed used capacity is the used capacity if data had not be compressed. It is used to calculate the compression ratio:

$$\frac{((\text{uncompressed used capacity} - \text{used capacity}) / \text{uncompressed used capacity})}{\text{compression ratio}}$$

Because there are at least two ways of calculating compression ratios, it is useful to remember that bigger is better, so a 90% compression ratio is better than 50% compression ratio.

Thin Provisioning and overallocation

As stated, Thin Provisioning means “don’t store the zeros,” so what does *overallocation* mean? To put it simply, a storage pool is only overallocated after the sum of all volume capacities exceeds the size of the storage pool.

One of the things that worries administrators the most is the question “what if I run out of space?”.

The first thing to remember is that if you already have enough capacity on disk to store fully allocated volumes, and then you convert to Thin Provisioning, you will have enough space to store everything even if the server writes to every byte of virtual capacity. So, this is not going to be a problem for the short term, and you will have time to monitor your system and understand how your capacity grows. But, you must monitor it.

Even if you are creating a storage pool, it is likely that you will not start over provisioning for a few weeks after you start writing to that pool. You do not actually need to overallocate until you feel comfortable that you have a handle on Thin Provisioning.

How do I monitor Thin Provisioning?

The basics of capacity planning for Thin Provisioning or compressed volumes are no different than capacity planning for fully allocated volumes. The capacity planner needs to monitor the amount of capacity being used versus the capacity of the storage pool. Make sure that you purchase more capacity before you run out.

The main difference is that in a fully allocated world, the used capacity normally only increases during working hours because the increase is caused by an administrator creating more volumes. In a Thin Provisioning world, the used capacity can increase at any time as long as the File Systems grow. Thus you need to approach capacity planning carefully.

To avoid unpleasant situations where some volumes can go offline due to lack of space, the storage administrator needs to monitor the real capacity rather than the volume capacity. And that is the main difference. Of course, they need to monitor it regularly because the real capacity can increase at any time of day for any reason.

Tools like IBM Spectrum Control™ can capture the real capacity of a storage pool and enable you to graph the real capacity so you can see how it is growing over time. Having a tool to show how the real capacity is growing over time is an important requirement to be able to predict when the space will run out.

IBM Spectrum Virtualize or Storwize also alert you by putting an event into the event log when the storage pool breaches a configurable threshold, called the warning level. The GUI sets this threshold to 80% of the capacity of the storage pool by default, although you can change it.

Have event notifications turned on so that someone gets an email or pop up on your monitoring system when an error is added to the event log. Note that this event will not call home to IBM. You need to respond to this notification yourself.

What to do if you run out of space

There are numerous options here. You can use just one of these options, or a combination of as many as you like.

Consider if one server decides to write to the space that you allocated to it and it uses up all of the free space in the storage pool. If the system does not have any more capacity to store the host writes, then the volume goes offline. But it is not only that one volume that goes offline. All the volumes in the storage pool are now at risk of going offline.

The following mechanisms and processes can help you deal with this situation:

- ▶ Automatic out of space protection provided by the product

If the storage pool runs out of space, each volume now has its own emergency capacity. That emergency capacity is normally sizable (2% is the default). The emergency capacity that is dedicated to a volume could allow that volume to stay online for anywhere between minutes to days depending on the change rate of that volume. This feature means that when you run out of space, you do have some time to repair things before everything starts going offline.

So you might implement a policy of 10% emergency capacity per volume if you wanted to be safer. Also, remember that you do not need to have the same contingency capacity for every volume.

Note: This automatic protection will probably solve most immediate problems, but remember that after you are informed that you have run out of space, you have a limited amount of time to react. You need a plan about what to do next.

- ▶ Buy more storage

This is the simplest solution when you run out of space. It doesn't matter if you use thin-provisioned or fully allocated volumes, buying additional storage will always resolve your capacity issues. You should calculate how much of additional storage you need, and estimate for how long it will support your needs. This option will cost you money for procuring, and also will take some days to have new storage available in your IBM SVC or Storwize, which should not be acceptable for your business.

- ▶ Have unallocated storage on standby

You can always have spare drives or managed disks ready to be added to whichever storage pool runs out of space within only a few minutes. This capacity gives you some breathing room while you take other actions. The more managed disks or drives that you have available, the more time you have to solve the problem.

► Move or delete volumes

You can migrate volumes to other pools to free up space. This technique is useful. However, data migration on IBM Spectrum Virtualize and Storwize is designed to go slowly to avoid causing performance problems. Therefore, it might be impossible to complete this migration before your applications go offline.

A very rapid but extreme solution is to delete one or more volumes to make space. If you have unused volumes, it shouldn't be a problem to delete them. But, if they are in use, only in a very critical situation the deletion will be an option. This can be used if you are sharing the storage pool with both production and development. You might choose to sacrifice less important volumes to preserve the critical volumes.

► Policy-based solutions

No policy is going to solve the problem if you run out of space, but you can use policies to reduce the likelihood of that ever happening to the point where you feel comfortable doing less of the other options.

You can use these types of policies for Thin Provisioning:

Note: The following policies use arbitrary numbers. These arbitrary numbers are designed to make the suggested policies more readable. We do not give any recommended numbers to insert into these policies because they are determined by business risk, and this consideration is different for every client.

- Manage free space such that there is always enough free capacity for your 10 biggest volumes to reach 100% full without running out of free space.
- Never overallocate more than 200%. In other words, if you have 100 TB of capacity in the storage pool, then the sum of the volume capacities in the same pool must not exceed 200 TB.
- Always start the process of buying more capacity when the storage pool reaches 60% full.

► Child Pools

Version 7.4 introduced a feature called child pools that allows you to make a storage pool that takes its capacity from a parent storage pool rather than from managed disks. This has a couple of possible use cases for this Thin Provisioning:

- You could separate different applications into different child pools. This technique prevents any problems with a server in child pool A affecting a server in child pool B. If Child Pool A runs out of space, and the parent pool still has space, then you can easily grow the child pool.
- You can use child pools to create a child pool that is called something descriptive like "DO NOT USE" and allocate (for example) 10% of the storage pool capacity to that child pool. Then, if the parent pool ever runs out, you have emergency capacity that can be given back to the parent pool. In this technique, you must figure out which server was eating up all the space and stop whatever it was doing.

For considerations and recommendations about Thin Provisioning usage, see the following link:

<https://ibm.biz/BdschV>

4.4.4 Limits on virtual capacity of thin-provisioned volumes

The extent and grain size factors limit the virtual capacity of thin-provisioned volumes beyond the factors that limit the capacity of regular volumes. Table 4-2 shows the maximum thin-provisioned volume virtual capacities for an extent size.

Table 4-2 Maximum thin volume virtual capacities for an extent size

Extent size in MB	Maximum volume real capacity in GB	Maximum thin virtual capacity in GB
16	2,048	2,000
32	4,096	4,000
64	8,192	8,000
128	16,384	16,000
256	32,768	32,000
512	65,536	65,000
1024	131,072	130,000
2048	262,144	260,000
4096	262,144	262,144
8192	262,144	262,144

Table 4-3 shows the maximum thin-provisioned volume virtual capacities for a grain size.

Table 4-3 Maximum thin volume virtual capacities for a grain size

Grain size in KB	Maximum thin virtual capacity in GB
32	260,000
64	520,000
128	1,040,000
256	2,080,000

4.5 Volume migration

A volume can be migrated from one storage pool to another storage pool regardless of the virtualization type (image, striped, or sequential). The command varies depending on the type of migration, as shown in Table 4-4.

Table 4-4 Migration types and associated commands

Storage pool-to-storage pool type	Command
Managed-to-managed or Image-to-managed	<code>migratevdisk</code>
Managed-to-image or Image-to-image	<code>migratetoimage</code>

Migrating a volume from one storage pool to another is nondisruptive to the host application using the volume. Depending on the workload of IBM Spectrum Virtualize or Storwize, there might be a slight performance impact. For this reason, migrate a volume from one storage pool to another when the SAN Volume Controller has a relatively low load.

Migrating a volume from one storage pool to another storage pool: For the migration to be acceptable, the source and destination storage pool *must* have the same extent size. Volume mirroring can also be used to migrate a volume between storage pools. You can use this method if the extent sizes of the two pools are not the same.

This section provides guidance for migrating volumes.

4.5.1 Image-type to striped-type migration

When you are migrating existing storage into the IBM Spectrum Virtualize cluster, the existing storage is brought in as *image-type volumes*, which means that the volume is based on a single MDisk. The CLI command that can be used is **migratevdisk**.

Example 4-6 shows the **migratevdisk** command that can be used to migrate an *image-type volume* to a *striped-type volume*, and can be used to migrate a *striped-type volume* to a *striped-type volume* as well.

Example 4-6 The migratevdisk command

```
IBM_2145:ITSO_DH8_B:superuser>svctask migratevdisk -mdiskgrp MDG1DS4K -threads 4  
-vdisk Migrate_sample
```

This command migrates the volume `Migrate_sample` to the storage pool `MDG1DS4K`, and uses four threads when migrating. Instead of using the volume name, you can use its ID number. For more information about this process, see *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.1*, SG24-7933.

You can monitor the migration process by using the **svcinfolsmigrate** command, as shown in Example 4-7.

Example 4-7 Monitoring the migration process

```
IBM_2145:ITSO_DH8_B:superuser>svcinfolsmigrate  
migrate_type MDisk_Group_Migration  
progress 0  
migrate_source_vdisk_index 3  
migrate_target_mdisk_grp 2  
max_thread_count 4  
migrate_source_vdisk_copy_id 0  
IBM_2145:ITSO_DH8_B:superuser>
```

4.5.2 Migrating to image-type volume

An *image-type volume* is a direct, “straight-through” mapping to one image mode MDisk. If a volume is migrated to another MDisk, the volume is represented as being in managed mode during the migration (because it is striped on two MDisks). It is only represented as an *image-type volume* after it reaches the state where it is a straight-through mapping. An image-type volume cannot be expanded.

Image-type disks are used to migrate existing data to an IBM Spectrum Virtualize or Storwize and to migrate data out of virtualization. In general, the reason for migrating a volume to an image type volume is to move the data on the disk to a nonvirtualized environment.

If the migration is interrupted by a cluster recovery, the migration resumes after the recovery completes.

The **migratetoimage** command migrates the data of a user-specified volume by consolidating its extents (which might be on one or more MDisks) onto the extents of the target MDisk that you specify. After migration is complete, the volume is classified as an image type volume, and the corresponding MDisk is classified as an image mode MDisk.

The managed disk that is specified as the target must be in an *unmanaged* state at the time that the command is run. Running this command results in the inclusion of the MDisk into the user-specified storage pool.

Remember: This command cannot be used if the source volume copy is in a child pool or if the target MDisk group that is specified is a child pool. This command does not work if the volume is fast formatting.

The **migratetoimage** command fails if the target or source volume is offline. Correct the offline condition before attempting to migrate the volume.

Remember: This command cannot be used on a volume that is owned by a file system or if the source MDisk is an SAS MDisk (which works in image mode only).

If the volume (or volume copy) is a target of a FlashCopy mapping with a source volume in an active-active relationship, the new managed disk group must be in the same site as the source volume. If the volume is in an active-active relationship, the new managed disk group must be located in the same site as the source volume. Additionally, the site information for the MDisk being added must be well-defined and match the site information for other MDisks in the storage pool.

Note: You cannot migrate data from a volume if the target volume's formatting attribute value is yes.

An encryption key cannot be used when migrating an image mode MDisk. To use encryption (when the MDisk has an encryption key), the MDisk must be self-encrypting.

IBM Spectrum Virtualize and Storwize **migratetoimage** command is useful when you want to use your system as a *data mover*. To better understand all requirements and specification for that command, see IBM Knowledge Center:

<https://ibm.biz/BdjJQm>

4.5.3 Migrating with volume mirroring

Volume mirroring offers the ability to migrate volumes between storage pools with different extent sizes. Complete the following steps to migrate volumes between storage pools:

1. Add a copy to the target storage pool.
2. Wait until the synchronization is complete.
3. Remove the copy in the source storage pool.

To migrate from a thin-provisioned volume to a fully allocated volume, the following steps are similar:

1. Add a target fully allocated copy.
2. Wait for synchronization to complete.
3. Remove the source thin-provisioned copy.

In both cases, if you set the **autodelete** option to *yes* when creating the volume copy, the source copy is automatically deleted, and you can skip the steps 3 mentioned previously. The preferred practice on this type of migration is to try not to overload the systems with a high *syncrate*, and not overload the system with too many migrations at the same time.

The **syncrate** parameter specifies the copy synchronization rate. A value of zero (0) prevents synchronization. The default value is 50. See Table 4-5 for the supported **-syncrate** values and their corresponding rates.

Table 4-5 Sample syncrate values

User-specified syncrate attribute value	Data copied/sec
1 - 10	128 KB
11 - 20	256 KB
21 - 30	512 KB
31 - 40	1 MB
41 - 50	2 MB
51 - 60	4 MB
61 - 70	8 MB
71 - 80	16 MB
81 - 90	32 MB
91 - 100	64 MB

For more information, see IBM Knowledge Center:

<https://ibm.biz/BdjJQb>

4.6 Preferred paths to a volume

For I/O purposes, IBM Spectrum Virtualize and Storwize nodes within the cluster are grouped into pairs, which are called *I/O groups* (sometimes cache I/O groups). A single pair is responsible for serving I/O on a specific volume. One node within the I/O group represents the preferred path for I/O to a specific volume. The other node represents the nonpreferred path. This preference alternates between nodes as each volume is created within an I/O group, to balance the workload evenly between the two nodes.

IBM Spectrum Virtualize and Storwize implement the concept of each volume having a preferred owner node, which improves cache efficiency and cache usage. The cache component read/write algorithms depend on one node that owns all the blocks for a specific track. The preferred node is set at the time of volume creation manually by the user or automatically by IBM Spectrum Virtualize and Storwize.

Because read-miss performance is better when the host issues a read request to the owning node, you want the host to know which node owns a track. The SCSI command set provides a mechanism for determining a preferred path to a specific volume. Because a track is part of a volume, the cache component distributes ownership by volume. The preferred paths are then all the paths through the owning node. Therefore, a *preferred path* is any port on a preferred controller, assuming that the SAN zoning is correct.

Tip: Performance can be better if the access is made on the preferred node. The data can still be accessed by the partner node in the I/O group if a failure occurs.

By default, IBM Spectrum Virtualize and Storwize assign ownership of even-numbered volumes to one node of a caching pair and the ownership of odd-numbered volumes to the other node. It is possible for the ownership distribution in a caching pair to become unbalanced if volume sizes are different between the nodes or if the volume numbers that are assigned to the caching pair are predominantly even or odd.

To provide flexibility in making plans to avoid this problem, the ownership for a specific volume can be explicitly assigned to a specific node when the volume is created. A node that is explicitly assigned as an owner of a volume is known as the *preferred node*. If a node becomes overloaded, the preferred node of a volume can be changed to the other node in the same I/O group, or to a node in another I/O group. This procedure can be performed concurrently with I/O operations if the host supports non-disruptive volume move.

For more information, see 4.7, “Changing the preferred node within or across I/O groups” on page 127.

IBM multipathing software (SDDPCM or SDDDSM) on hosts are aware of the preferred paths that IBM Spectrum Virtualize or Storwize sets per volume. They use algorithms to select paths and balance the load across them. In cases where all paths to preferred and non-preferred nodes are all available, the host will perform I/O operations using the paths to the preferred node. If all paths to preferred node become unavailable, the multipath software will make the host use the non-preferred paths. If all paths become unavailable, the host will set the device offline.

For more details, see 6.3, “Host pathing” on page 254.

Sometimes when debugging performance problems, it can be useful to look at the Non-Preferred Node Usage Percentage metric in IBM Spectrum Control. I/O to the non-preferred node might cause performance problems for the I/O group. This metric identifies any usage of non-preferred nodes to the user.

For more information about this metric and more, see IBM Spectrum Control in IBM Knowledge Center:

<https://ibm.biz/BdjJtq>

4.7 Changing the preferred node within or across I/O groups

The change of preferred node of a volume either within an I/O group or to another I/O group is a nondisruptive process.

Changing the preferred node within an I/O group can be done with concurrent I/O. However, it can lead to some delay in performance and, in case of some specific operating systems or applications, they could detect some time outs.

Changing the preferred node within an I/O group can be done by using both CLI and GUI, but if you have only one I/O group, this is not possible using the GUI. To change the preferred node within an I/O group using CLI, use the command `movevdisk -node <node_id or node_name> <vdisk_id or vdisk_name>`.

There are some limitations to change the preferred node across I/O groups, which is named Non-Disruptive Volume Move (NDVM). These limitations are mostly in Host Cluster environments, and you can find them in the *IBM System Storage Interoperation Center (SSIC)* website:

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

Attention: These migration tasks can be nondisruptive if performed correctly and the hosts that are mapped to the volume support NDVM. The cached data that is held within the system must first be written to disk before the allocation of the volume can be changed.

Modifying the I/O group that services the volume can be done concurrently with I/O operations if the host supports NDVM. This process also requires a rescan at the host level to ensure that the multipathing driver is notified that the allocation of the preferred node changed and the ports by which the volume is accessed changed. This can be done when one pair of nodes becomes over-used. If there is any host mapping for the volume, the host must be member of the target I/O group or the migration will fail.

Ensure that you create paths to I/O groups on the host system. After the system has successfully added the new I/O group to the volume's access set and you have moved the selected volumes to another I/O group, detect the new paths to the volumes on the host. The commands and actions on the host vary depending on the type of host and the connection method that is used. These steps must be completed on all hosts to which the selected volumes are currently mapped.

To move a volume between I/O groups by using the CLI, complete the steps listed in IBM Knowledge Center for IBM Spectrum Virtualize:

<https://ibm.biz/BdjJQx>

4.8 Volume throttling

Volume throttling effectively throttles the number of I/O operations per second (IOPS) or bandwidth (MBps) that can be achieved to and from a specific volume. You might want to use I/O throttling if you have a volume that has an access pattern that adversely affects the performance of other volumes. For example, volumes that are used for backup or archive operations can have I/O intensive workloads, potentially taking bandwidth from production volumes. Volume throttle can be used to limit I/Os for these types volumes so that I/O operations for production volumes are not affected.

When deciding between using IOPS or bandwidth as the I/O governing throttle, consider the disk access pattern of the application. Database applications often issue large amounts of I/O, but they transfer only a relatively small amount of data. In this case, setting an I/O governing throttle that is based on MBps does not achieve the expected result. It would be better to set an IOPS limit.

On the other hand, a streaming video application often issues a small amount of I/O, but it transfers large amounts of data. In contrast to the database example, defining an I/O throttle based in IOPS does not achieve a good result. For a streaming video application, it would be better to set an MBps limit.

You can set either IOPS or bandwidth limit, and starting with V7.7, you can set both limits for a volume. Standard SCSI read and write operations are throttled. Also, throttle is enforced on node level. For example, if a throttle limit is set for a volume at 100 IOPS, each node on the system that has access to the volume allows 100 IOPS for that volume. Any I/O operation that exceeds the throttle limit are queued at the receiving nodes.

4.8.1 Managing throttles for volumes

As mentioned previously, with Volume throttling, the IOPS limit, the bandwidth limit, or both, can be set for a volume.

Throttling at a volume level can be set by using the two commands below:

- ▶ **mkthrottle**: To set I/O throttles for volumes using this command, it must be used with **-type vdisk** parameter, followed by **-bandwidth bandwidth_limit_in_mbdisk** and/or **-iops iops_limit** to define MBps and IOPS limits.
- ▶ **chvdisk**: This command used with **-rate throttle_rate** parameter specifies the IOPS and MBps limits. The default throttle_rate units are I/Os. To change the **throttle_rate** units to megabits per second (MBps), specify the **-unitmb** parameter. If **throttle_rate** value is zero, the throttle rate is disabled. By default, the **throttle_rate** parameter is disabled.

Note: The command **mkthrottle** is used not to create throttles for volumes only, but also for hosts, host clusters, pools, and system offload.

When the IOPS limit is configured on a volume, and it is smaller than 100 IOPS, the throttling logic rounds it to 100 IOPS. Even if throttle is set to a value smaller than 100 IOPS, the actual throttling occurs at 100 IOPS.

After using any of the commands shown previously to set volume throttling, a throttle object is created. Then, you can list your created throttle objects by using the **lsthrottle** command, and change their parameters with the **chthrottle** command. Example 4-8 shows some command examples.

Example 4-8 Throttle command example

```
IBM_2145:ITS0_DH8_A:superuser>mkthrottle -type vdisk -bandwidth 100 -vdisk
testvol10
Throttle, id [0], successfully created.
IBM_2145:ITS0_DH8_A:superuser>lsthrottle
throttle_id throttle_name object_id object_name throttle_type IOPs_limit
bandwidth_limit_MB
0          throttle0      25          testvol10   vdisk          100

IBM_2145:ITS0_DH8_A:superuser>chthrottle -iops 1000 throttle0

IBM_2145:ITS0_DH8_A:superuser>lsthrottle
throttle_id throttle_name object_id object_name throttle_type IOPs_limit
bandwidth_limit_MB
0          throttle0      25          testvol10   vdisk          1000      100
```

```
IBM_2145:ITSO_DH8_A:superuser>lsthrottle throttle0
id 0
throttle_name throttle0
object_id 25
object_name testvol10
throttle_type vdisk
IOPs_limit 1000
bandwidth_limit_MB 100
```

For more information, and the procedure to set volume throttling, see IBM Knowledge Center:

<https://ibm.biz/BdjAHP>

4.9 Volume cache mode

Cache in IBM Spectrum Virtualize and Storwize can be set at a single volume granularity. For each volume, the cache can be *readwrite*, *readonly*, or *none*. The meaning of each parameter is self-explanatory. By default, when a volume is created, the cache mode is set to *readwrite*.

In most cases, the volume with readwrite cache mode is recommended, because disabling cache for a volume can result in performance issues to the host. But, there are some specific scenarios that it is recommended to disable the readwrite cache.

You use cache-disabled (*none*) volumes primarily when you have remote copy or FlashCopy in the underlying storage controller, and these volumes are virtualized in IBM Spectrum Virtualize or Storwize devices as image vdisks. You might want to use cache-disabled volumes where intellectual capital is in existing copy services automation scripts. Keep the use of cache-disabled volumes to minimum for normal workloads.

You can also use cache-disabled volumes to control the allocation of cache resources. By disabling the cache for certain volumes, more cache resources are available to cache I/Os to other volumes in the same I/O group. This technique of using cache-disabled volumes is effective where an I/O group serves volumes that benefit from cache and other volumes, where the benefits of caching are small or nonexistent.

Also, a case in which you can use cache-disabled volumes is in a scenario of an application which requires very low response time and uses IBM Spectrum Virtualize volumes which are in mdisks from all-flash storage. If this application generates so many IOPS and requires very low response time, disabling the cache of the volumes in IBM Spectrum Virtualize would take advantage of the all-flash performance capabilities, and consume less resources of the Spectrum Virtualize device.

More details of these cases are shown in the next sections.

4.9.1 Changing the cache mode of a volume

The cache mode of a volume can be concurrently changed (with I/O) by using the **svctask chvdisk** command. This command will not fail I/O to the user, and the command must be allowed to run on any volume. If used correctly without the **-force** flag, the command will not result in a corrupted volume. Therefore, the cache must be flushed and you must discard cache data if the user disables cache on a volume.

Example 4-9 shows an image volume VDISK_IMAGE_1 that changed the cache parameter after it was created.

Example 4-9 Changing the cache mode of a volume

```
IBM_2145:svccg8:admin>svctask mkvdisk -name VDISK_IMAGE_1 -iogrp 0 -mdiskgrp
IMAGE_Test -vtype image -mdisk D8K_L3331_1108
Virtual Disk, id [9], successfully created
IBM_2145:svccg8:admin>svcinfolsvdisk VDISK_IMAGE_1
id 9
.
lines removed for brevity
.
fast_write_state empty
cache readwrite
.
lines removed for brevity

IBM_2145:svccg8:admin>svctask chvdisk -cache none VDISK_IMAGE_1
IBM_2145:svccg8:admin>svcinfolsvdisk VDISK_IMAGE_1
id 9
.
lines removed for brevity
.
cache none
.
lines removed for brevity
```

Tip: By default, the volumes are created with cache mode enabled (read/write), but you can specify the cache mode when the volume is created by using the **-cache** option.

4.9.2 Underlying controller remote copy with IBM Spectrum Virtualize and Storwize cache-disabled volumes

When synchronous or asynchronous remote copy is used in the underlying storage controller, you must map the controller logical unit numbers (LUNs) at the source and destination through IBM Spectrum Virtualize and Storwize as image mode disks. IBM Spectrum Virtualize and Storwize cache must be disabled.

You can access the source or the target of the remote copy from a host directly, rather than through IBM Spectrum Virtualize and Storwize. You can use IBM Spectrum Virtualize and Storwize copy services with the image mode volume that represents the primary site of the controller remote copy relationship.

Do not use IBM Spectrum Virtualize and Storwize copy services with the volume at the secondary site because IBM Spectrum Virtualize and Storwize do not detect the data that is flowing to this LUN through the controller.

Figure 4-4 shows the relationships between IBM Spectrum Virtualize and Storwize, the volume, and the underlying storage controller for a cache-disabled volume.

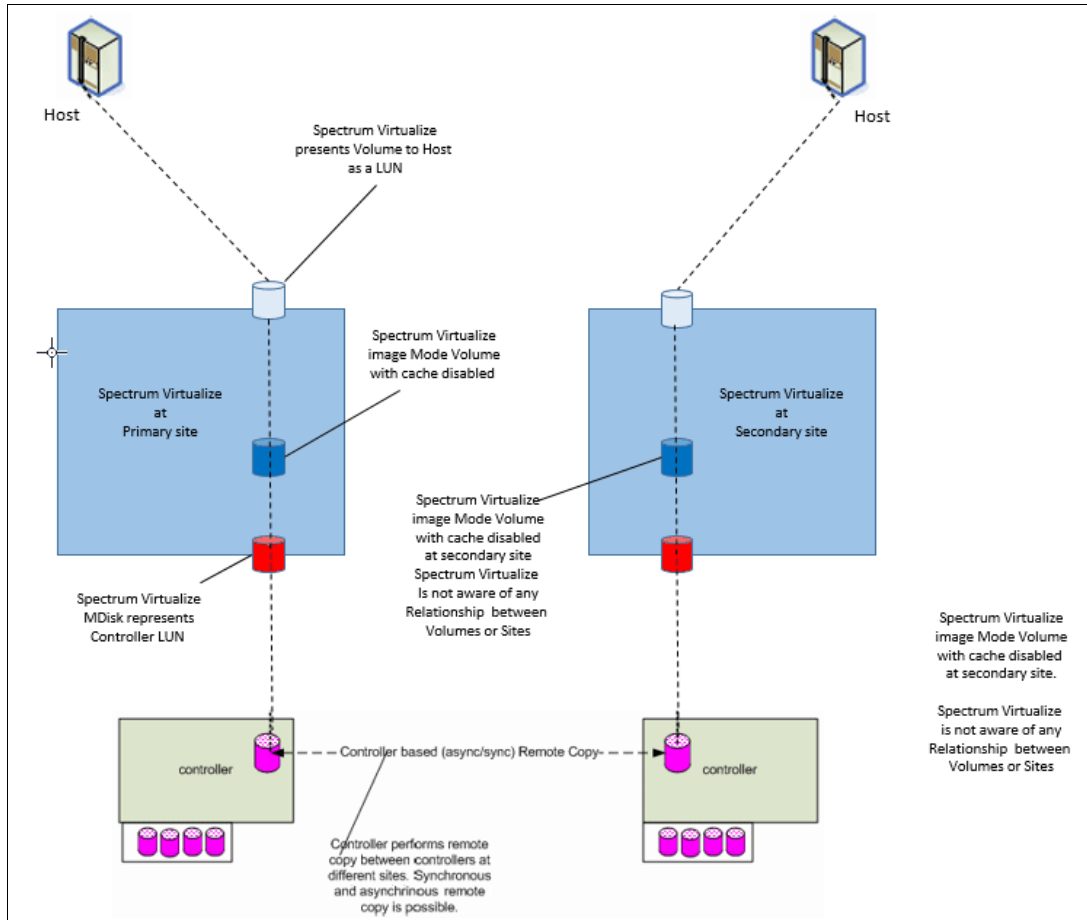


Figure 4-4 Cache-disabled volume in a remote copy relationship

4.9.3 Using underlying controller FlashCopy with IBM Spectrum Virtualize and Storwize cache-disabled volumes

When FlashCopy is used in the underlying storage controller, you must map the controller LUNs for the source and the target through IBM Spectrum Virtualize and Storwize as image mode disks, as shown in Figure 4-5. IBM Spectrum Virtualize and Storwize cache must be disabled. You can access the source or the target of the FlashCopy from a host directly rather than through IBM Spectrum Virtualize and Storwize.

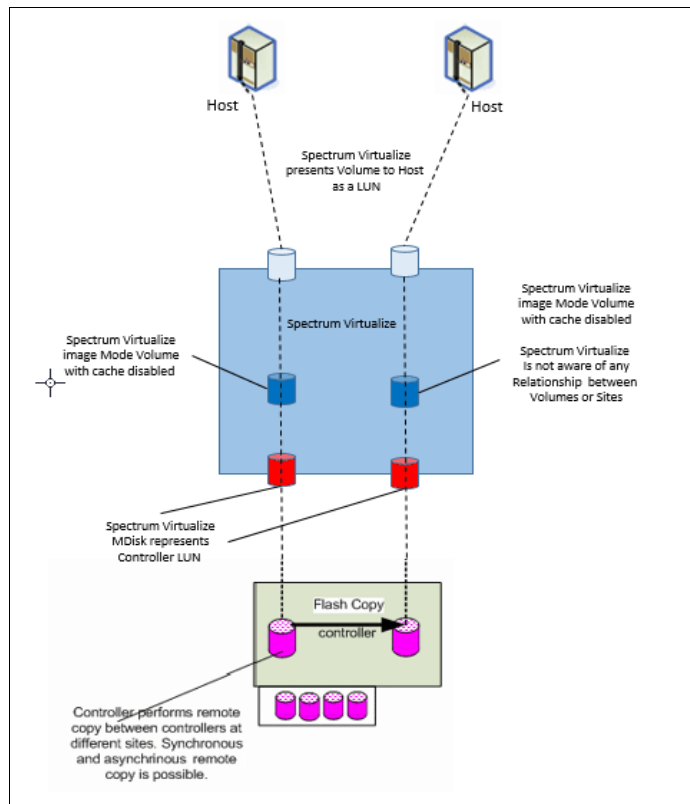


Figure 4-5 FlashCopy with cache-disabled volumes

4.9.4 Using IBM Spectrum Virtualize or Storwize with FlashSystem

There can be some specific scenarios where you want to virtualize IBM or OEM all-flash array (AFA) because you want to have specific performance for specific workloads. The MDisk supplied by those AFA will be encompassed in a dedicated storage pool where Volumes are configured.

In this scenario, perform some optimization on the IBM Spectrum Virtualize or Storwize Volumes cache depending on the infrastructure you are building.

Figure 4-6 shows write operation behavior when volume cache is activated (*readwrite*).

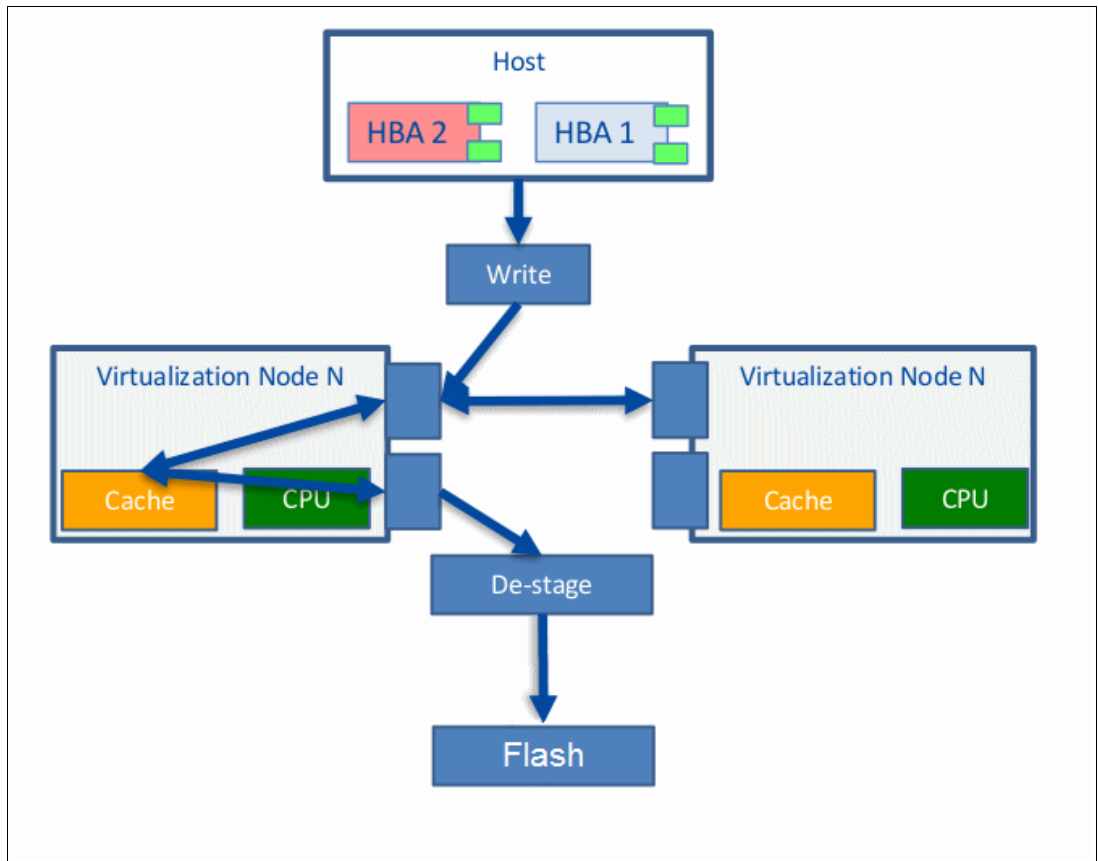


Figure 4-6 Cache activated

Figure 4-7 shows a write operation behavior when volume cache is deactivated (*none*).

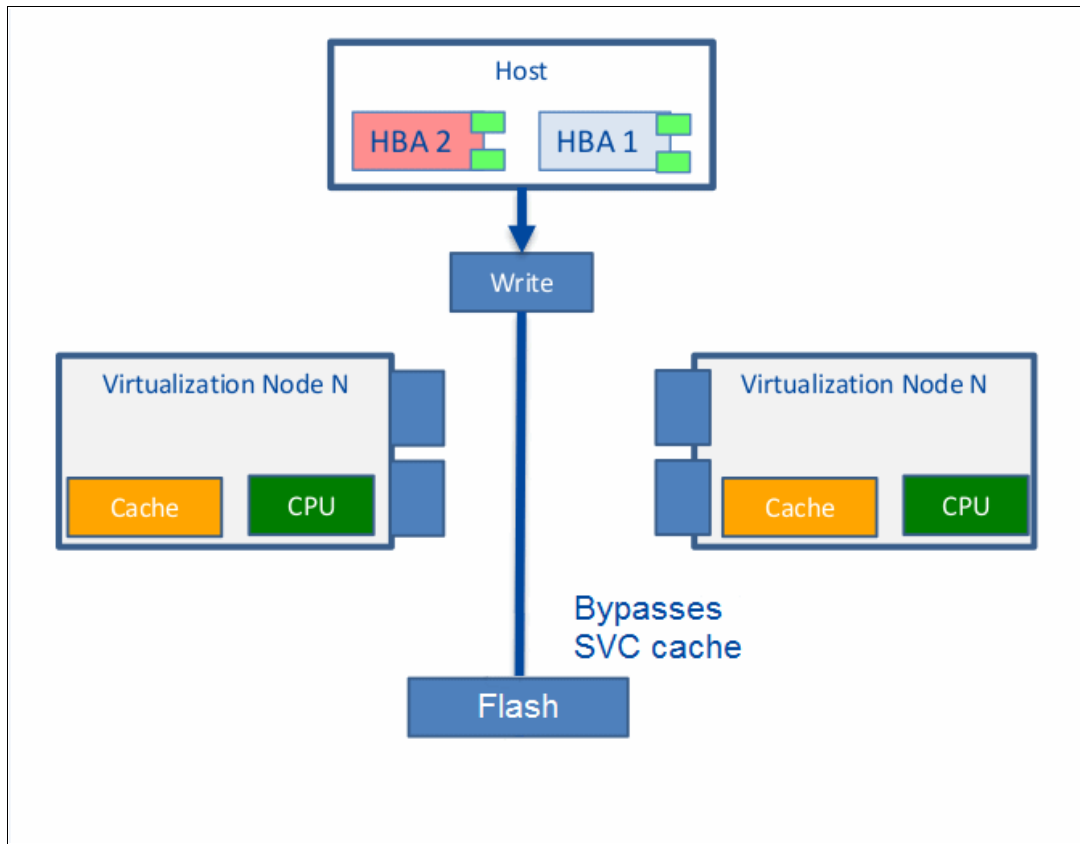


Figure 4-7 Cache deactivated

In this case, an environment with Copy Services (FlashCopy, Metro Mirror, Global Mirror, and Volume Mirroring) and typical workloads, disabling SVC cache is detrimental to overall performance. In cases where there are no advanced functions and extremely high IOPS rate is required, disabling the cache might help.

Attention: Carefully evaluate the impact to the entire system with quantitative analysis before and after making this change

4.10 VMware Virtual Volumes

IBM Spectrum Virtualize and VMware's Virtual Volumes (VVols) are paving the way towards a truly Software-Defined Environment. IBM Spectrum Virtualize is at the very core of Software-Defined Storage. The addition of Virtual Volumes enables a fundamentally more efficient operational model for storage in virtualized environments, centering it around the virtual machine (VM) rather than the physical infrastructure.

Without the use of Virtual Volumes, a virtual machine disk (VMDK) is presented to a VM in the form of a file. This file represents a disk to the VM. The VM is then accessed by the guest operating system in the same way as a physical disk is accessed by a physical server. This VMDK is stored on a VMware Virtual Machine File System (VMFS) formatted data store.

The VMFS data store is hosted by a single volume on a storage system such as IBM Spectrum Virtualize or Storwize. A single VMFS data store, sometimes referred to as the VMFS blender, can have hundreds or even thousands of VMDKs.

Virtual Volumes provide a one-to-one mapping between the VM's disks and the volumes (VVols) hosted by the storage system. This VVol is wholly owned by the VM. Exposing the VVol at the storage level enables storage-system-based operations at the granular VM level. For example, capabilities such as compression and encryption can be applied to an individual VM. Similarly, IBM FlashCopy can be used at the VVol level when performing snapshot and clone operations.

For more information about VVols prerequisites, implementation, and configuration in IBM Spectrum Virtualize or Storwize environments, see *Configuring VMware Virtual Volumes for Systems Powered by IBM Spectrum Virtualize*, SG24-8328, and *Quick-start Guide to Configuring VMware Virtual Volumes for Systems Powered by IBM Spectrum Virtualize*, REDP-5321.

4.11 Additional considerations

The following section describes additional and brief considerations regarding volumes.

4.11.1 Volume protection

From V7.4 onwards, consider enabling volume protection by using `chsystem vdiskprotectionenabled yes -vdiskprotectiontime <value_in_minutes>`. Volume protection ensures that some CLI actions (most of those that either explicitly or implicitly remove host-volume mappings or delete volumes) are policed to prevent the removal of mappings to volumes or deletion of volumes that are considered active (the system has detected I/O activity within the specified time in minutes to the volume from any host).

Note: Volume protection cannot be overridden by the use of the `-force` flag in the affected CLI commands. Volume protection must be disabled to permit an activity that is currently blocked.

4.11.2 Volume resize

Fully allocated and thin-provisioned volumes can have their sizes increased or decreased. A volume can be expanded with concurrent I/Os for some operating systems, but you should never attempt to shrink a volume in use that contains data, because in IBM Spectrum Virtualize, volume capacity is removed from the end of the disk, whether or not that capacity is in use by a server. Remember that a volume cannot be expanded or shrunk during its quick initialization process.

Expanding a volume

You can expand volumes for the following reasons:

- ▶ To increase the available capacity on a particular volume that is already mapped to a host.
- ▶ To increase the size of a volume aiming to make it match the size of the source or master volume so that it can be used in a FlashCopy mapping or Metro Mirror relationship. In this case, there are some considerations to be taken, which are shown in Chapter 5, “Copy services” on page 139.

The CLI or the GUI can be used to expand a volume concurrently with I/O operations on Windows, AIX, or Linux systems.

Note: You should see operating system documentation to check supported versions and requirements to have an IBM Spectrum Virtualize volume expanded with concurrent I/O.

Use the command `expandvdisksize -size <size_change> -unit <b | kb | mb | gb | tb | pb> <vdisk_name>` to expand a volume size.

Shrinking a volume

Volumes can be reduced in size if necessary. If a volume does not contain any data, there should be no issues to shrink its size.

However, if a volume is in use and contains data, do not shrink its size, because IBM Spectrum Virtualize will not be aware if it is removing used or non-used capacity.

Attention: It is difficult to anticipate how an operating system or file system uses the capacity in a volume. When you shrink a volume, capacity is removed from the end of the disk, whether or not that capacity is in use. Even if a volume has free capacity, do not assume that only unused capacity is removed when you shrink a volume.

If the volume contains data that is being used, do not attempt under any circumstances to shrink a volume without first backing up your data.

The command to shrink a volume is `shrinkvdisksize -size <size_change> -unit <b | kb | mb | gb | tb | pb> <vdisk_name>`.



Copy services

Copy services are a collection of functions that provide capabilities for disaster recovery, data migration, and data duplication solutions. This chapter provides an overview and the preferred practices of IBM Spectrum Virtualize and Storwize family copy services capabilities, including FlashCopy, Metro Mirror and Global Mirror, and Volume Mirroring.

This chapter includes the following sections:

- ▶ Introduction to copy services
- ▶ FlashCopy
- ▶ Remote Copy services
- ▶ Native IP replication
- ▶ Volume Mirroring

5.1 Introduction to copy services

IBM Spectrum Virtualize and Storwize family products offer a complete set of copy services functions that provide capabilities for disaster recovery, business continuity, data movement, and data duplication solutions.

5.1.1 FlashCopy

FlashCopy is a function that allows you to create a point-in-time copy of one of your volumes. This function might be helpful when performing backups or application testing. These copies can be cascaded on one another, read from, written to, and even reversed. These copies are able to conserve storage, if needed, by being space-efficient copies that only record items that have changed from the originals instead of full copies.

5.1.2 Metro Mirror and Global Mirror

Metro Mirror and Global Mirror are technologies that enable you to keep a real-time copy of a volume at a remote site that contains another IBM Spectrum Virtualize or Storwize system:

- ▶ Metro Mirror makes *synchronous* copies, which means that the original writes are not considered complete until the write to the destination disk has been confirmed. The distance between your two sites is usually determined by how much latency your applications can handle.
- ▶ Global Mirror makes *asynchronous* copies of your disk. This fact means that the write is considered complete after it is complete at the local disk. It does not wait for the write to be confirmed at the remote system as Metro Mirror does. This requirement greatly reduces the latency experienced by your applications if the other system is far away. However, it also means that during a failure, the data on the remote copy might not have the most recent changes committed to the local disk.

5.1.3 Global Mirror with Change Volumes

This function (also known as Cycle-Mode Global Mirror), introduced in V6.3, can best be described as “Continuous Remote FlashCopy.” If you use this feature, the system takes periodic FlashCopies of a disk and write them to your remote destination.

This feature completely isolates the local copy from wide area network (WAN) issues and from sudden spikes in workload that might occur. The drawback is that your remote copy might lag behind the original by a significant amount, depending on how you have set up the cycle time.

5.1.4 Volume Mirroring function

Volume Mirroring is a function that is designed to increase high availability of the storage infrastructure. It provides the ability to create up to two local copies of a volume. Volume Mirroring can use space from two storage pools, and preferably from two separate back-end disk subsystems.

Primarily, you use this function to insulate hosts from the failure of a storage pool and also from the failure of a back-end disk subsystem. During a storage pool failure, the system continues to provide service for the volume from the other copy on the other storage pool, with no disruption to the host.

You can also use Volume Mirroring to migrate from a thin-provisioned volume to a non-thin-provisioned volume, and to migrate data between storage pools of different extent sizes.

5.2 FlashCopy

By using the IBM FlashCopy function of the IBM Spectrum Virtualize and Storwize systems, you can perform a *point-in-time copy* of one or more volumes. This section describes the inner workings of FlashCopy, and provides some preferred practices for its use.

You can use FlashCopy to help you solve critical and challenging business needs that require duplication of data of your source volume. Volumes can remain online and active while you create consistent copies of the data sets. Because the copy is performed at the block level, it operates below the host operating system and its cache. Therefore, the copy is not apparent to the host.

Important: Because FlashCopy operates at the block level below the host operating system and cache, those levels do need to be flushed for consistent FlashCopies.

While the FlashCopy operation is performed, the source volume is stopped briefly to initialize the FlashCopy bitmap, and then input/output (I/O) can resume. Although several FlashCopy options require the data to be copied from the source to the target in the background, which can take time to complete, the resulting data on the target volume is presented so that the copy appears to complete immediately.

This process is performed by using a bitmap (or bit array) that tracks changes to the data after the FlashCopy is started, and an indirection layer that enables data to be read from the source volume transparently.

5.2.1 FlashCopy use cases

When you are deciding whether FlashCopy addresses your needs, you must adopt a combined business and technical view of the problems that you want to solve. First, determine the needs from a business perspective. Then, determine whether FlashCopy can address the technical needs of those business requirements.

The business applications for FlashCopy are wide-ranging. In the following sections, a short description of the most common use cases is provided.

Backup improvements with FlashCopy

FlashCopy does not reduce the time that it takes to perform a backup to traditional backup infrastructure. However, it can be used to minimize and, under certain conditions, eliminate application downtime that is associated with performing backups. FlashCopy can also transfer the resource usage of performing intensive backups from production systems.

After the FlashCopy is performed, the resulting image of the data can be backed up to tape as though it were the source system. After the copy to tape is complete, the image data is redundant and the target volumes can be discarded. For time-limited applications, such as these examples, “no copy” or incremental FlashCopy is used most often. The use of these methods puts less load on your infrastructure.

When FlashCopy is used for backup purposes, the target data usually is managed as read-only at the operating system level. This approach provides extra security by ensuring that your target data was not modified and remains true to the source.

Restore with FlashCopy

FlashCopy can perform a restore from any existing FlashCopy mapping. Therefore, you can restore (or copy) from the target to the source of your regular FlashCopy relationships. It might be easier to think of this method as reversing the direction of the FlashCopy mappings. This capability has the following benefits:

- ▶ There is no need to worry about pairing mistakes because you trigger a restore.
- ▶ The process appears instantaneous.
- ▶ You can maintain a pristine image of your data while you are restoring what was the primary data.

This approach can be used for various applications, such as recovering your production database application after an errant batch process that caused extensive damage.

Preferred practices: Although restoring from a FlashCopy is quicker than a traditional tape media restore, do not use restoring from a FlashCopy as a substitute for good archiving practices. Instead, keep one to several iterations of your FlashCopies so that you can near-instantly recover your data from the most recent history. Keep your long-term archive as appropriate for your business.

In addition to the restore option, which copies the original blocks from the target volume to modified blocks on the source volume, the target can be used to perform a restore of individual files. To do that, you must make the target available on a host. Do not make the target available to the source host, because seeing duplicates of disks causes problems for most host operating systems. Copy the files to the source by using the normal host data copy methods for your environment.

Moving and migrating data with FlashCopy

FlashCopy can be used to facilitate the movement or migration of data between hosts while minimizing downtime for applications. By using FlashCopy, application data can be copied from source volumes to new target volumes while applications remain online. After the volumes are fully copied and synchronized, the application can be brought down and then immediately brought back up on the new server that is accessing the new FlashCopy target volumes.

This method differs from the other migration methods, which are described later in this chapter. Common uses for this capability are host and back-end storage hardware refreshes.

Application testing with FlashCopy

It is often important to test a new version of an application or operating system that is using actual production data. This testing ensures the highest quality possible for your environment. FlashCopy makes this type of testing easy to accomplish without putting the production data at risk or requiring downtime to create a constant copy.

Create a FlashCopy of your source and use that for your testing. This copy is a duplicate of your production data down to the block level so that even physical disk identifiers are copied. Therefore, it is impossible for your applications to tell the difference.

5.2.2 FlashCopy capabilities overview

FlashCopy occurs between a source volume and a target volume in the same storage system. The minimum granularity that IBM Spectrum Virtualize and Storwize systems support for FlashCopy is an entire volume. It is not possible to use FlashCopy to copy only part of a volume.

To start a FlashCopy operation, a relationship between the source and the target volume must be defined. This relationship is called *FlashCopy Mapping*.

FlashCopy mappings can be stand-alone or a member of a Consistency Group. You can perform the actions of preparing, starting, or stopping FlashCopy on either a stand-alone mapping or a Consistency Group.

Figure 5-1 shows the concept of FlashCopy mapping.

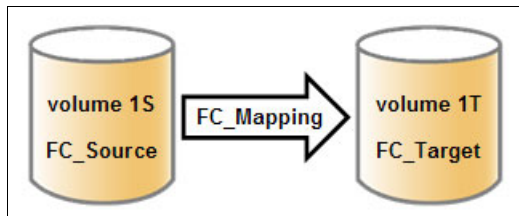


Figure 5-1 FlashCopy mapping

A FlashCopy mapping has a set of attributes and settings that define the characteristics and the capabilities of the FlashCopy.

These characteristics are explained in more detail in the following sections.

Background copy

The *background copy rate* is a property of a FlashCopy mapping that allows to specify whether a background physical copy of the source volume to the corresponding target volume occurs. A value of 0 disables the background copy. If the FlashCopy background copy is disabled, only data that has changed on the source volume is copied to the target volume. A FlashCopy with background copy disabled is also known as *No-Copy* FlashCopy.

The benefit of using a FlashCopy mapping with background copy enabled is that the target volume becomes a real clone (independent from the source volume) of the FlashCopy mapping source volume after the copy is complete. When the background copy function is not performed, the target volume remains a valid copy of the source data while the FlashCopy mapping remains in place.

Prior to V7.8.1, valid values for the background copy rate were 0 - 100. IBM Spectrum Virtualize V7.8.1 increased the background copy rate up to 150. The background copy rate can be defined and changed dynamically for individual FlashCopy mappings.

Table 5-1 shows the relationship of the background copy rate value to the attempted amount of data to be copied per second.

Table 5-1 Relationship between the rate and data rate per second

Value	Data copied per second
1 - 10	128 KB
11 - 20	256 KB
21 - 30	512 KB
31 - 40	1 MB
41 - 50	2 MB
51 - 60	4 MB
61 - 70	8 MB
71 - 80	16 MB
81 - 90	32 MB
91 - 100	64 MB
101-110	128 MB
111-120	256 MB
121-130	512 MB
131-140	1024 MB
141-150	2048 MB

Note: A background copy rate value from 101 - 150 can be specified only with IBM Spectrum Virtualize V7.8.1 or later. To ensure optimal performance of all IBM Spectrum Virtualize/Storwize features, it is advised not to exceed a copyrate value of 130.

FlashCopy Consistency Groups

Consistency Groups can be used to help create a consistent point-in-time copy across multiple volumes. They are used to manage the consistency of dependent writes that are run in the application following the correct sequence.

When Consistency Groups are used, the FlashCopy commands are issued to the Consistency Groups. The groups perform the operation on all FlashCopy mappings contained within the Consistency Groups at the same time.

Figure 5-2 illustrates a Consistency Group consisting of two volume mappings.

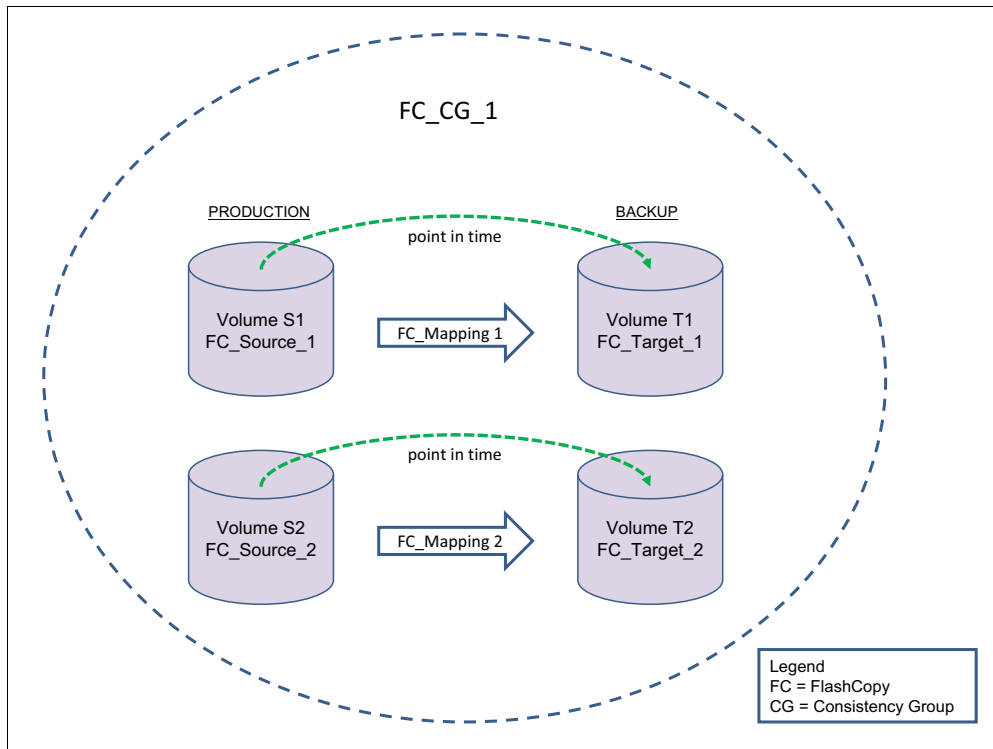


Figure 5-2 Multiple volumes mapping in a Consistency Group

FlashCopy mapping considerations: If the FlashCopy mapping has been added to a Consistency Group, it can only be managed as part of the group. This limitation means that FlashCopy operations are no longer allowed on the individual FlashCopy mappings.

Incremental FlashCopy

Using Incremental FlashCopy, you can reduce the required time of copy. Also, because less data must be copied, the workload put on the system and the back-end storage is reduced.

Basically, Incremental FlashCopy does not require that you copy an entire disk source volume every time the FlashCopy mapping is started. It means that only the changed regions on source volumes are copied to target volumes, as shown in Figure 5-3.

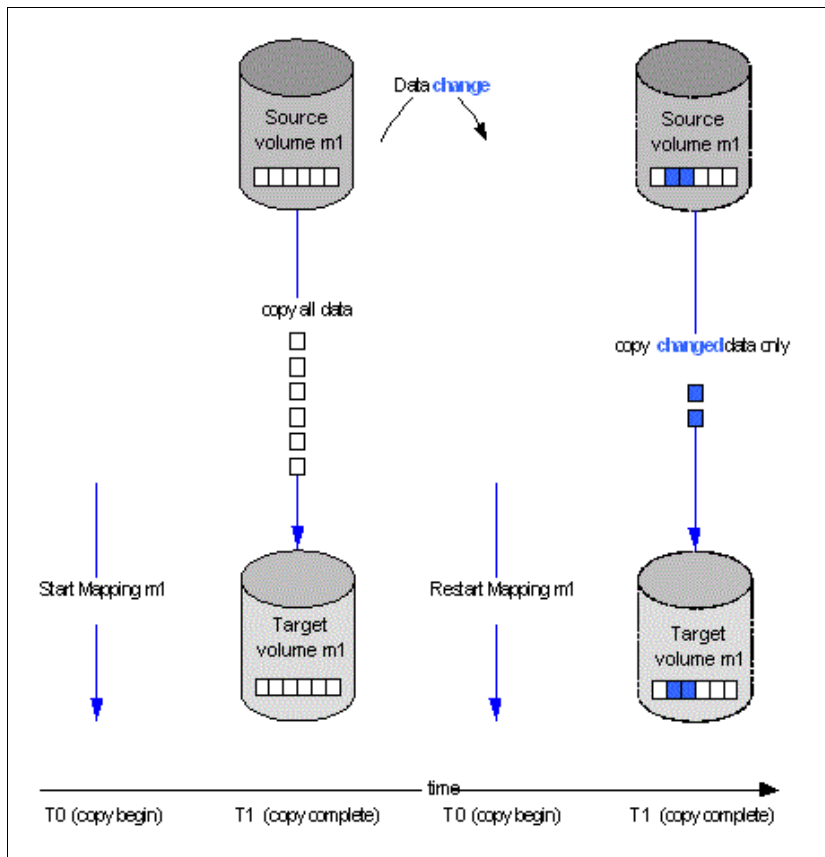


Figure 5-3 Incremental FlashCopy

If the FlashCopy mapping was stopped before the background copy completed, then when the mapping is restarted, the data that was copied before the mapping was stopped will not be copied again. For example, if an incremental mapping reaches 10 percent progress when it is stopped and then it is restarted, that 10 percent of data will not be recopied when the mapping is restarted, assuming that it was not changed.

Stopping an incremental FlashCopy mapping: If you are planning to stop an incremental FlashCopy mapping, make sure that the copied data on the source volume will not be changed, if possible. Otherwise, you might have an inconsistent point-in-time copy.

A “difference” value is provided in the query of a mapping, which makes it possible to know how much data has changed. This data must be copied when the Incremental FlashCopy mapping is restarted. The difference value is the percentage (0-100 percent) of data that has been changed. This data must be copied to the target volume to get a fully independent copy of the source volume.

An incremental FlashCopy can be defined setting the *incremental* attribute in the FlashCopy mapping.

Multiple Target FlashCopy

In Multiple Target FlashCopy, a source volume can be used in multiple FlashCopy mappings, while the target is a different volume, as shown in Figure 5-4.

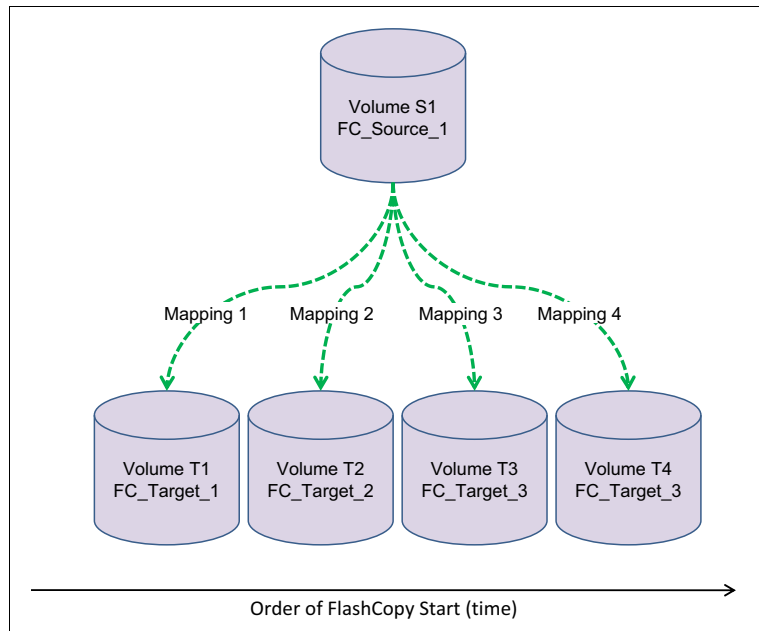


Figure 5-4 Multiple Target FlashCopy

Up to 256 different mappings are possible for each source volume. These mappings are independently controllable from each other. Multiple Target FlashCopy mappings can be members of the same or different Consistency Groups. In cases where all the mappings are in the same Consistency Group, the result of starting the Consistency Group will be to FlashCopy to multiple identical target volumes.

Cascaded FlashCopy

With Cascaded FlashCopy, you can have a source volume for one FlashCopy mapping and as the target for another FlashCopy mapping; this is referred to as a *Cascaded FlashCopy*. This function is illustrated in Figure 5-5.

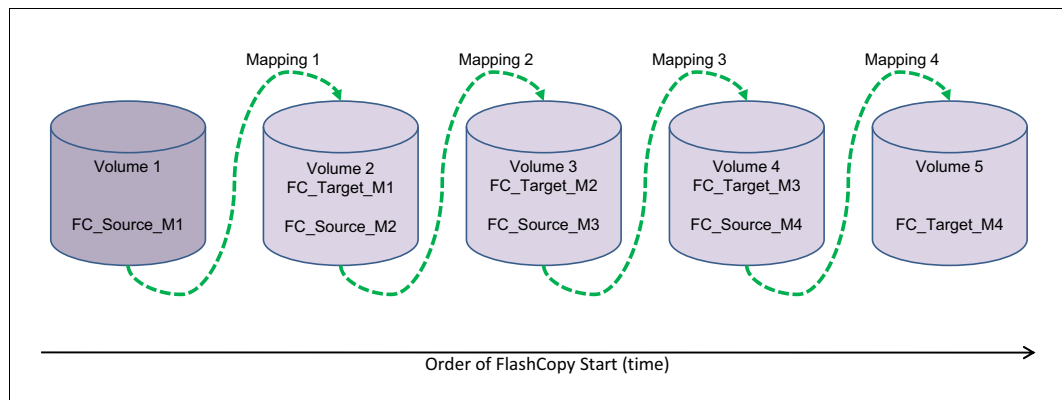


Figure 5-5 Cascaded FlashCopy

A total of 255 mappings are possible for each cascade.

Thin-provisioned FlashCopy

When a new volume is created, you can designate it as a *thin-provisioned volume*, and it has a virtual capacity and a real capacity.

Virtual capacity is the volume storage capacity that is available to a host. *Real capacity* is the storage capacity that is allocated to a volume copy from a storage pool. In a fully allocated volume, the virtual capacity and real capacity are the same. However, in a thin-provisioned volume, the virtual capacity can be much larger than the real capacity.

The virtual capacity of a thin-provisioned volume is typically larger than its real capacity. On IBM Spectrum Virtualize and Storwize systems, the real capacity is used to store data that is written to the volume, and metadata that describes the thin-provisioned configuration of the volume. As more information is written to the volume, more of the real capacity is used.

Thin-provisioned volumes can also help to simplify server administration. Instead of assigning a volume with some capacity to an application and increasing that capacity following the needs of the application if those needs change, you can configure a volume with a large virtual capacity for the application. You can then increase or shrink the real capacity as the application needs change, without disrupting the application or server.

When you configure a thin-provisioned volume, you can use the warning level attribute to generate a warning event when the used real capacity exceeds a specified amount or percentage of the total real capacity. For example, if you have a volume with 10 GB of total capacity and you set the warning to 80 percent, an event is registered in the event log when you use 80 percent of the total capacity. This technique is useful when you need to control how much of the volume is used.

If a thin-provisioned volume does not have enough real capacity for a write operation, the volume is taken offline and an error is logged (error code 1865, event ID 060001). Access to the thin-provisioned volume is restored by either increasing the real capacity of the volume or increasing the size of the storage pool on which it is allocated.

You can use thin volumes for cascaded FlashCopy and multiple target FlashCopy. It is also possible to mix thin-provisioned with normal volumes. It can be used for incremental FlashCopy too, but using thin-provisioned volumes for incremental FlashCopy only makes sense if the source and target are thin-provisioned.

Thin-provisioned incremental FlashCopy

The implementation of thin-provisioned volumes does not preclude the use of incremental FlashCopy on the same volumes. It does not make sense to have a fully allocated source volume and then use incremental FlashCopy, which is always a full copy at first, to copy this fully allocated source volume to a thin-provisioned target volume. However, this action is not prohibited.

Consider this optional configuration:

- ▶ A thin-provisioned source volume can be copied incrementally by using FlashCopy to a thin-provisioned target volume. Whenever the FlashCopy is performed, only data that has been modified is recopied to the target. Note that if space is allocated on the target because of I/O to the target volume, this space will not be reclaimed with subsequent FlashCopy operations.
- ▶ A fully allocated source volume can be copied incrementally using FlashCopy to another fully allocated volume at the same time as it is being copied to multiple thin-provisioned targets (taken at separate points in time). This combination allows a single full backup to be kept for recovery purposes, and separates the backup workload from the production workload. At the same time, it allows older thin-provisioned backups to be retained.

Reverse FlashCopy

Reverse FlashCopy enables FlashCopy targets to become restore points for the source without breaking the FlashCopy relationship, and without having to wait for the original copy operation to complete. Therefore, it supports multiple targets (up to 256) and multiple rollback points.

A key advantage of the Multiple Target Reverse FlashCopy function is that the reverse FlashCopy does not destroy the original target. This feature enables processes that are using the target, such as a tape backup, to continue uninterrupted.

IBM Spectrum Virtualize and Storwize family systems also allow you to create an optional copy of the source volume to be made before the reverse copy operation starts. This ability to restore back to the original source data can be useful for diagnostic purposes.

5.2.3 FlashCopy functional overview

Understanding how FlashCopy works internally helps you to configure it in a way that you want and enables you to obtain more benefits from it.

FlashCopy bitmaps and grains

A *bitmap* is an internal data structure stored in a particular I/O Group that is used to track which data in FlashCopy mappings has been copied from the source volume to the target volume. *Grains* are units of data grouped together to optimize the use of the bitmap. One bit in each bitmap represents the state of one grain. FlashCopy grain can be either 64 KB or 256 KB.

A FlashCopy bitmap takes up the bitmap space in the memory of the I/O group that must be shared with other features' bitmaps (such as Remote Copy bitmaps, Volume Mirroring bitmaps, and RAID bitmaps).

Indirection layer

The *FlashCopy indirection layer* governs the I/O to the source and target volumes when a FlashCopy mapping is started. This process is done by using a FlashCopy bitmap. The purpose of the FlashCopy indirection layer is to enable both the source and target volumes for read and write I/O immediately after FlashCopy starts.

The following description illustrates how the FlashCopy indirection layer works when a FlashCopy mapping is prepared and then started.

When a FlashCopy mapping is prepared and started, the following sequence is applied:

1. Flush the write cache to the source volume or volumes that are part of a Consistency Group.
2. Put the cache into write-through mode on the source volumes.
3. Discard the cache for the target volumes.
4. Establish a sync point on all of the source volumes in the Consistency Group (creating the FlashCopy bitmap).
5. Ensure that the indirection layer governs all of the I/O to the source volumes and target.
6. Enable the cache on source volumes and target volumes.

FlashCopy provides the semantics of a point-in-time copy that uses the indirection layer, which intercepts I/O that is directed at either the source or target volumes. The act of starting a FlashCopy mapping causes this indirection layer to become active in the I/O path, which occurs automatically across all FlashCopy mappings in the Consistency Group. The indirection layer then determines how each of the I/O is to be routed based on the following factors:

- ▶ The volume and the logical block address (LBA) to which the I/O is addressed
- ▶ Its direction (read or write)
- ▶ The state of an internal data structure, the FlashCopy bitmap

The indirection layer allows the I/O to go through the underlying volume. It redirects the I/O from the target volume to the source volume, or queues the I/O while it arranges for data to be copied from the source volume to the target volume. The process of queueing the write operations on the source volume while the indirection layer completes the grain copy on the target volume is called *copy-on-write*.

Table 5-2 summarizes the indirection layer algorithm.

Table 5-2 Summary table of the FlashCopy indirection layer algorithm

Volume being accessed	Has the grain been copied?	Host I/O operation	
		Read	Write
Source	No	Read from the source volume.	Copy grain to the most recently started target for this source, then write to the source.
	Yes	Read from the source volume.	Write to the source volume.
Target	No	If any newer targets exist for this source in which this grain has already been copied, read from the oldest of these targets. Otherwise, read from the source.	Hold the write. Check the dependency target volumes to see whether the grain has been copied. If the grain is not already copied to the next oldest target for this source, copy the grain to the next oldest target. Then, write to the target.
	Yes	Read from the target volume.	Write to the target volume.

Interaction with cache

Starting with V7.3, the entire cache subsystem was redesigned and changed. Cache has been divided into upper and lower cache. Upper cache serves mostly as write cache and hides the write latency from the hosts and application. Lower cache is a read/write cache and optimizes I/O to and from disks.

Figure 5-6 shows the IBM Spectrum Virtualize cache architecture.

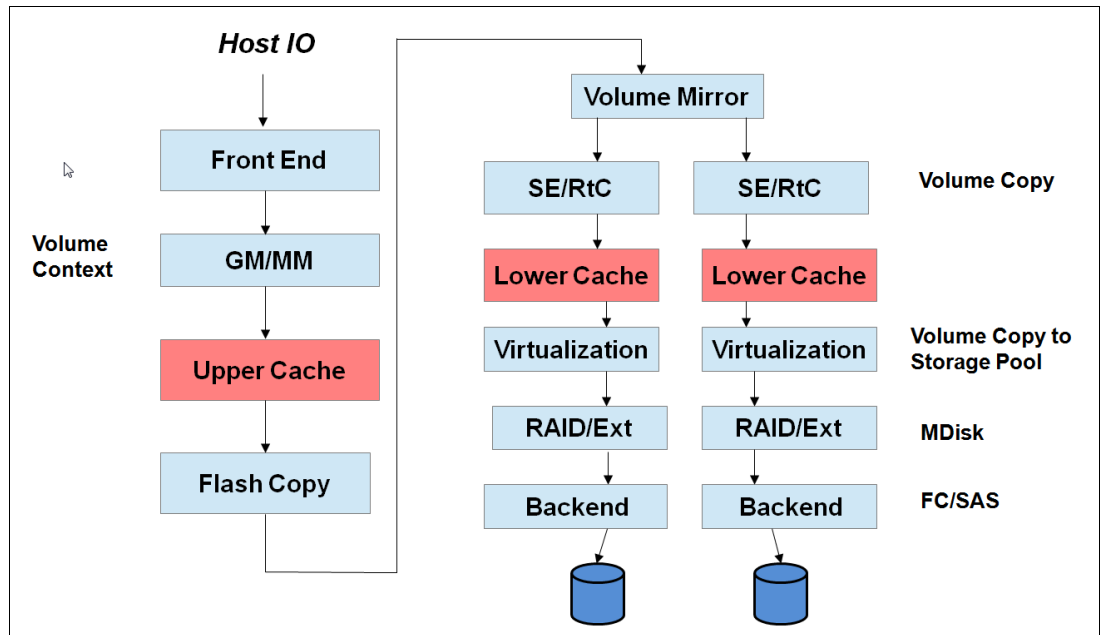


Figure 5-6 New cache architecture

The copy-on-write process introduces significant latency into write operations. To isolate the active application from this additional latency, the FlashCopy indirection layer is placed logically between the upper and lower cache. Therefore, the additional latency that is introduced by the copy-on-write process is encountered only by the internal cache operations, and not by the application.

The logical placement of the FlashCopy indirection layer is shown in Figure 5-7.

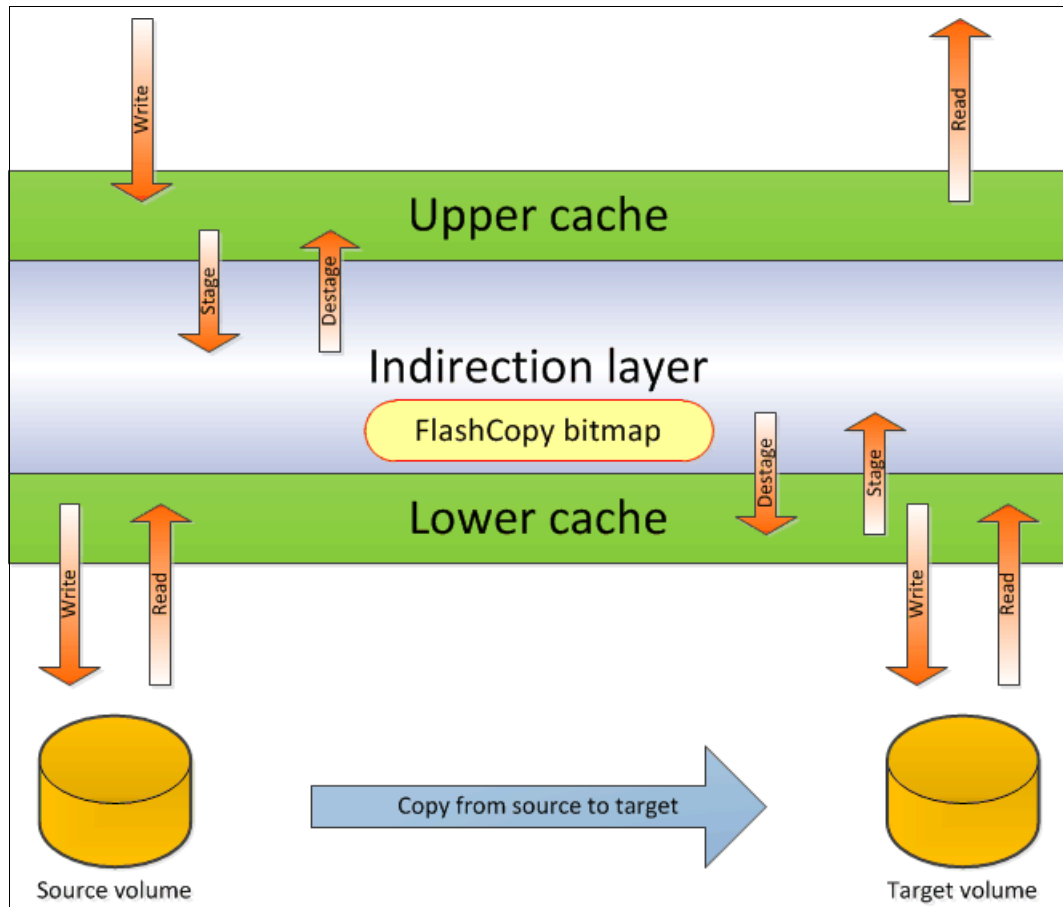


Figure 5-7 Logical placement of the FlashCopy indirection layer

The introduction of the two-level cache provides additional performance improvements to the FlashCopy mechanism. Because the FlashCopy layer is now above the lower cache in the IBM Spectrum Virtualize software stack, it can benefit from read pre-fetching and coalescing writes to back-end storage. Also, preparing FlashCopy is much faster because upper cache write data does not have to go directly to back-end storage, but just to the lower cache layer.

Additionally, in multi-target FlashCopy, the target volumes of the same image share cache data. This design is opposite to previous IBM Spectrum Virtualize code versions, where each volume had its own copy of cached data.

Interaction and dependency between Multiple Target FlashCopy mappings

Figure 5-8 on page 153 represents a set of three FlashCopy mappings that share a common source. The FlashCopy mappings target volumes Target 1, Target 2, and Target 3.

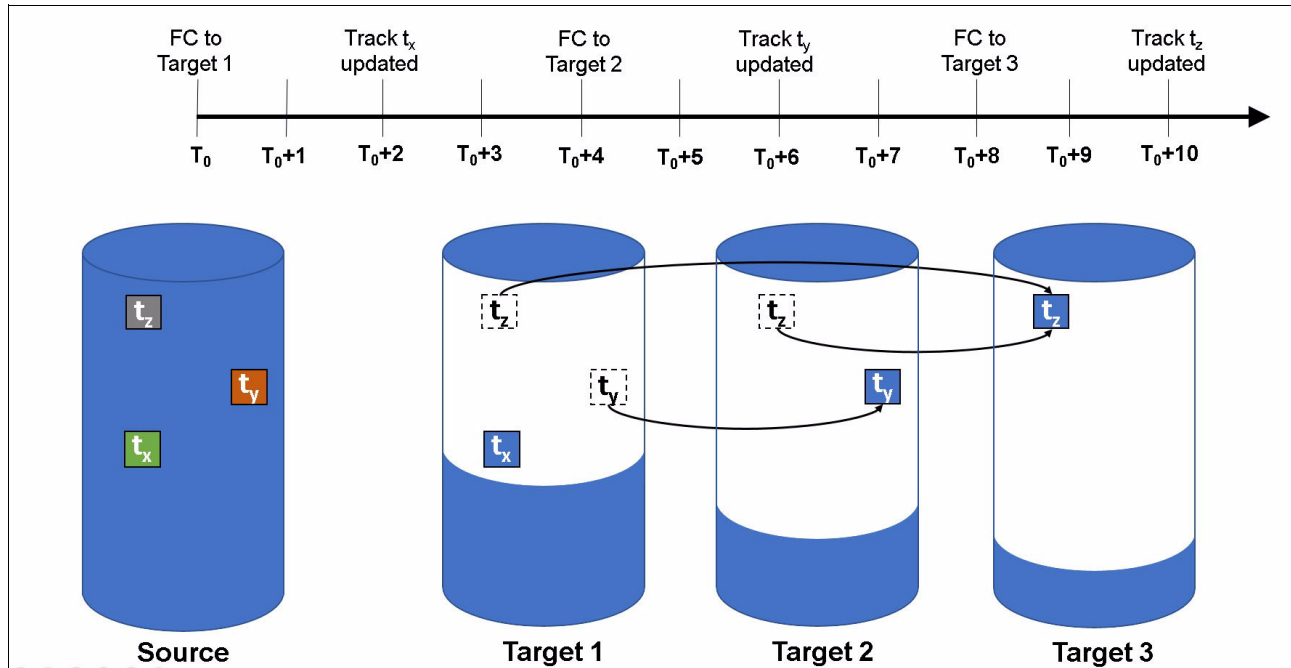


Figure 5-8 Interaction between Multiple Target FlashCopy mappings

Consider the following events timeline:

- ▶ At time T_0 a Flashcopy mapping is started between the source and the Target 1.
- ▶ At time T_0+2 the track t_x is updated in the source. Since this track has not yet been copied in background on Target 1, the copy-on-write process copies this track to the Target 1 before being updated on the source.
- ▶ At time T_0+4 a Flashcopy mapping is started between the source and the Target 2.
- ▶ At time T_0+6 the track t_y is updated in the source. Because this track has not yet been copied in background on Target 2, the copy-on-write process copies this track to the Target 2 only before being updated on the source.
- ▶ At time T_0+8 a Flashcopy mapping is started between the source and the Target 3.
- ▶ At time T_0+10 the track t_z is updated in the source. Because this track has not yet been copied in background on Target 3, the copy-on-write process copies this track to the Target 3 only before being updated on the source.

As result of this sequence of events, the configuration in Figure 5-8 has these characteristics:

- ▶ Target 1 is dependent upon Target 2 and Target 3. It remains dependent until all of Target 1 has been copied. No target depends on Target 1, so the mapping can be stopped without need to copy any data to maintain the consistency in the other targets.
- ▶ Target 2 depends on Target 3, and will remain dependent until all of Target 2 has been copied. Target 1 depends on Target 2, so if this mapping is stopped, the cleanup process is started to copy all data that is uniquely held on this mapping (that is t_y) to Target 1.
- ▶ Target 3 is not dependent on any target, but it has Target 1 and Target 2 depending on it, so if this mapping is stopped the cleanup process is started to copy all data that is uniquely held on this mapping (that is t_z) to Target 2.

Target writes with Multiple Target FlashCopy

A write to an intermediate or newest target volume must consider the state of the grain within its own mapping, and the state of the grain of the next oldest mapping:

- ▶ If the grain of the next oldest mapping has not been copied yet, it must be copied before the write is allowed to proceed to preserve the contents of the next oldest mapping. The data that is written to the next oldest mapping comes from a target or source.
- ▶ If the grain in the target being written has not yet been copied, the grain is copied from the oldest already copied grain in the mappings that are newer than the target, or the source if none are already copied. After this copy is done, the write can be applied to the target.

Target reads with Multiple Target FlashCopy

If the grain being read has already been copied from the source to the target, the read simply returns data from the target being read. If the grain has not been copied, each of the newer mappings is examined in turn and the read is performed from the first copy found. If none are found, the read is performed from the source.

5.2.4 FlashCopy planning considerations

The FlashCopy function, like all the advanced IBM Spectrum Virtualize and Storwize family product features, offers useful capabilities. However, some basic planning considerations are to be followed for a successful implementation.

FlashCopy configurations limits

To plan for and implement FlashCopy, you must check the configuration limits and adhere to them. Table 5-3 shows the system limits that apply to the latest version at the time of writing this book.

Table 5-3 *FlashCopy properties and maximum configurations*

FlashCopy property	Maximum	Comment
FlashCopy targets per source	256	This maximum is the maximum number of FlashCopy mappings that can exist with the same source volume.
FlashCopy mappings per system	5000	This property applies to these models: <ul style="list-style-type: none">▶ SAN Volume Controller 2145 models SV1, DH8, CG8, and CF8▶ Storwize V7000 2176 models 524 (Gen2) and 624 (Gen2+)
	4096	Any other Storwize models
FlashCopy Consistency Groups per system	500	This maximum is an arbitrary limit that is policed by the software.
FlashCopy volume space per I/O Group	4096 TB	This maximum is a limit on the quantity of FlashCopy mappings by using bitmap space from one I/O Group.
FlashCopy mappings per Consistency Group	512	This limit is due to the time that is taken to prepare a Consistency Group with many mappings.

Configuration Limits: The configuration limits always change with the introduction of new hardware and software capabilities. Check the IBM Spectrum Virtualize/Storwize online documentation for the latest configuration limits.

The total amount of cache memory reserved for the FlashCopy bitmaps limits the amount of capacity that can be used as a FlashCopy target. Table 5-4 illustrates the relationship of bitmap space to FlashCopy address space, depending on the size of the grain and the kind of FlashCopy service being used.

Table 5-4 Relationship of bitmap space to FlashCopy address space for the specified I/O Group

Copy Service	Grain size in KB	1 MB of memory provides the following volume capacity for the specified I/O Group
FlashCopy	256	2 TB of target volume capacity
FlashCopy	64	512 GB of target volume capacity
Incremental FlashCopy	256	1 TB of target volume capacity
Incremental FlashCopy	64	256 GB of target volume capacity

Mapping consideration: For multiple FlashCopy targets, you must consider the number of mappings. For example, for a mapping with a 256 KB grain size, 8 KB of memory allows one mapping between a 16 GB source volume and a 16 GB target volume. Alternatively, for a mapping with a 256 KB grain size, 8 KB of memory allows two mappings between one 8 GB source volume and two 8 GB target volumes.

When you create a FlashCopy mapping, if you specify an I/O Group other than the I/O Group of the source volume, the memory accounting goes towards the specified I/O Group, not towards the I/O Group of the source volume.

The default amount of memory for FlashCopy is 20 MB. This value can be increased or decreased by using the **chiogrp** command. The maximum amount of memory that can be specified for FlashCopy is 2048 MB (512 MB for 32-bit systems). The maximum combined amount of memory across all copy services features is 2600 MB (552 MB for 32-bit systems).

Bitmap allocation: When creating a FlashCopy mapping, you can optionally specify the I/O group where the bitmap is allocated. If you specify an I/O Group other than the I/O Group of the source volume, the memory accounting goes towards the specified I/O Group, not towards the I/O Group of the source volume. This option can be useful when an I/O group is exhausting the memory that is allocated to the FlashCopy bitmaps and no more free memory is available in the I/O group.

FlashCopy general restrictions

The following implementation restrictions apply to FlashCopy:

- ▶ The size of source and target volumes in a FlashCopy mapping must be the same.
- ▶ Multiple FlashCopy mappings that use the same target volume can be defined, but only one of these mappings can be started at a time. This limitation means that no multiple FlashCopy can be active to the same target volume.
- ▶ Expansion or shrinking of volumes defined in a FlashCopy mapping is not allowed. To modify the size of a source or target volume, first remove the FlashCopy mapping.

- ▶ In a cascading FlashCopy, the grain size of all the FlashCopy mappings that participate must be the same.
- ▶ In a multi-target FlashCopy, the grain size of all the FlashCopy mappings that participate must be the same.
- ▶ In a reverse FlashCopy, the grain size of all the FlashCopy mappings that participate must be the same.
- ▶ No FlashCopy mapping can be added to a consistency group while the FlashCopy mapping status is Copying.
- ▶ No FlashCopy mapping can be added to a consistency group while the consistency group status is Copying.
- ▶ The use of Consistency Groups is restricted when using Cascading FlashCopy. A Consistency Group serves the purpose of starting FlashCopy mappings at the same point in time. Within the *same* Consistency Group, it is not possible to have mappings with these conditions:
 - The source volume of one mapping is the target of another mapping.
 - The target volume of one mapping is the source volume for another mapping.

These combinations are not useful because within a Consistency Group, mappings cannot be established in a certain order. This limitation renders the content of the target volume undefined. For instance, it is not possible to determine whether the first mapping was established before the target volume of the first mapping that acts as a source volume for the second mapping.

Even if it were possible to ensure the order in which the mappings are established within a Consistency Group, the result is equal to Multi Target FlashCopy (two volumes holding the same target data for one source volume). In other words, a cascade is useful for copying volumes in a certain order (and copying the changed content targets of FlashCopies), rather than at the same time in an undefined order (from within one single Consistency Group).

- ▶ Both source and target volumes can be used as primary in a Remote Copy relationship. For more details about the FlashCopy and the Remote Copy possible interactions see “Interaction between Remote Copy and FlashCopy” on page 189.

FlashCopy presets

The IBM Spectrum Virtualize/Storwize GUI interface provides three FlashCopy presets (Snapshot, Clone, and Backup) to simplify the more common FlashCopy operations. Figure 5-9 on page 157 shows the preset selection panel in the GUI.

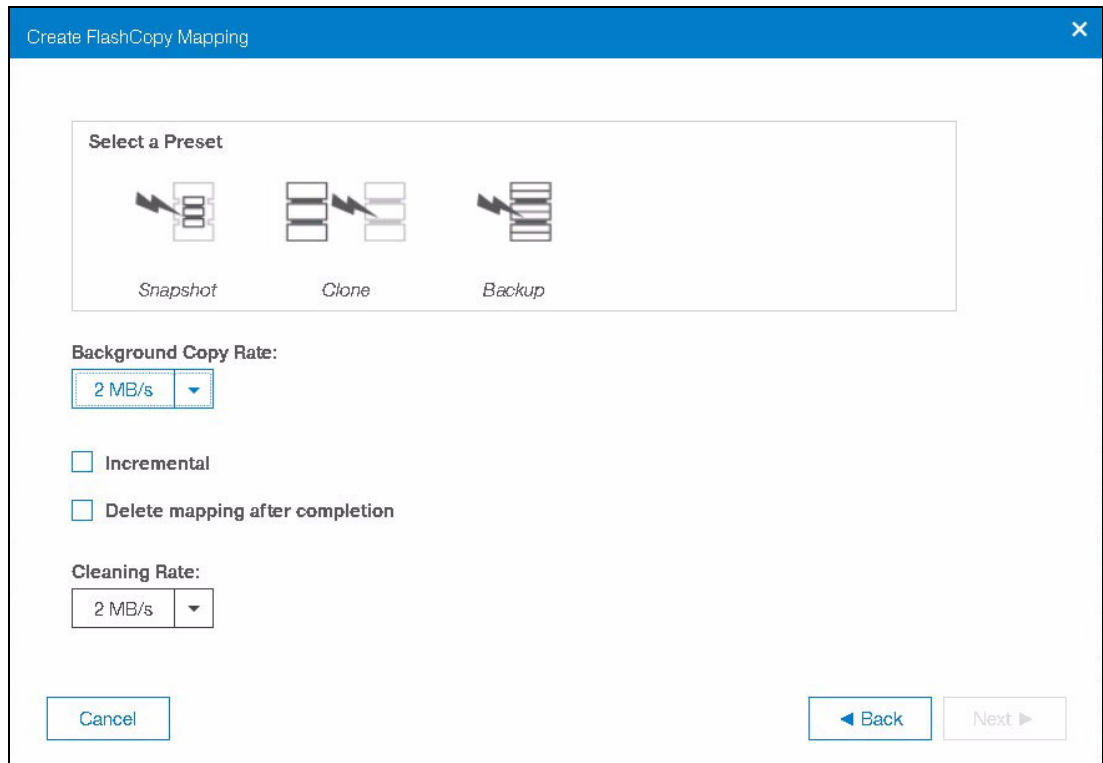


Figure 5-9 GUI Flashcopy Presets

Although these presets meet most FlashCopy requirements, they do not provide support for all possible FlashCopy options. If more specialized options are required that are not supported by the presets, the options must be performed by using CLI commands.

This section describes the three preset options and their use cases.

Snapshot

This preset creates a copy-on-write point-in-time copy. The snapshot is not intended to be an independent copy. Instead, the copy is used to maintain a view of the production data at the time that the snapshot is created. Therefore, the snapshot holds only the data from regions of the production volume that have changed since the snapshot was created. Because the snapshot preset uses thin provisioning, only the capacity that is required for the changes is used.

Snapshot uses the following preset parameters:

- ▶ Background copy: None
- ▶ Incremental: No
- ▶ Delete after completion: No
- ▶ Cleaning rate: No
- ▶ Primary copy source pool: Target pool

A typical use case for the Snapshot is when the user wants to produce a copy of a volume without affecting the availability of the volume. The user does not anticipate many changes to be made to the source or target volume. A significant proportion of the volumes remains unchanged.

By ensuring that only changes require a copy of data to be made, the total amount of disk space that is required for the copy is reduced. Therefore, many Snapshot copies can be used in the environment.

Snapshots are useful for providing protection against corruption or similar issues with the validity of the data. However, they do not provide protection from physical controller failures. Snapshots can also provide a vehicle for performing repeatable testing (including “what-if” modeling that is based on production data) without requiring a full copy of the data to be provisioned.

Clone

The clone preset creates a replica of the volume, which can then be changed without affecting the original volume. After the copy completes, the mapping that was created by the preset is automatically deleted.

Clone uses the following preset parameters:

- ▶ Background copy rate: 50
- ▶ Incremental: No
- ▶ Delete after completion: Yes
- ▶ Cleaning rate: 50
- ▶ Primary copy source pool: Target pool

A typical use case for the Snapshot is when users want a copy of the volume that they can modify without affecting the original volume. After the clone is established, there is no expectation that it is refreshed or that there is any further need to reference the original production data again. If the source is thin-provisioned, the target is thin-provisioned for the auto-create target.

Backup

The backup preset creates a point-in-time replica of the production data. After the copy completes, the backup view can be refreshed from the production data, with minimal copying of data from the production volume to the backup volume.

Backup uses the following preset parameters:

- ▶ Background Copy rate: 50
- ▶ Incremental: Yes
- ▶ Delete after completion: No
- ▶ Cleaning rate: 50
- ▶ Primary copy source pool: Target pool

The Backup preset can be used when the user wants to create a copy of the volume that can be used as a backup if the source becomes unavailable. This unavailability can happen during loss of the underlying physical controller. The user plans to periodically update the secondary copy, and does not want to suffer from the resource demands of creating a new copy each time.

Incremental FlashCopy times are faster than full copy, which helps to reduce the window where the new backup is not yet fully effective. If the source is thin-provisioned, the target is also thin-provisioned in this option for the auto-create target.

Another use case, which is not supported by the name, is to create and maintain (periodically refresh) an independent image. This image can be subjected to intensive I/O (for example, data mining) without affecting the source volume’s performance.

Grain size considerations

When creating a mapping a grain size of 64 KB can be specified as compared to the default 256 KB. This smaller grain size has been introduced specifically for the incremental FlashCopy, even though its use is not restricted to the incremental mappings.

In an incremental FlashCopy, the modified data is identified by using the bitmaps. The amount of data to be copied when refreshing the mapping depends on the grain size. If the grain size is 64 KB, as compared to 256 KB, there might be less data to copy to get a fully independent copy of the source again.

Incremental FlashCopy: For incremental FlashCopy, the 64 KB grain size is preferred.

Similar to the FlashCopy, the Thin Provisioned volumes also have a grain size attribute that represents the size of chunk of storage to be added to used capacity.

The following are the preferred settings for thin-provisioned FlashCopy:

- ▶ Thin-provisioned volume grain size must be equal to the FlashCopy grain size.
- ▶ Thin-provisioned volume grain size must be 64 KB for the best performance and the best space efficiency.

The exception is where the thin target volume is going to become a production volume (and is likely to be subjected to ongoing heavy I/O). In this case, the 256 KB thin-provisioned grain size is preferable because it provides better long-term I/O performance at the expense of a slower initial copy.

FlashCopy grain size considerations: Even if the 256 KB thin-provisioned volume grain size is chosen, it is still beneficial to limit the FlashCopy grain size to 64 KB. It is possible to minimize the performance impact to the source volume, even though this size increases the I/O workload on the target volume.

However, clients with very large numbers of FlashCopy/Remote Copy relationships might still be forced to choose a 256 KB grain size for FlashCopy to avoid constraints on the amount of bitmap memory.

Volume placement considerations

The source and target volumes placement among the pools and the I/O groups must be planned to minimize the effect of the underlying FlashCopy processes. In normal condition (that is with all the nodes/canisters fully operative), the FlashCopy background copy workload distribution follows this schema:

- ▶ The preferred node of the source volume is responsible for the background copy read operations.
- ▶ The preferred node of the target volume is responsible for the background copy write operations.

Table 5-5 shows how the backend I/O operations are distributed across the nodes.

Table 5-5 Workload distribution for backend I/O operations

	Read from source	Read from target	Write to source	Write to target
Node that performs the back-end I/O if the grain is copied	Preferred node in source volume's I/O group	Preferred node in target volume's I/O group	Preferred node in source volume's I/O group	Preferred node in target volume's I/O group
Node that performs the back-end I/O if the grain is not yet copied	Preferred node in source volume's I/O group	Preferred node in source volume's I/O group	The preferred node in source volume's I/O group will read and write, and the preferred node in target volume's I/O group will write	The preferred node in source volume's I/O group will read, and the preferred node in target volume's I/O group will write

Note that the data transfer among the source and the target volume's preferred nodes occurs through the node-to-node connectivity. Consider the following volume placement alternatives:

- ▶ Source and target volumes use the same preferred node.
In this scenario, the node that is acting as preferred for both source and target volume manages all the read and write FlashCopy operations. Only resources from this node are consumed for the FlashCopy operations, and no node-to-node bandwidth is used.
- ▶ Source and target volumes use the different preferred node.
In this scenario, both nodes that are acting as preferred nodes manage read and write FlashCopy operations according to the schemes described above. The data that is transferred between the two preferred nodes goes through the node-to-node network.

Both alternatives described have advantages and disadvantages, but in general option 1 is preferred. We explore the following scenarios:

1. IBM Spectrum Virtualize or Storwize systems with multiple I/O groups where the source volumes are evenly spread across all the nodes.
Assuming also that the I/O workload is evenly distributed across the nodes, alternative 1 is preferable. In fact, the amount of read and write FlashCopy operations will again be evenly spread across the nodes without using any node-to-node bandwidth.
2. IBM Spectrum Virtualize or Storwize system with multiple I/O groups where the source volumes and most of the workload are concentrated in some nodes.
In this case alternative 2 can also be considered. In fact, defining the target volumes preferred node in the less utilized nodes will relieve the source volumes preferred node of some additional FlashCopy workload (especially during the background copy).
3. IBM Spectrum Virtualize system with multiple I/O groups in Enhanced Stretched Cluster configuration where the source volumes are evenly spread across all the nodes.
In this case the preferred node placement should follow the location of the source and target volumes on the backend storage. For example, if the source volume is on site A and the target volume is on site B, then the target volumes preferred node must be in site B. Placing the target volumes preferred node in site A will cause the re-direction of the FlashCopy write operation through the node-to-node network.

4. A clustered Storwize system with multiple control enclosures where the source volumes are evenly spread across all the canisters.

In this case the preferred node placement should follow the location of source and target volumes on the internal storage. For example, if the source volume is on the internal storage attached to control enclosure A and the target volume is on internal storage attached to control enclosure B, then the target volumes preferred node must be in one canister of control enclosure B. Placing the target volumes preferred node on control enclosure A will cause the re-direction of the FlashCopy write operation through the node-to-node network.

Placement on the back-end storage is mainly driven by the availability requirements. Generally, use different back-end storage controllers or arrays for the source and target volumes.

Background copy considerations

The background copy process uses internal resources such as CPU, memory, and bandwidth. This copy process tries to reach the target copy data rate for every volume according to the background copy rate parameter setting (as reported in Table 5-1 on page 144).

If the copy process is unable to achieve these goals, it starts contending resources to the foreground I/O (that is the I/O coming from the hosts). As result, both background copy and foreground I/O will tend to see an increase in latency and therefore reduction in throughput compared to the situation when the bandwidth not been limited. Degradation is graceful. Both background copy and foreground I/O continue to make progress, and will not stop, hang, or cause the node to fail.

To avoid any impact on the foreground I/O, that is in the hosts response time, carefully plan the background copy activity, taking in account the overall workload running in the systems. The background copy basically reads and writes data to managed disks. Usually, the most affected component is the back-end storage. CPU and memory are not normally significantly affected by the copy activity.

The theoretical added workload due to the background copy is easily estimable. For instance, starting 20 FlashCopy with a background copy rate of 70 each adds a maximum throughput of 160 MBps for the reads and 160 MBps for the writes.

The source and target volumes distribution on the back-end storage determines where this workload is going to be added. The duration of the background copy depends on the amount of data to be copied. This amount is the total size of volumes for full background copy or the amount of data that is modified for incremental copy refresh.

Performance monitoring tools like IBM Spectrum Control can be used to evaluate the existing workload on the back-end storage in a specific time window. By adding this workload to the foreseen background copy workload, you can estimate the overall workload running toward the back-end storage. Disk performance simulation tools, like Disk Magic, can be used to estimate the effect, if any, of the added back-end workload to the host service time during the background copy window. The outcomes of this analysis can provide useful hints for the background copy rate settings.

When performance monitoring and simulation tools are not available, use a conservative and progressive approach. Consider that the background copy setting can be modified at any time, even when the FlashCopy is already started. The background copy process can even be completely stopped by setting the background copy rate to 0.

Initially set the background copy rate value to add a limited workload to the backend (for example less than 100 MBps). If no effects on hosts are noticed, the background copy rate value can be increased. Do this process until you see negative effects. Note that the background copy rate setting follows an exponential scale, so changing for instance from 50 to 60 doubles the data rate goal from 2 MBps to 4 MBps.

Cleaning process and Cleaning Rate

The Cleaning Rate is the rate at which the data is copied among dependent FlashCopies such as Cascaded and Multi Target FlashCopy. The Cleaning process aims to release the dependency of a mapping in such a way that it can be stopped immediately (without going to the stopping state). The typical use case for setting the Cleaning Rate is when it is required to stop a Cascaded or Multi Target FlashCopy that is not the oldest in the FlashCopy chain. In this case to avoid the stopping state lasting for a long time, the cleaning rate can be adjusted accordingly.

There is an interaction between the background copy rate and the Cleaning Rate settings:

- ▶ Background copy = 0 and Cleaning Rate = 0
No background copy or cleaning take place. When the mapping is stopped, it goes into stopping state and a cleaning process starts with the default cleaning rate, which is 50 or 2 MBps.
- ▶ Background copy > 0 and Cleaning Rate = 0
The background copy takes place at the background copy rate but no cleaning process is started. When the mapping is stopped, it goes into stopping state and a cleaning process starts with the default cleaning rate (50 or 2 MBps).
- ▶ Background copy = 0 and Cleaning Rate > 0
No background copy takes place, but the cleaning process runs at the cleaning rate. When the mapping is stopped, the cleaning completes (if not yet completed) at the cleaning rate.
- ▶ Background copy > 0 and Cleaning Rate > 0
The background copy takes place at the background copy rate but no cleaning process is started. When the mapping is stopped, it goes into stopping state and a cleaning process starts with the specified cleaning rate.

Regarding the workload considerations for the cleaning process, the same guidelines as for background copy apply.

Host and application considerations to ensure FlashCopy integrity

Because FlashCopy is at the block level, it is necessary to understand the interaction between your application and the host operating system. From a logical standpoint, it is easiest to think of these objects as “layers” that sit on top of one another. The application is the topmost layer, and beneath it is the operating system layer.

Both of these layers have various levels and methods of caching data to provide better speed. Because IBM Spectrum Virtualize systems, and therefore FlashCopy, sit below these layers, they are unaware of the cache at the application or operating system layers.

To ensure the integrity of the copy that is made, it is necessary to flush the host operating system and application cache for any outstanding reads or writes before the FlashCopy operation is performed. Failing to flush the host operating system and application cache produces what is referred to as a *crash consistent* copy.

The resulting copy requires the same type of recovery procedure, such as log replay and file system checks, that is required following a host crash. FlashCopies that are crash consistent often can be used following file system and application recovery procedures.

Note: Although the best way to perform FlashCopy is to flush host cache first, some companies, like Oracle, support using snapshots without it, as stated in Metalink note 604683.1.

Various operating systems and applications provide facilities to stop I/O operations and ensure that all data is flushed from host cache. If these facilities are available, they can be used to prepare for a FlashCopy operation. When this type of facility is not available, the host cache must be flushed manually by quiescing the application and unmounting the file system or drives.

Preferred practice: From a practical standpoint, when you have an application that is backed by a database and you want to make a FlashCopy of that application's data, it is sufficient in most cases to use the write-suspend method that is available in most modern databases. You can use this method because the database maintains strict control over I/O.

This method is as opposed to flushing data from both the application and the backing database, which is always the suggested method because it is safer. However, this method can be used when facilities do not exist or your environment includes time sensitivity.

5.3 Remote Copy services

IBM Spectrum Virtualize and Storwize technology offers various remote copy services functions that address Disaster Recovery and Business Continuity needs.

Metro Mirror is designed for metropolitan distances with a zero recovery point objective (RPO), which is zero data loss. This objective is achieved with a synchronous copy of volumes. Writes are not acknowledged until they are committed to both storage systems. By definition, any vendors' synchronous replication makes the host wait for write I/Os to complete at both the local and remote storage systems, and includes round-trip network latencies. Metro Mirror has the following characteristics:

- ▶ Zero RPO
- ▶ Synchronous
- ▶ Production application performance that is affected by round-trip latency

Global Mirror is designed to minimize application performance impact by replicating asynchronously. That is, writes are acknowledged as soon as they can be committed to the local storage system, sequence-tagged, and passed on to the replication network. This technique allows Global Mirror to be used over longer distances. By definition, any vendors' asynchronous replication results in an RPO greater than zero. However, for Global Mirror, the RPO is quite small, typically anywhere from several milliseconds to some number of seconds.

Although Global Mirror is asynchronous, the RPO is still small, and thus the network and the remote storage system must both still be able to cope with peaks in traffic. Global Mirror has the following characteristics:

- ▶ Near-zero RPO
- ▶ Asynchronous
- ▶ Production application performance that is affected by I/O sequencing preparation time

Global Mirror with Change Volumes provides an option to replicate point-in-time copies of volumes. This option generally requires lower bandwidth because it is the average rather than the peak throughput that must be accommodated. The RPO for Global Mirror with Change Volumes is higher than traditional Global Mirror. Global Mirror with Change Volumes has the following characteristics:

- ▶ Larger RPO
- ▶ Point-in-time copies
- ▶ Asynchronous
- ▶ Possible system performance effect because point-in-time copies are created locally

Successful implementation depends on taking a holistic approach in which you consider all components and their associated properties. The components and properties include host application sensitivity, local and remote SAN configurations, local and remote system and storage configuration, and the intersystem network.

5.3.1 Remote copy functional overview

In this section, the terminology and the basic functional aspects of the remote copy services are presented.

Common terminology and definitions

When such a breadth of technology areas is covered, the same technology component can have multiple terms and definitions. This document uses the following definitions:

- ▶ *Local system or master system*
The system on which the foreground applications run.
- ▶ *Local hosts*
Hosts that run on the foreground applications.
- ▶ *Master volume or source volume*
The local volume that is being mirrored. The volume has nonrestricted access. Mapped hosts can read and write to the volume.
- ▶ *Intersystem link or intersystem network*
The network that provides connectivity between the local and the remote site. It can be a Fibre Channel network (SAN), an IP network, or a combination of the two.
- ▶ *Remote system or auxiliary system*
The system that holds the remote mirrored copy.
- ▶ *Auxiliary volume or target volume*
The remote volume that holds the mirrored copy. It is read-access only.
- ▶ *Remote copy*
A generic term that is used to describe a Metro Mirror or Global Mirror relationship in which data on the source volume is mirrored to an identical copy on a target volume. Often the two copies are separated by some distance, which is why the term *remote* is used to describe the copies. However, having remote copies is not a prerequisite. A remote copy relationship includes the following states:
 - Consistent relationship
A remote copy relationship where the data set on the target volume represents a data set on the source volumes at a certain point.

- Synchronized relationship

A relationship is *synchronized* if it is consistent *and* the point that the target volume represents is the current point. The target volume contains identical data as the source volume.

- ▶ *Synchronous remote copy* (Metro Mirror)

Writes to the source and target volumes that are committed in the foreground before confirmation is sent about completion to the local host application.

- ▶ *Asynchronous remote copy* (Global Mirror)

A foreground write I/O is acknowledged as complete to the local host application before the mirrored foreground write I/O is cached at the remote system. Mirrored foreground writes are processed asynchronously at the remote system, but in a committed sequential order as determined and managed by the Global Mirror remote copy process.

- ▶ *Global Mirror Change Volume*

Holds earlier consistent revisions of data when changes are made. A change volume must be created for the master volume and the auxiliary volume of the relationship.

- ▶ The *background copy* process manages the initial synchronization or resynchronization processes between source volumes to target mirrored volumes on a remote system.
- ▶ *Foreground I/O* reads and writes I/O on a local SAN, which generates a mirrored foreground write I/O that is across the intersystem network and remote SAN.

Figure 5-10 shows some of the concepts of remote copy.

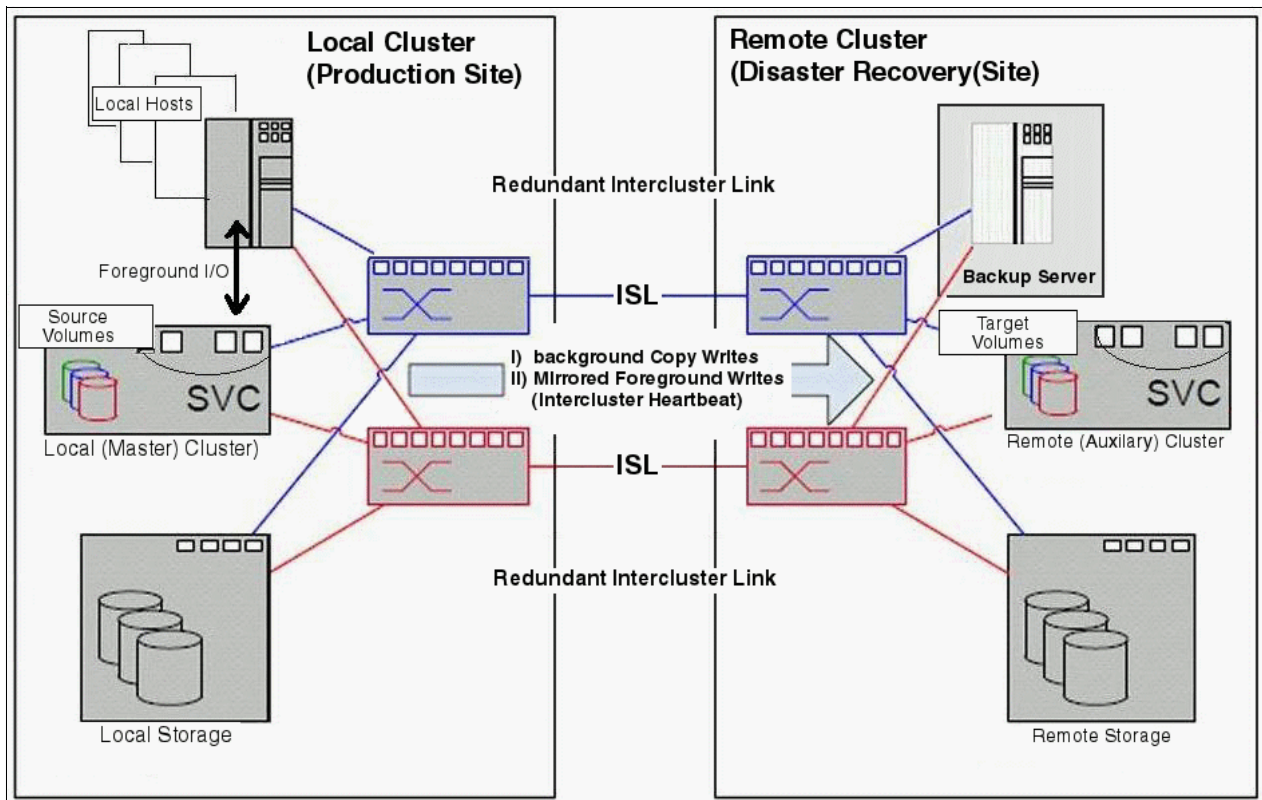


Figure 5-10 Remote copy components and applications

A successful implementation of intersystem remote copy services significantly depends on the quality and configuration of the intersystem network.

Remote Copy partnerships and relationships

A remote copy *partnership* is a partnership that is established between a master (local) system and an auxiliary (remote) system, as shown in Figure 5-11.

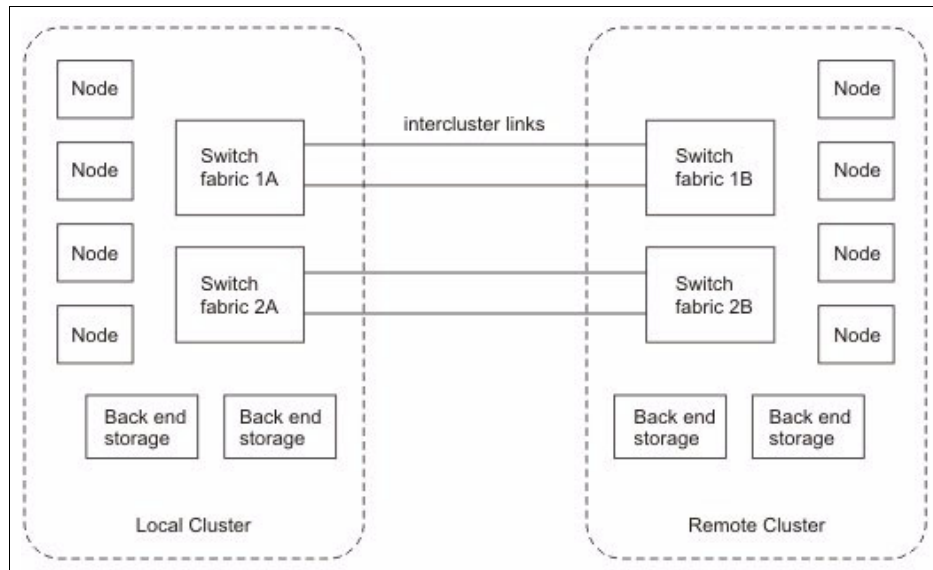


Figure 5-11 Remote copy partnership

Partnerships are established between two systems by issuing the **mkfcpartnership** or **mkippartnership** command once from each end of the partnership. The parameters that need to be specified are the remote system name (or ID), the available bandwidth (in Mbps), and the maximum background copy rate as a percentage of the available bandwidth. The background copy parameter determines the maximum speed of the initial synchronization and resynchronization of the relationships.

Tip: To establish a fully functional Metro Mirror or Global Mirror partnership, issue the **mkfcpartnership** or **mkippartnership** command from both systems.

A remote copy *relationship* is a relationship that is established between a source (primary) volume in the local system and a target (secondary) volume in the remote system. Usually when a remote copy relationship is started, a background copy process that copies the data from source to target volumes is started as well.

In addition to the background copy rate setting, the initial synchronization can be adjusted at relationship level with the `relationship_bandwidth_limit` parameter. The `relationship_bandwidth_limit` is a system-wide parameter that sets the maximum bandwidth that can be used to initially synchronize a single relationship.

After background synchronization or resynchronization is complete, a Global Mirror relationship provides and maintains a consistent mirrored copy of a source volume to a target volume.

Copy directions and default roles

When you create a remote copy relationship, the source or master volume is initially assigned the role of the master, and the target auxiliary volume is initially assigned the role of the auxiliary. This design implies that the initial copy direction of mirrored foreground writes and background resynchronization writes (if applicable) is from master to auxiliary.

After the initial synchronization is complete, you can change the copy direction (see Figure 5-12). The ability to change roles is used to facilitate disaster recovery.

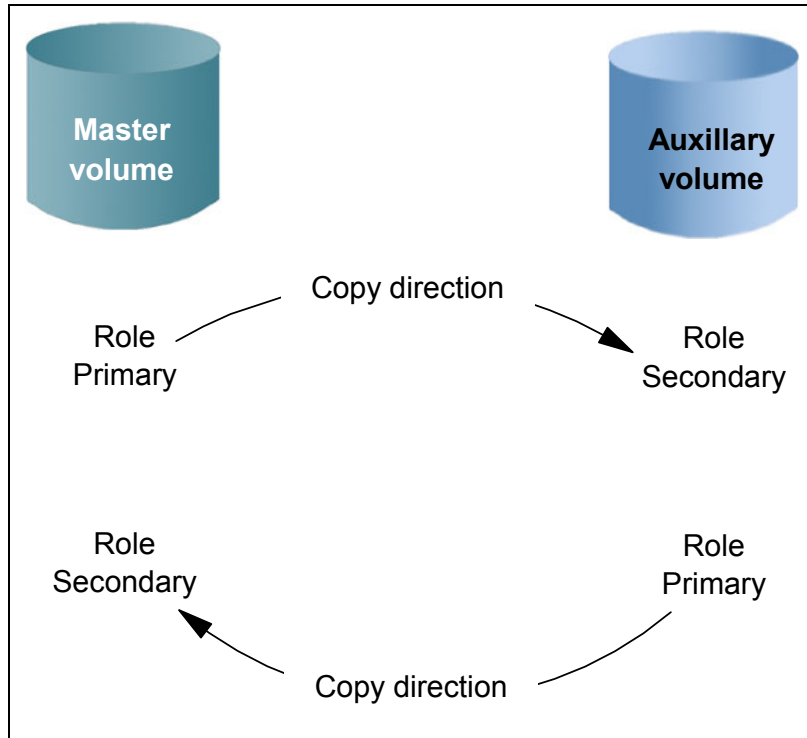


Figure 5-12 Role and direction changes

Attention: When the direction of the relationship is changed, the roles of the volumes are altered. A consequence is that the read/write properties are also changed, meaning that the master volume takes on a secondary role and becomes read-only.

Consistency Groups

A Consistency Group (CG) is a collection of relationships that can be treated as one entity. This technique is used to preserve write order consistency across a group of volumes that pertain to one application, for example, a database volume and a database log file volume.

After a remote copy relationship is added into a Consistency Group, you cannot manage the relationship in isolation from the Consistency Group. So, for example, issuing a **stoprelationship** command on the stand-alone volume would fail because the system knows that the relationship is part of a Consistency Group.

Note the following points regarding Consistency Groups:

- ▶ Each volume relationship can belong to only one Consistency Group.
- ▶ Volume relationships can also be stand-alone, that is, not in any Consistency Group.
- ▶ Consistency Groups can also be created and left empty, or can contain one or many relationships.
- ▶ You can create up to 256 Consistency Groups on a system.
- ▶ All volume relationships in a Consistency Group must have matching primary and secondary systems, but they do not need to share I/O groups.

- ▶ All relationships in a Consistency Group have the same copy direction and state.
- ▶ Each Consistency Group is either for Metro Mirror or for Global Mirror relationships, but not both. This choice is determined by the first volume relationship that is added to the Consistency Group.

Consistency Group consideration: A Consistency Group relationship does not have to be in a directly matching I/O group number at each site. A Consistency Group owned by I/O group 1 at the local site does not have to be owned by I/O group 1 at the remote site. If you have more than one I/O group at either site, you can create the relationship between any two I/O groups. This technique spreads the workload, for example, from local I/O group 1 to remote I/O group 2.

Streams

Consistency Groups can also be used as a way to spread replication workload across multiple streams within a partnership.

The Metro or Global Mirror partnership architecture allocates traffic from each Consistency Group in a round-robin fashion across 16 streams. That is, cg0 traffic goes into stream0, and cg1 traffic goes into stream1.

Any volume that is *not* in a Consistency Group also goes into stream0. You might want to consider creating an empty Consistency Group 0 so that stand-alone volumes do not share a stream with active Consistency Group volumes.

It can also pay to optimize your streams by creating more Consistency Groups. Within each stream, each batch of writes must be processed in tag sequence order and any delays in processing any particular write also delays the writes behind it in the stream. Having more streams (up to 16) reduces this kind of potential congestion.

Each stream is sequence-tag-processed by one node, so generally you would want to create at least as many Consistency Groups as you have IBM Spectrum Virtualize nodes/Storwize canisters, and, ideally, perfect multiples of the node count.

Layer concept

Version 6.3 introduced the concept of *layer*, which allows you to create partnerships among IBM Spectrum Virtualize and Storwize products. The key points concerning layers are listed here:

- ▶ IBM Spectrum Virtualize is always in the *Replication* layer.
- ▶ By default, Storwize products are in the *Storage* layer.
- ▶ A system can only form partnerships with systems in the same layer.
- ▶ An IBM Spectrum Virtualize can virtualize a Storwize system only if the Storwize is in Storage layer.
- ▶ With version 6.4, a Storwize system in the Replication layer can virtualize a Storwize system in the Storage layer.

Figure 5-13 illustrates the concept of layers.

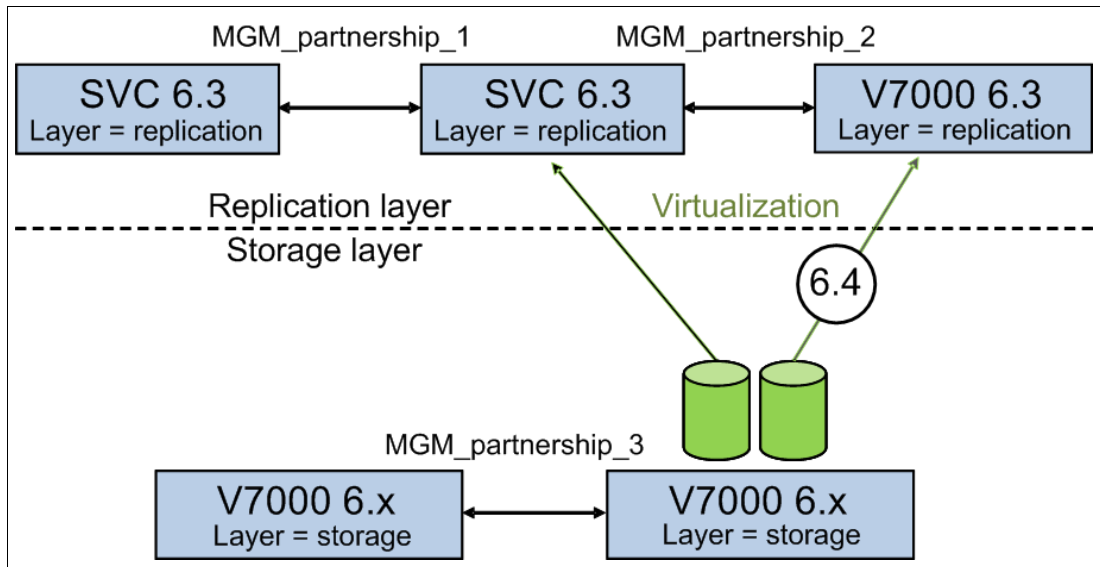


Figure 5-13 Conceptualization of layers

Generally, changing the layer is only performed at initial setup time or as part of a major reconfiguration. To change the layer of a Storwize system, the system must meet the following pre-conditions:

- ▶ The Storwize system must not have any IBM Spectrum Virtualize or Storwize host objects defined, and must not be virtualizing any other Storwize controllers.
- ▶ The Storwize system must not be visible to any other IBM Spectrum Virtualize or Storwize system in the SAN fabric, which might require SAN zoning changes.
- ▶ The Storwize system must not have any system partnerships defined. If it is already using Metro Mirror or Global Mirror, the existing partnerships and relationships must be removed first.

Changing a Storwize system from Storage layer to Replication layer can only be performed by using the CLI. After you are certain that all of the pre-conditions have been met, issue the following command:

```
chsystem -layer replication
```

Partnership topologies

Each system can be connected to a maximum of three other systems for the purposes of Metro or Global Mirror.

Figure 5-14 shows examples of the principal supported topologies for Metro and Global Mirror partnerships. Each box represents an IBM Spectrum Virtualize or Storwize system.

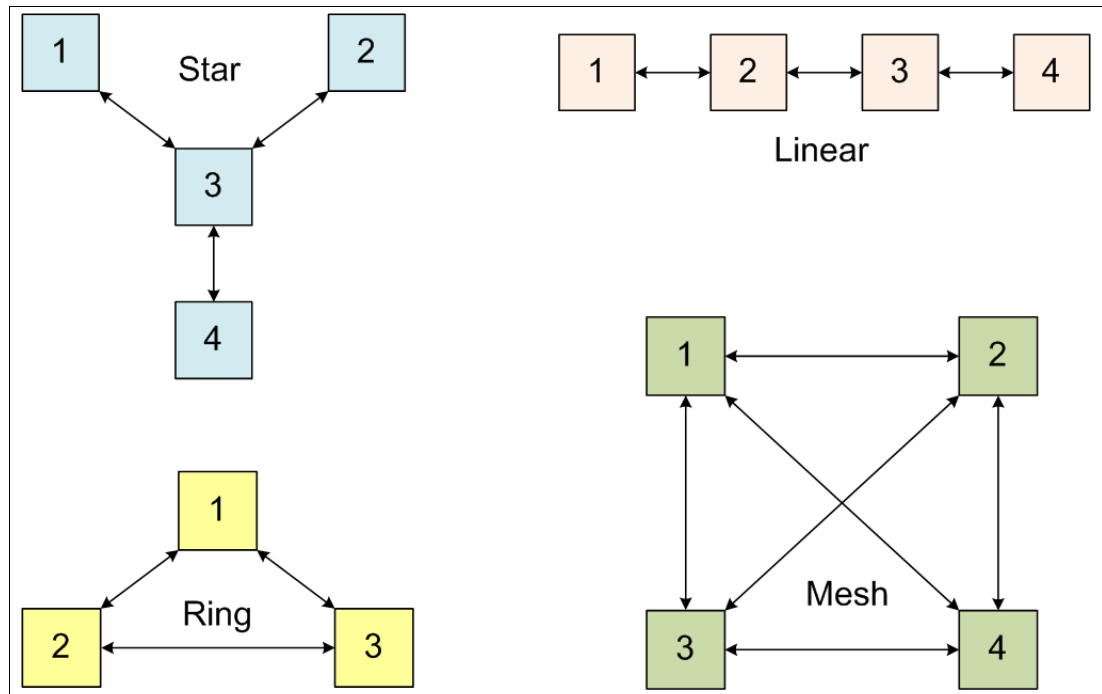


Figure 5-14 Supported topologies for Metro and Global Mirror

Star topology

A star topology can be used, for example, to share a centralized disaster recovery system (3, in this example) with up to three other systems, for example replicating 1 → 3, 2 → 3, and 4 → 3.

Ring topology

A ring topology (3 or more systems) can be used to establish a one-in, one-out implementation. For example, the implementation can be 1 → 2, 2 → 3, 3 → 1 to spread replication loads evenly among three systems.

Linear topology

A linear topology of two or more sites is also possible. However, it would generally be simpler to create partnerships between system 1 and system 2, and separately between system 3 and system 4.

Mesh topology

A fully connected mesh topology is where every system has a partnership to each of the three other systems. This topology allows flexibility in that volumes can be replicated between any two systems.

Topology considerations:

- ▶ Although systems can have up to three partnerships, any one volume can be part of only a single relationship. That is, you cannot replicate any given volume to multiple remote sites.
- ▶ Although various topologies are supported, it is advisable to keep your partnerships as simple as possible, which in most cases means system pairs or a star.

Intrasystem versus intersystem

Although remote copy services are available for intrasystem, it has no functional value for production use. Intrasystem Metro Mirror provides the same capability with less overhead. However, leaving this function in place simplifies testing and allows for experimentation and testing. For example, you can validate server failover on a single test system.

Intrasystem remote copy: Intrasystem remote copy is not supported on IBM Spectrum Virtualize/Storwize systems that run V6 or later.

Metro Mirror functional overview

Metro Mirror provides synchronous replication. It is designed to ensure that updates are committed to both the primary and secondary volumes before sending an acknowledgment (Ack) of the completion to the server.

If the primary volume fails completely for any reason, Metro Mirror is designed to ensure that the secondary volume holds the same data as the primary did immediately before the failure.

Metro Mirror provides the simplest way to maintain an identical copy on both the primary and secondary volumes. However, as with any synchronous copy over long distance, there can be a performance impact to host applications due to network latency.

Metro Mirror supports relationships between volumes that are up to 300 km apart. Latency is an important consideration for any Metro Mirror network. With typical fiber optic round-trip latencies of 1 ms per 100 km, you can expect a minimum of 3 ms extra latency, due to the network alone, on each I/O if you are running across the 300 km separation.

Figure 5-15 shows the order of Metro Mirror write operations.

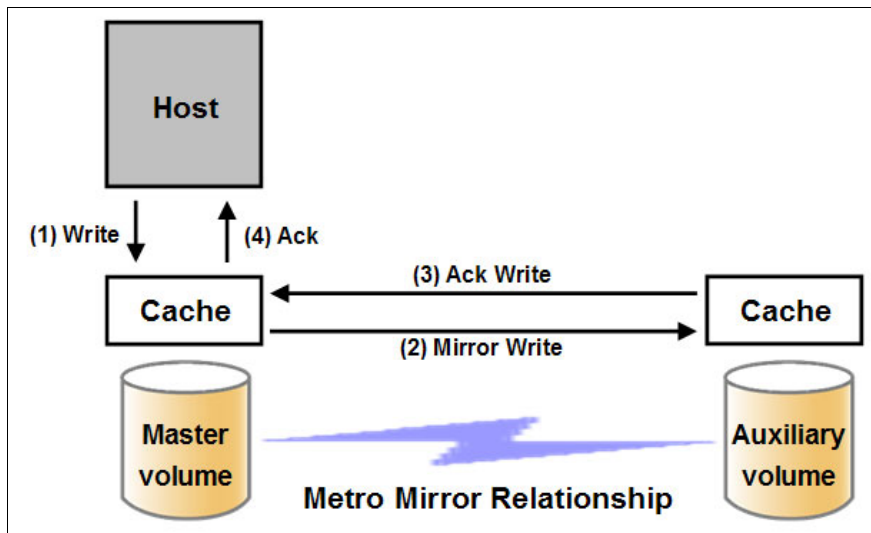


Figure 5-15 Metro Mirror write sequence

A write into mirrored cache on an IBM Spectrum Virtualize or Storwize system is all that is required for the write to be considered as committed. De-staging to disk is a natural part of I/O management, but it is not generally in the critical path for a Metro Mirror write acknowledgment.

Global Mirror functional overview

Global Mirror provides asynchronous replication. It is designed to reduce the dependency on round-trip network latency by acknowledging the primary write in parallel with sending the write to the secondary volume.

If the primary volume fails completely for any reason, Global Mirror is designed to ensure that the secondary volume holds the same data as the primary did at a point a short time before the failure. That short period of data loss is typically between 10 ms and 10 seconds, but varies according to individual circumstances.

Global Mirror provides a way to maintain a write-order-consistent copy of data at a secondary site only slightly behind the primary. Global Mirror has minimal impact on the performance of the primary volume.

Although Global Mirror is an asynchronous remote copy technique, foreground writes at the local system and mirrored foreground writes at the remote system are not wholly independent of one another. IBM Spectrum Virtualize/Storwize implementation of asynchronous remote copy uses algorithms to maintain a consistent image at the target volume always.

They achieve this image by identifying sets of I/Os that are active concurrently at the source, assigning an order to those sets, and applying these sets of I/Os in the assigned order at the target. The multiple I/Os within a single set are applied concurrently.

The process that marshals the sequential sets of I/Os operates at the remote system, and therefore is not subject to the latency of the long-distance link.

Figure 5-16 shows that a write operation to the master volume is acknowledged back to the host that issues the write before the write operation is mirrored to the cache for the auxiliary volume.

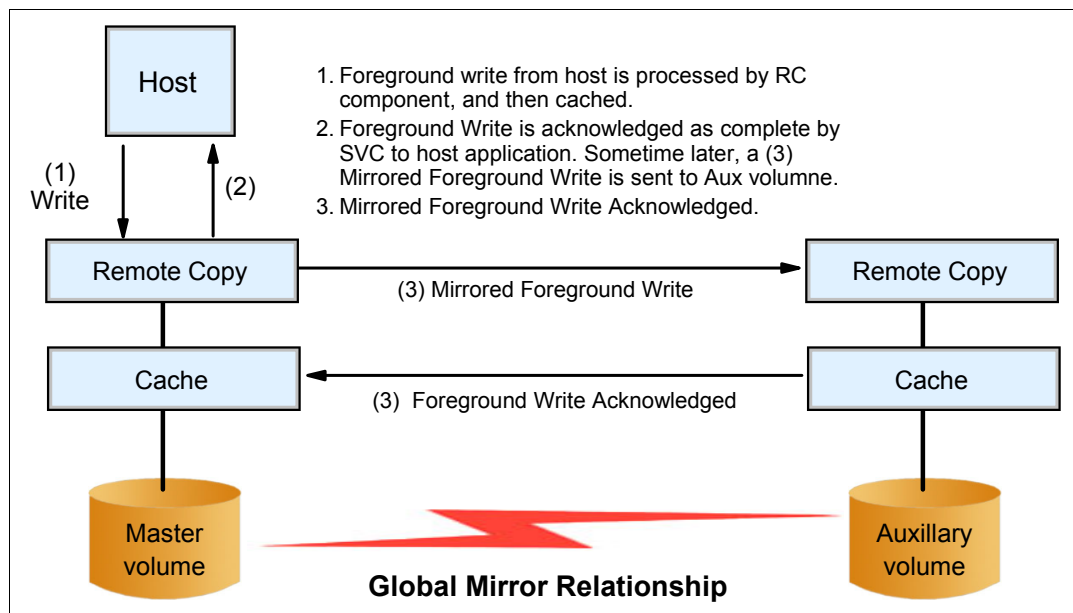


Figure 5-16 Global Mirror relationship write operation

With Global Mirror, a confirmation is sent to the host server before the host receives a confirmation of the completion at the auxiliary volume. The GM function identifies sets of write I/Os that are active concurrently at the primary volume. It then assigns an order to those sets, and applies these sets of I/Os in the assigned order at the auxiliary volume.

Further writes might be received from a host when the secondary write is still active for the same block. In this case, although the primary write might complete, the new host write on the auxiliary volume is delayed until the previous write is completed.

Write ordering

Many applications that use block storage are required to survive failures, such as a loss of power or a software crash. They are also required to not lose data that existed before the failure. Because many applications must perform many update operations in parallel to that storage block, maintaining write ordering is key to ensuring the correct operation of applications after a disruption.

An application that performs a high volume of database updates is often designed with the concept of dependent writes. Dependent writes ensure that an earlier write completes before a later write starts. Reversing the order of dependent writes can undermine the algorithms of the application and can lead to problems, such as detected or undetected data corruption.

Colliding writes

Colliding writes are defined as new write I/Os that overlap existing active write I/Os.

Before V4.3.1, the Global Mirror algorithm required only a single write to be active on any 512-byte LBA of a volume. If another write was received from a host while the auxiliary write was still active, the new host write was delayed until the auxiliary write was complete (although the master write might complete). This restriction was needed if a series of writes to the auxiliary must be retried (which is known as *reconstruction*). Conceptually, the data for reconstruction comes from the master volume.

If multiple writes were allowed to be applied to the master for a sector, only the most recent write had the correct data during reconstruction. If reconstruction was interrupted for any reason, the intermediate state of the auxiliary was inconsistent.

Applications that deliver such write activity do not achieve the performance that Global Mirror is intended to support. A volume statistic is maintained about the frequency of these collisions. Starting with V4.3.1, an attempt is made to allow multiple writes to a single location to be outstanding in the Global Mirror algorithm.

A need still exists for master writes to be serialized. The intermediate states of the master data must be kept in a non-volatile journal while the writes are outstanding to maintain the correct write ordering during reconstruction. Reconstruction must never overwrite data on the auxiliary with an earlier version. The colliding writes of volume statistic monitoring are now limited to those writes that are not affected by this change.

Figure 5-17 shows a colliding write sequence.

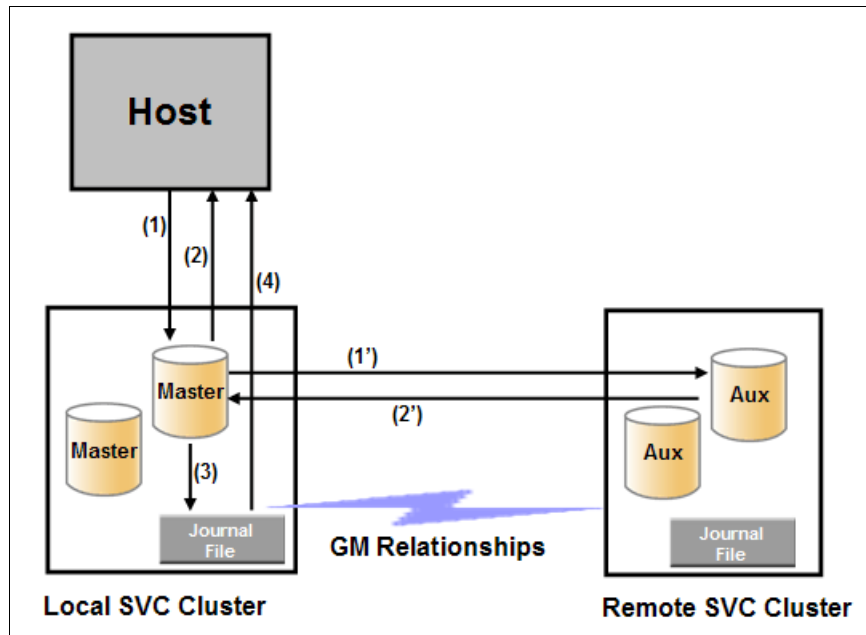


Figure 5-17 Colliding writes

The following numbers correspond to the numbers that are shown in Figure 5-17:

1. A first write is performed from the host to LBA X.
2. A host is provided acknowledgment that the write is complete, even though the mirrored write to the auxiliary volume is not yet completed.

The first two actions (1 and 2) occur asynchronously with the first write.

3. A second write is performed from the host to LBA X. If this write occurs before the host receives acknowledgment (2), the write is written to the journal file.
4. A host is provided acknowledgment that the second write is complete.

Global Mirror Change Volumes functional overview

Global Mirror with Change Volumes (GM/CV) provides asynchronous replication based on point-in-time copies of data. It is designed to allow for effective replication over lower bandwidth networks and to reduce any impact on production hosts.

Metro Mirror and Global Mirror both require the bandwidth to be sized to meet the peak workload. Global Mirror with Change Volumes must only be sized to meet the average workload across a cycle period.

Figure 5-18 shows a high-level conceptual view of Global Mirror with Change Volumes. GM/CV uses FlashCopy to maintain image consistency and to isolate host volumes from the replication process.

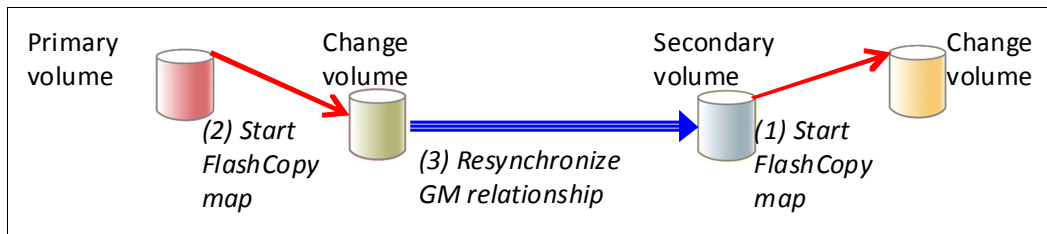


Figure 5-18 Global Mirror with Change Volumes

Global Mirror with Change Volumes also only sends one copy of a changed grain that might have been rewritten many times within the cycle period.

If the primary volume fails completely for any reason, GM/CV is designed to ensure that the secondary volume holds the same data as the primary did at a specific point in time. That period of data loss is typically between 5 minutes and 24 hours, but varies according to the design choices that you make.

Change Volumes hold point-in-time copies of 256 KB grains. If any of the disk blocks in a grain change, that grain is copied to the change volume to preserve its contents. Change Volumes are also maintained at the secondary site so that a consistent copy of the volume is always available even when the secondary volume is being updated.

Primary and Change Volumes are always in the same I/O group and the Change Volumes are always thin-provisioned. Change Volumes cannot be mapped to hosts and used for host I/O, and they cannot be used as a source for any other FlashCopy or Global Mirror operations.

Figure 5-19 shows how a Change Volume is used to preserve a point-in-time data set, which is then replicated to a secondary site. The data at the secondary site is in turn preserved by a Change Volume until the next replication cycle has completed.

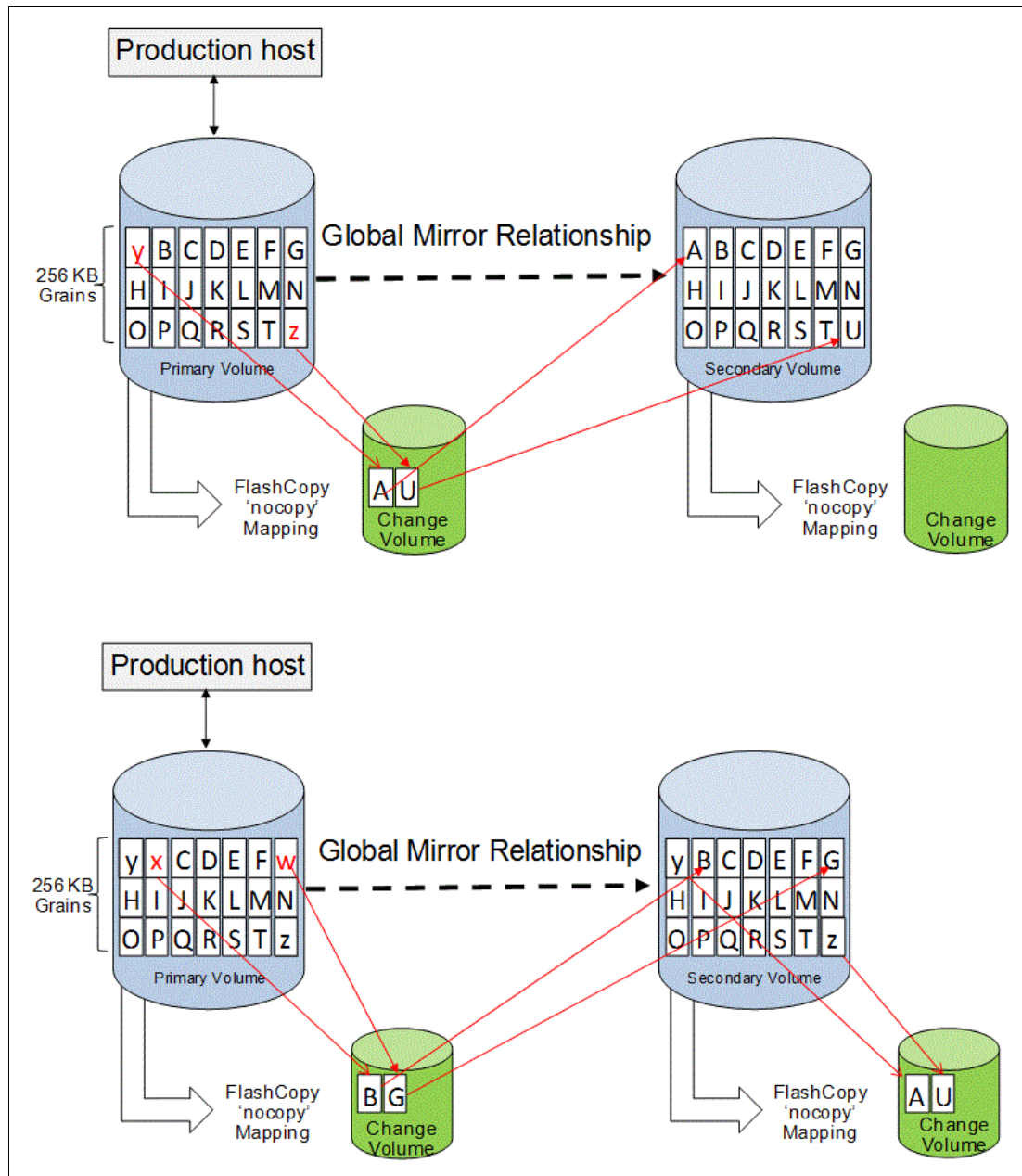


Figure 5-19 Global Mirror with Change Volumes uses FlashCopy point-in-time copy technology

FlashCopy mapping note: These FlashCopy mappings are not standard FlashCopy volumes and are not accessible for general use. They are internal structures that are dedicated to supporting Global Mirror with Change Volumes.

The options for `-cyclingmode` are `none` and `multi`.

Specifying or taking the default `none` means that Global Mirror acts in its traditional mode without Change Volumes.

Specifying `multi` means that Global Mirror starts cycling based on the cycle period, which defaults to 300 seconds. The valid range is from 60 seconds to 24*60*60 seconds (86,400 seconds = one day).

If all of the changed grains cannot be copied to the secondary site within the specified time, then the replication is designed to take as long as it needs and to start the next replication as soon as the earlier one completes. You can choose to implement this approach by deliberately setting the cycle period to a short amount of time, which is a perfectly valid approach. However, remember that the shorter the cycle period, the less opportunity there is for peak write I/O smoothing, and the more bandwidth you need.

The `-cyclingmode` setting can only be changed when the Global Mirror relationship is in a stopped state.

Recovery point objective using Change Volumes

RPO is the maximum tolerable period in which data might be lost if you switch over to your secondary volume.

If a cycle completes within the specified cycle period, then the RPO is not more than 2x cycle long. However, if it does not complete within the cycle period, then the RPO is not more than the sum of the last two cycle times.

The current RPO can be determined by looking at the `1srcr1ationship` freeze time attribute. The freeze time is the time stamp of the last primary Change Volume that has completed copying to the secondary site. Note the following example:

1. The cycle period is the default of 5 minutes and a cycle is triggered at 6:00 AM. At 6:03 AM, the cycle completes. The freeze time would be 6:00 AM, and the RPO is 3 minutes.
2. The cycle starts again at 6:05 AM. The RPO now is 5 minutes. The cycle is still running at 6:12 AM, and the RPO is now up to 12 minutes because 6:00 AM is still the freeze time of the last complete cycle.
3. At 6:13 AM, the cycle completes and the RPO now is 8 minutes because 6:05 AM is the freeze time of the last complete cycle.
4. Because the cycle period has been exceeded, the cycle immediately starts again.

5.3.2 Remote Copy network planning

Remote copy partnerships and relationships do not work reliably if the connectivity on which they are running is configured incorrectly. This section focuses on the intersystem network, giving an overview of the remote system connectivity options.

Terminology

The intersystem network is specified in terms of *latency* and *bandwidth*. These parameters define the capabilities of the link regarding the traffic that is on it. They must be chosen so that they support all forms of traffic, including mirrored foreground writes, background copy writes, and intersystem heartbeat messaging (node-to-node communication).

Link latency is the time that is taken by data to move across a network from one location to another and is measured in milliseconds. The longer the time, the greater the performance impact.

Tip: SCSI write over FC requires two round trips per I/O operation, as shown in the following example:

$2 \text{ (round trips)} \times 2 \text{ (operations)} \times 5 \text{ microsec/km} = 20 \text{ microsec/km}$

At 50 km, you have another latency, as shown in the following example:

$20 \text{ microsec/km} \times 50 \text{ km} = 1000 \text{ microsec} = 1 \text{ msec}$ (msec represents millisecond)

Each SCSI I/O has 1 ms of more service time. At 100 km, it becomes 2 ms for more service time.

Link bandwidth is the network capacity to move data as measured in millions of bits per second (Mbps) or billions of bits per second (Gbps).

The term *bandwidth* is also used in the following context:

- ▶ Storage bandwidth: The ability of the back-end storage to process I/O. Measures the amount of data (in bytes) that can be sent in a specified amount of time.
- ▶ Remote copy partnership bandwidth (parameter): The rate at which background write synchronization is attempted (unit of MBps).

Intersystem connectivity supports mirrored foreground and background I/O. A portion of the link is also used to carry traffic that is associated with the exchange of low-level messaging between the nodes of the local and remote systems. A *dedicated amount* of the link bandwidth is required for the exchange of heartbeat messages and the initial configuration of intersystem partnerships.

Interlink bandwidth must support the following traffic:

- ▶ Mirrored foreground writes, as generated by foreground processes at peak times
- ▶ Background write synchronization, as defined by the Global Mirror bandwidth parameter
- ▶ Intersystem communication (*heartbeat messaging*)

Fibre Channel connectivity is the standard connectivity that is used for the remote copy intersystem networks. It uses the Fibre Channel protocol and SAN infrastructures to interconnect the systems.

Native IP connectivity has been introduced with IBM Spectrum Virtualize version 7.2 to implement intersystem networks by using standard TPC/IP infrastructures.

Network latency considerations

The maximum supported round-trip latency between sites depends on the type of partnership between systems, the version of software, and the system hardware that is used. Table 5-6 on page 179 lists the maximum round-trip latency. This restriction applies to all variants of remote mirroring.

Table 5-6 Maximum round trip

IBM Spectrum Virtualize version	System node hardware	Partnership		
		FC	1 Gbps IP	10 Gbps IP
7.3 or earlier	All	80 ms	80 ms	10 ms
7.4 or later	CG8 nodes (with a second four-port Fibre Channel adapter installed) DH8 and SV1 nodes	250 ms		
	All other models	80 ms		

More configuration requirements and guidelines apply to systems that perform remote mirroring over extended distances, where the round-trip time is greater than 80 ms. If you use remote mirroring between systems with 80 - 250 ms round-trip latency, you must meet the following additional requirements:

- ▶ The RC buffer size setting must be 512 MB on each system in the partnership. This setting can be accomplished by running the `chsystem -rcbuffersize 512` command on each system.

Important: Changing this setting is disruptive to Metro Mirror and Global Mirror operations. Use this command only before partnerships are created between systems, or when all partnerships with the system are stopped.

- ▶ Two Fibre Channel ports on each node that will be used for replication must be dedicated for replication traffic. This configuration can be achieved by using SAN zoning and port masking.
- ▶ SAN zoning should be applied to provide separate intrasystem zones for each local-remote I/O group pair that is used for replication. See “Remote system ports and zoning considerations” on page 185 for further zoning guidelines.

Link bandwidth that is used by internode communication

IBM Spectrum Virtualize uses part of the bandwidth for its internal intersystem heartbeat. The amount of traffic depends on how many nodes are in each of the local and remote systems. Table 5-7 shows the amount of traffic (in megabits per second) that is generated by different sizes of systems.

Table 5-7 IBM Spectrum Virtualize intersystem heartbeat traffic (megabits per second)

Local or remote system	Two nodes	Four nodes	Six nodes	Eight nodes
Two nodes	5	6	6	6
Four nodes	6	10	11	12
Six nodes	6	11	16	17
Eight nodes	6	12	17	21

These numbers represent the total traffic between the two systems when *no* I/O is occurring to a mirrored volume on the remote system. Half of the data is sent by one system, and half of the data is sent by the other system. The traffic is divided evenly over all available connections. Therefore, if you have two redundant links, half of this traffic is sent over each link during fault-free operation.

If the link between the sites is configured with redundancy to tolerate single failures, size the link so that the bandwidth and latency statements continue to be accurate even during single failure conditions.

Network sizing considerations

Proper network sizing is essential for the remote copy services operations. Failing to estimate the network sizing requirements can lead to poor performance in remote copy services and the production workload.

Consider that intersystem bandwidth should be capable of supporting the combined traffic of the following items:

- ▶ Mirrored foreground writes, as generated by your server applications at peak times
- ▶ Background resynchronization, for example, after a link outage
- ▶ Inter-system heartbeat

Calculating the required bandwidth is essentially a question of mathematics based on your current workloads, so it is advisable to start by assessing your current workloads.

For Metro or Global Mirror, you need to know your peak write rates and I/O sizes down to at least a 5-minute interval. This information can be easily gained from tools like IBM Spectrum Control. Finally, you need to allow for unexpected peaks.

There are also unsupported tools to help with sizing available from IBM:

<http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD105947>

Do not compromise on bandwidth or network quality when planning a Metro or Global Mirror deployment. If bandwidth is likely to be an issue in your environment, consider Global Mirror with Change Volumes.

Bandwidth sizing examples

As an example, consider a business with the following I/O profile:

- ▶ Average write size 8 KB (= 8 x 8 bits/1024 = 0.0625 Mb).
- ▶ For most of the day between 8 AM and 8 PM, the write activity is around 1500 writes per second.
- ▶ Twice a day (once in the morning and once in the afternoon), the system bursts up to 4500 writes per second for up to 10 minutes.
- ▶ Outside of the 8 AM to 8 PM window, there is little or no I/O write activity.

This example is intended to represent a general traffic pattern that might be common in many medium-sized sites. Furthermore, 20% of bandwidth must be left available for the background synchronization.

Here we consider options for Metro Mirror, Global Mirror, and for Global Mirror with Change Volumes based on a cycle period of 30 minutes and 60 minutes.

Metro Mirror or Global Mirror require bandwidth on the instantaneous peak of 4500 writes per second as follows:

$4500 \times 0.0625 = 282 \text{ Mbps} + 20\% \text{ resync allowance} + 5 \text{ Mbps heartbeat} = 343 \text{ Mbps}$
dedicated plus any safety margin plus growth

In the following two examples, the bandwidth for GM/CV needs to be able to handle the peak 30-minute period, or the peak 60-minute period.

GMCV peak 30-minute period example

If we look at this time broken into 10-minute periods, the peak 30-minute period is made up of one 10-minute period of 4500 writes per second, and two 10-minute periods of 1500 writes per second. The average write rate for the 30-minute cycle period can then be expressed mathematically as follows:

$$(4500 + 1500 + 1500) / 3 = 2500 \text{ writes/sec for a 30-minute cycle period}$$

The minimum bandwidth that is required for the cycle period of 30 minutes is as follows:

$$2500 \times 0.0625 = 157 \text{ Mbps} + 20\% \text{ resync allowance} + 5 \text{ Mbps heartbeat} = 195 \text{ Mbps}$$

dedicated plus any safety margin plus growth

GMCV peak 60-minute period example

For a cycle period of 60 minutes, the peak 60-minute period is made up of one 10-minute period of 4500 writes per second, and five 10-minute periods of 1500 writes per second. The average write for the 60-minute cycle period can be expressed as follows:

$$(4500 + 5 \times 1500) / 6 = 2000 \text{ writes/sec for a 60-minute cycle period}$$

The minimum bandwidth that is required for a cycle period of 60 minutes is as follows:

$$2000 \times 0.0625 = 125 \text{ Mbps} + 20\% \text{ resync allowance} + 5 \text{ Mbps heartbeat} = 155 \text{ Mbps}$$

dedicated plus any safety margin plus growth

Now consider whether the business does not have aggressive RPO requirements and does not want to provide dedicated bandwidth for Global Mirror. But the network is available and unused at night, so Global Mirror can use that. There is an element of risk here, which is if the network is unavailable for any reason, GM/CV cannot keep running during the day until it catches up. Therefore, you would need to allow a much higher resync allowance in your replication window, for example, 100 percent.

A GM/CV replication based on daily point-in-time copies at 8 PM each night, and replicating until 8 AM at the latest would probably require at least the following bandwidth:

$$(9000 + 70 \times 1500) / 72 = 1584 \times 0.0625 = 99 \text{ Mbps} + 100\% + 5 \text{ Mbps heartbeat}$$

= 203 Mbps at night plus any safety margin plus growth, non-dedicated, time-shared with daytime traffic

Global Mirror with Change Volumes provides a way to maintain point-in-time copies of data at a secondary site where insufficient bandwidth is available to replicate the peak workloads in real time.

Another factor that can reduce the bandwidth that is required for Global Mirror with Change Volumes is that it only sends one copy of a changed grain, which might have been rewritten many times within the cycle period.

Remember that these are examples. The central principle of sizing is that you need to know your data write rate, which is the number of write I/Os and the average size of those I/Os. For Metro Mirror and Global Mirror, you need to know the peak write I/O rates. For GM/CV, you need to know the average write I/O rates.

GMCV bandwidth: In the above samples, the bandwidth estimation for the GMCV is based on the assumption that the write operations occurs in such a way that a change volume grain (that has a size of 256 KB) is completely changed before it is transferred to the remote site. In the real life, this situation is unlikely to occur.

Usually only a portion of a grain is changed during a GMCV cycle, but the transfer process always copies the whole grain to the remote site. This behavior can lead to an unforeseen processor burden in the transfer bandwidth that, in the edge case, can be even higher than the one required for a standard Global Mirror.

Fibre Channel connectivity

You must remember several considerations when you use Fibre Channel technology for the intersystem network:

- ▶ Redundancy
- ▶ Basic topology and problems
- ▶ Switches and ISL oversubscription
- ▶ Distance extensions options
- ▶ Optical multiplexors
- ▶ Long-distance SFPs and XFPs
- ▶ Fibre Channel over IP
- ▶ Hops
- ▶ Buffer credits
- ▶ Remote system ports and zoning considerations

Redundancy

The intersystem network must adopt the same policy toward redundancy as for the local and remote systems to which it is connecting. The ISLs must have redundancy, and the individual ISLs must provide the necessary bandwidth in isolation.

Basic topology and problems

Because of the nature of Fibre Channel, you must avoid ISL congestion whether within individual SANs or across the intersystem network. Although FC (and IBM Spectrum Virtualize) can handle an overloaded host or storage array, the mechanisms in FC are ineffective for dealing with congestion in the fabric in most circumstances. The problems that are caused by fabric congestion can range from dramatically slow response time to storage access loss. These issues are common with all high-bandwidth SAN devices and are inherent to FC. They are not unique to the IBM Spectrum Virtualize/Storwize products.

When an FC network becomes congested, the FC switches stop accepting more frames until the congestion clears. They can also drop frames. Congestion can quickly move upstream in the fabric and clog the end devices from communicating anywhere.

This behavior is referred to as *head-of-line blocking*. Although modern SAN switches internally have a nonblocking architecture, head-of-line-blocking still exists as a SAN fabric problem. Head-of-line blocking can result in IBM Spectrum Virtualize nodes that cannot communicate with storage subsystems or to mirror their write caches because you have a single congested link that leads to an edge switch.

Switches and ISL oversubscription

As specified in Chapter 2, “Back-end storage” on page 49, the suggested maximum host port to ISL ratio is 7:1. With modern 8 Gbps or 16 Gbps SAN switches, this ratio implies an average bandwidth (in one direction) per host port of approximately 230 MBps (16 Gbps).

You must take peak loads (not average loads) into consideration. For example, while a database server might use only 20 MBps during regular production workloads, it might perform a backup at higher data rates.

Congestion to one switch in a large fabric can cause performance issues throughout the entire fabric, including traffic between IBM Spectrum Virtualize nodes and storage subsystems, even if they are not directly attached to the congested switch. The reasons for these issues are inherent to FC flow control mechanisms, which are not designed to handle fabric congestion. Therefore, any estimates for required bandwidth before implementation must have a safety factor that is built into the estimate.

On top of the safety factor for traffic expansion, implement a spare ISL or ISL trunk. The spare ISL or ISL trunk can provide a fail-safe that avoids congestion if an ISL fails because of issues, such as a SAN switch line card or port blade failure.

Exceeding the standard 7:1 oversubscription ratio requires you to implement fabric bandwidth threshold alerts. When one of your ISLs exceeds 70%, you must schedule fabric changes to distribute the load further.

You must also consider the bandwidth consequences of a complete fabric outage. Although a complete fabric outage is a fairly rare event, insufficient bandwidth can turn a single-SAN outage into a total access loss event.

Take the bandwidth of the links into account. It is common to have ISLs run faster than host ports, which reduces the number of required ISLs.

Distance extensions options

To implement remote mirroring over a distance by using the Fibre Channel, you have the following choices:

- ▶ Optical multiplexors, such as dense wavelength division multiplexing (DWDM) or coarse wavelength division multiplexing (CWDM) devices
- ▶ Long-distance Small Form-factor Pluggable (SFP) transceivers and XFPs
- ▶ Fibre Channel-to-IP conversion boxes

Of these options, the optical distance extension is the preferred method. IP distance extension introduces more complexity, is less reliable, and has performance limitations. However, optical distance extension can be impractical in many cases because of cost or unavailability.

For the list of supported SAN routers and FC extenders, see the support page at this website:

<https://ibm.biz/BdiZa6>

Optical multiplexors

Optical multiplexors can extend a SAN up to hundreds of kilometers (or miles) at high speeds. For this reason, they are the preferred method for long-distance expansion. If you use multiplexor-based distance extension, closely monitor your physical link error counts in your switches. Optical communication devices are high-precision units. When they shift out of calibration, you will start to see errors in your frames.

Long-distance SFPs and XFPs

Long-distance optical transceivers have the advantage of extreme simplicity. You do not need any expensive equipment, and you have only a few configuration steps to perform. However, ensure that you only use transceivers that are designed for your particular SAN switch.

Fibre Channel over IP

Fibre Channel over IP (FCIP) is by far the most common and least expensive form of distance extension. It is also complicated to configure. Relatively subtle errors can have severe performance implications.

With IP-based distance extension, you must dedicate bandwidth to your FCIP traffic if the link is shared with other IP traffic. Do not assume that because the link between two sites has low traffic or is used only for email, this type of traffic is always the case. FC is far more sensitive to congestion than most IP applications.

Also, when you are communicating with the networking architects for your organization, make sure to distinguish between *megabytes per second* as opposed to *megabits per second*. In the storage world, bandwidth often is specified in megabytes per second (MBps), and network engineers specify bandwidth in megabits per second (Mbps).

Hops

The hop count is not increased by the intersite connection architecture. For example, if you have a SAN extension that is based on DWDM, the DWDM components are not apparent to the number of hops. The hop count limit within a fabric is set by the fabric devices (switch or director) operating system. It is used to derive a frame hold time value for each fabric device.

This hold time value is the maximum amount of time that a frame can be held in a switch before it is dropped or the fabric is busy condition is returned. For example, a frame might be held if its destination port is unavailable. The hold time is derived from a formula that uses the error detect timeout value and the resource allocation timeout value. It is considered that every extra hop adds about 1.2 microseconds of latency to the transmission.

Currently, IBM Spectrum Virtualize and Storwize remote copy services support three hops when protocol conversion exists. Therefore, if you have DWDM extended between primary and secondary sites, three SAN directors or switches can exist between the primary and secondary systems.

Buffer credits

SAN device ports need memory to temporarily store frames as they arrive, assemble them in sequence, and deliver them to the upper layer protocol. The number of frames that a port can hold is called its *buffer credit*. Fibre Channel architecture is based on a flow control that ensures a constant stream of data to fill the available pipe.

When two FC ports begin a conversation, they exchange information about their buffer capacities. An FC port sends only the number of buffer frames for which the receiving port gives credit. This method avoids overruns and provides a way to maintain performance over distance by filling the pipe with in-flight frames or buffers.

The following types of transmission credits are available:

- ▶ **Buffer_to_Buffer Credit**

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

- ▶ **End_to_End Credit**

In the same way during login, all N_Ports establish End-to-End Credit (EE_Credit) with each other. During data transmission, a port must not send more frames than the buffer of the receiving port can handle before you receive an indication from the receiving port that it processed a previously sent frame. Two counters are used: BB_Credit_CNT and EE_Credit_CNT. Both counters are initialized to zero during login.

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

At any time, if BB_Credit_CNT becomes equal to the BB_Credit, or EE_Credit_CNT becomes equal to the EE_Credit of the receiving port, the transmitting port stops sending frames until the respective count is decremented.

The previous statements are true for Class 2 service. Class 1 is a dedicated connection. Therefore, BB_Credit is not important, and only EE_Credit is used (EE Flow Control). However, Class 3 is an unacknowledged service. Therefore, it uses only BB_Credit (BB Flow Control), but the mechanism is the same in all cases.

Here, you see the importance that the number of buffers has in overall performance. You need enough buffers to ensure that the transmitting port can continue to send frames without stopping to use the full bandwidth, which is true with distance. The total amount of buffer credit needed to optimize the throughput depends on the link speed and the average frame size.

For example, consider an 8 Gbps link connecting two switches that are 100 km apart. At 8 Gbps, a full frame (2148 bytes) occupies about 0.51 km of fiber. In a 100 km link, you can send 198 frames before the first one reaches its destination. You need an ACK to go back to the start to fill EE_Credit again. You can send another 198 frames before you receive the first ACK.

You need at least 396 buffers to allow for nonstop transmission at 100 km distance. The maximum distance that can be achieved at full performance depends on the capabilities of the FC node that is attached at either end of the link extenders, which are vendor-specific. A match should occur between the buffer credit capability of the nodes at either end of the extenders.

Remote system ports and zoning considerations

Ports and zoning requirements for the remote system partnership have changed over time. The current preferred configuration is described in the following Flash Alert:

<https://www.ibm.com/support/docview.wss?uid=ssg1S1003634>

The preferred practice for the IBM Spectrum Virtualize and Storwize systems is to provision dedicated node ports for local node-to-node traffic (by using port masking) and isolate Global Mirror node-to-node traffic between the local nodes from other local SAN traffic.

Remote port masking: To isolate the node-to-node traffic from the remote copy traffic, the local and remote port masking implementation is preferable.

This configuration of local node port masking is less of a requirement on non-clustered Storwize systems, where traffic between node canisters in an I/O group is serviced by the dedicated PCI inter-canister link in the enclosure. The following guidelines apply to the remote system connectivity:

- ▶ The minimum requirement to establish a remote copy partnership is to connect at least one node per system. When remote connectivity among all the nodes of both systems is not available, the nodes of the local system not participating to the remote partnership will use the node/nodes defined in the partnership as a bridge to transfer the replication data to the remote system.

This replication data transfer occurs through the node-to-node connectivity. Note that this configuration, even though supported, allows the replication traffic to go through the node-to-node connectivity and this is not recommended.

- ▶ Partnered systems should use the same number of nodes in each system for replication.
- ▶ For maximum throughput, all nodes in each system should be used for replication, both in terms of balancing the preferred node assignment for volumes and for providing intersystem Fibre Channel connectivity.
- ▶ Where possible, use the minimum number of partnerships between systems. For example, assume site A contains systems A1 and A2, and site B contains systems B1 and B2. In this scenario, creating separate partnerships between pairs of systems (such as A1-B1 and A2-B2) offers greater performance for Global Mirror replication between sites than a configuration with partnerships defined between all four systems.

For zoning, the following rules for the remote system partnership apply:

- ▶ For Metro Mirror and Global Mirror configurations where the round-trip latency between systems is less than 80 milliseconds, zone two Fibre Channel ports on each node in the local system to two Fibre Channel ports on each node in the remote system.
- ▶ For Metro Mirror and Global Mirror configurations where the round-trip latency between systems is more than 80 milliseconds, apply SAN zoning to provide separate intrasystem zones for each local-remote I/O group pair that is used for replication, as shown in Figure 5-20.

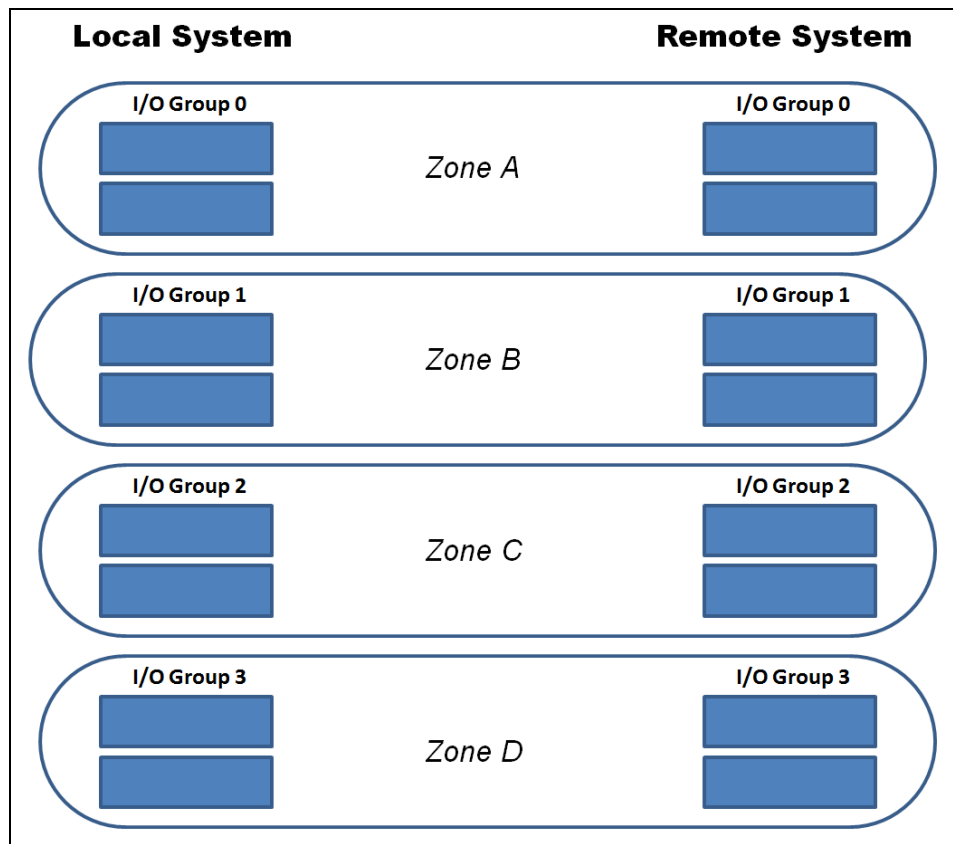


Figure 5-20 Zoning scheme for >80 ms remote copy partnerships

NPIV: IBM Spectrum Virtualize and Storwize systems with the NPIV feature enabled provide virtual WWPN for the host zoning. Those WWPNs are intended for host zoning only and can not be used for the remote copy partnership.

Native IP connectivity

Remote Mirroring over IP communication is supported on the IBM Spectrum Virtualize and Storwize Family systems by using Ethernet communication links. The IBM Spectrum Virtualize Software IP replication uses innovative Bridgeworks SANSlide technology to optimize network bandwidth and utilization.

This new function enables the use of a lower-speed and lower-cost networking infrastructure for data replication. Bridgeworks' SANSlide technology, which is integrated into the IBM Spectrum Virtualize Software, uses artificial intelligence to help optimize network bandwidth use and adapt to changing workload and network conditions. This technology can improve remote mirroring network bandwidth usage up to three times, which can enable clients to deploy a less costly network infrastructure, or speed up remote replication cycles to enhance disaster recovery effectiveness.

The native IP replication is covered in detail in 5.4, "Native IP replication" on page 214.

5.3.3 Remote Copy services planning

When you plan for remote copy services, you must keep in mind the considerations that are outlined in the following sections.

Remote Copy configurations limits

To plan for and implement remote copy services, you must check the configuration limits and adhere to them. Table 5-8 shows the limits for a system that apply to IBM Spectrum Virtualize V7.8 (still valid with V8.1 for the system supported).

Table 5-8 Remote copy maximum limits

Remote copy property	Maximum	Apply to	Comment
Remote Copy (Metro Mirror and Global Mirror) relationships per system	10000	<ul style="list-style-type: none"> ▶ SAN Volume Controller models SV1, DH8, CG8, and CF8 ▶ Storwize V7000 models 524 (Gen2) and 624 (Gen2+) 	This configuration can be any mix of Metro Mirror and Global Mirror relationships.
	8192	Any other Storwize model	This configuration can be any mix of Metro Mirror and Global Mirror relationships. Maximum requires an 8-node system (volumes per I/O group limit applies).

Remote copy property	Maximum	Apply to	Comment
Active-Active Relationships	1250	<ul style="list-style-type: none"> ▶ SAN Volume Controller models SV1, DH8, CG8, and CF8 ▶ Storwize V7000 models 524 (Gen2) and 624 (Gen2+) 	This is the limit for the number of HyperSwap volumes in a system.
	1024	Any other Storwize model	This is the limit for the number of HyperSwap volumes in a system.
Remote Copy relationships per consistency group	None	All models	No limit is imposed beyond the Remote Copy relationships per system limit.
Remote Copy consistency groups per system	256	All models	
Total Metro Mirror and Global Mirror volume capacity per I/O group	1024 TB	All models	This limit is the total capacity for all master and auxiliary volumes in the I/O group.
Total number of Global Mirror with Change Volumes relationships per system	256	All models	
Inter-system IP partnerships per system	1	All models	A system can be partnered with up to three remote systems. A maximum of one of those can be IP and the other two FC.
I/O groups per system in IP partnerships	2	All models	The nodes from a maximum of two I/O groups per system can be used for IP partnership.
Inter site links per IP partnership	2	All models	A maximum of two inter site links can be used between two IP partnership sites.
Ports per node	1	All models	A maximum of one port per node can be used for IP partnership.
IP partnership Software Compression Limit	70 MBps	<ul style="list-style-type: none"> ▶ SAN Volume Controller models CG8 and CF8 ▶ Storwize V7000 model 124 (Gen1) 	
	140 MBps	<ul style="list-style-type: none"> ▶ SAN Volume Controller models SV1 and DH8 ▶ Storwize V7000 models 524 (Gen2) and 624 (Gen2+) 	

Similar to FlashCopy, the remote copy services require memory to allocate the bitmap structures used to track the updates while volume are suspended or synchronizing. The default amount of memory for remote copy services is 20 MB. This value can be increased or decreased by using the **chiogrp** command. The maximum amount of memory that can be specified for remote copy services is 512 MB. The grain size for the remote copy services is 256 KB.

Remote Copy general restrictions

To use Metro Mirror and Global Mirror, you must adhere to the following rules:

- ▶ You must have the same size for source and target volume when defining a remote copy relationship. However, the target volume can be a different type (image, striped, or sequential mode) or have different cache settings (cache-enabled or cache-disabled).
- ▶ You cannot move Metro Mirror or Global Mirror source or target volumes to different I/O groups.
- ▶ Metro Mirror and Global Mirror volumes can be resized with the following restrictions:
 - The code level must be V7.8.1 or later.
 - The volumes must be thin-provisioned or compressed.
 - Apply to Metro Mirror and Global Mirror only, GMCV is not supported.
 - The Remote Copy Consistency Protection feature is not allowed and must be removed before resizing the volumes.
 - No active FlashCopy allowed.
 - The remote copy relationship must be in synchronized status.
 - The resize order must guarantee the target volume to be always larger than the source volume.
- ▶ You can mirror intrasystem Metro Mirror or Global Mirror only between volumes in the same I/O group.

Intrasystem remote copy: The intrasystem remote copy is not supported on IBM Spectrum Virtualize/Storwize systems running version 6 or later.

- ▶ Global Mirror is not recommended for cache-disabled volumes that are participating in a Global Mirror relationship.

Changing the remote copy type

Changing the remote copy type for an existing relationship is quite an easy task. It is enough to stop the relationship, if it is active, and change the properties to set the new remote copy type. Do not forget to create the change volumes in case of change from Metro or Global Mirror to Global Mirror Change Volumes.

Interaction between Remote Copy and FlashCopy

Remote Copy functions can be used in conjunction with the Flash Copy function so that you can have both operating concurrently on the same volume. The possible combinations between Remote Copy and FlashCopy follow:

- ▶ Remote copy source:
 - A remote copy source can be a FlashCopy source.
 - A remote copy source can be a FlashCopy target with the following restrictions:
 - IBM Spectrum Virtualize/Storwize must be V6.2 or later.
 - A FlashCopy target volume cannot be updated while it is the source volume of a Metro or Global Mirror relationship that is actively mirroring. A FlashCopy mapping cannot be started while the target volume is in an active remote copy relationship.
 - The I/O group for the FlashCopy mappings must be the same as the I/O group for the FlashCopy target volume (that is the I/O group of the Remote copy source).

- ▶ Remote copy target:
 - A remote copy target can be a FlashCopy source.
 - A remote copy target can be a FlashCopy target with the following restrictions:
 - A FlashCopy mapping must be in the `idle_copied` state when its target volume is the target volume of an active Metro Mirror or Global Mirror relationship.

Native backend controller copy functions considerations

As discussed in the above sections, the IBM Spectrum Virtualize/Storwize technology provides a widespread set of copy services functions that cover most of the clients requirements.

However, some storage controllers can provide specific copy services capabilities not available with the current version of IBM Spectrum Virtualize software. The IBM Spectrum Virtualize/Storwize technology addresses these situations by using cache-disabled image mode volumes that virtualize LUN participating to the native backend controller's copy services relationships.

Keeping the cache disabled guarantees data consistency throughout the I/O stack, from the host to the backend controller. Otherwise, by leaving the cache enabled on a volume, the underlying controller does not receive any write I/Os as the host writes them. IBM Spectrum Virtualize caches them and processes them later. This process can have more ramifications if a target host depends on the write I/Os from the source host as they are written.

Note: Native copy services are not supported on all storage controllers. For more information about the known limitations, see *Using Native Controller Copy Services*, S1002852, at this website:

<http://www.ibm.com/support/docview.wss?&uid=ssg1S1002852>

As part of its copy services function, the storage controller might take a LUN offline or suspend reads or writes. As IBM Spectrum Virtualize/Storwize does not recognize why this happens; therefore, it might log errors when these events occur. For this reason, if the IBM Spectrum Virtualize/Storwize must detect the LUN, ensure to keep that LUN in the unmanaged state until full access is granted.

Native backend controller copy services can also be used for LUNs not managed by the IBM Spectrum Virtualize/Storwize. Note that accidental incorrect configurations of the backend controller copy services involving IBM Spectrum Virtualize/Storwize attached LUN can produce unpredictable results.

For example, if you accidentally use a LUN with IBM Spectrum Virtualize/Storwize data on it as a point-in-time target LUN, you can corrupt that data. Moreover, if that LUN was a managed disk in a managed disk group with striped or sequential volumes on it, the managed disk group might be brought offline. This situation, in turn, makes all of the volumes that belong to that group go offline, leading to a widespread host access disruption.

Remote Copy and code upgrade considerations

When you upgrade system software where the system participates in one or more intersystem relationships, upgrade only one cluster at a time. That is, do not upgrade the systems concurrently.

Attention: Upgrading both systems concurrently is not monitored by the software upgrade process.

Allow the software upgrade to complete one system before it is started on the other system. Upgrading both systems concurrently can lead to a loss of synchronization. In stress situations, it can further lead to a loss of availability.

Usually, pre-existing remote copy relationships are unaffected by a software upgrade that is performed correctly. However, always check in the target code release notes for special considerations on the copy services.

Even if it is not a best practice, a remote copy partnership can be established, with some restriction, among systems with different IBM Spectrum Virtualize versions. For more information about a compatibility table for intersystem Metro Mirror and Global Mirror relationships between IBM Spectrum Virtualize code levels, see *SAN Volume Controller Inter-system Metro Mirror and Global Mirror Compatibility Cross Reference*, S1003646. This publication is available at this website:

<http://www.ibm.com/support/docview.wss?rs=591&uid=ssg1S1003646>

Volume placement considerations

You can optimize the distribution of volumes within I/O groups at the local and remote systems to maximize performance.

Although defined at a system level, the partnership bandwidth, and consequently the background copy rate, is evenly divided among the cluster's I/O groups. The available bandwidth for the background copy can be used by either node, or shared by both nodes within the I/O Group.

This bandwidth allocation is independent from the number of volumes for which a node is responsible. Each node, in turn, divides its bandwidth evenly between the (multiple) remote copy relationships with which it associates volumes that are performing a background copy.

Volume preferred node

Conceptually, a connection (path) goes between each node on the primary system to each node on the remote system. Write I/O, which is associated with remote copying, travels along this path. Each node-to-node connection is assigned a finite amount of remote copy resource and can sustain only in-flight write I/O to this limit.

The node-to-node in-flight write limit is determined by the number of nodes in the remote system. The more nodes that exist at the remote system, the lower the limit is for the in-flight write I/Os from a local node to a remote node. That is, less data can be outstanding from any one local node to any other remote node. Therefore, to optimize performance, Global Mirror volumes must have their preferred nodes distributed evenly between the nodes of the systems.

The preferred node property of a volume helps to balance the I/O load between nodes in that I/O group. This property is also used by remote copy to route I/O between systems.

The IBM Spectrum Virtualize node/Storwize canister that receives a write for a volume is normally the preferred node of the volume. For volumes in a remote copy relationship, that node is also responsible for sending that write to the preferred node of the target volume. The primary preferred node is also responsible for sending any writes that relate to the background copy. Again, these writes are sent to the preferred node of the target volume.

Each node of the remote system has a fixed pool of remote copy system resources for *each node* of the primary system. That is, each remote node has a separate queue for I/O from each of the primary nodes. This queue is a fixed size and is the same size for every node. If preferred nodes for the volumes of the remote system are set so that every combination of primary node and secondary node is used, remote copy performance is maximized.

Figure 5-21 shows an example of remote copy resources that are not optimized. Volumes from the local system are replicated to the remote system. All volumes with a preferred node of node 1 are replicated to the remote system, where the target volumes also have a preferred node of node 1.

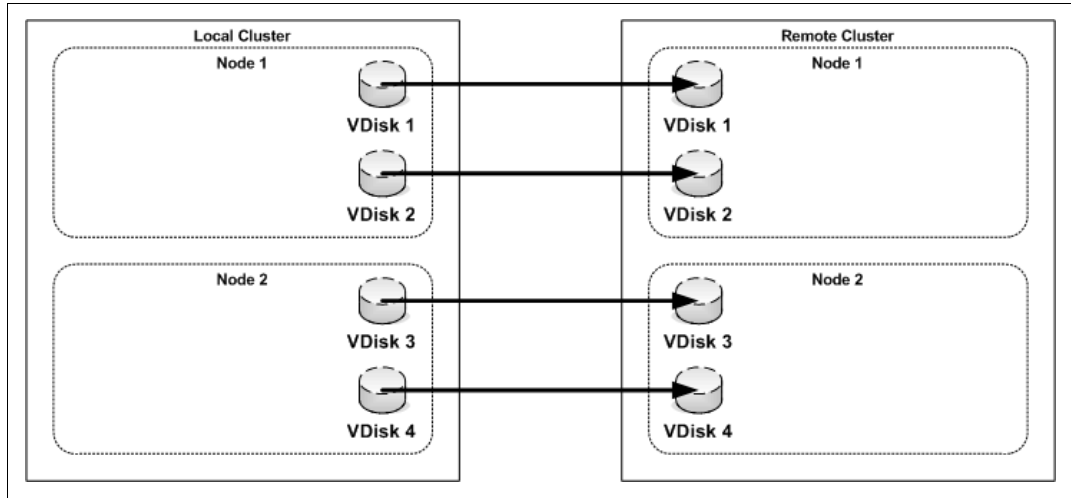


Figure 5-21 Remote copy resources that are not optimized

With this configuration, the resources for remote system node 1 that are reserved for local system node 2 are not used. The resources for local system node 1 that are used for remote system node 2 also are not used.

If the configuration changes to the configuration that is shown in Figure 5-22, all remote copy resources for each node are used, and remote copy operates with better performance.

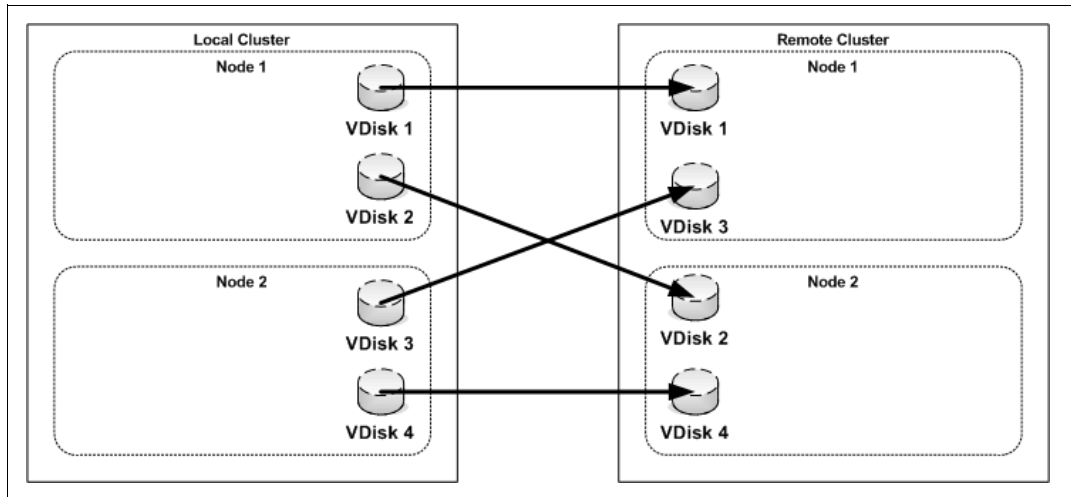


Figure 5-22 Optimized Global Mirror resources

Background copy considerations

The remote copy partnership bandwidth parameter *explicitly* defines the rate at which the background copy is attempted, but also *implicitly* affects foreground I/O. Background copy bandwidth can affect foreground I/O latency in one of the following ways:

- ▶ Increasing latency of foreground I/O

If the remote copy partnership bandwidth parameter is set too high for the actual intersystem network capability, the background copy resynchronization writes use too much of the intersystem network. It starves the link of the ability to service synchronous or asynchronous mirrored foreground writes. Delays in processing the mirrored foreground writes increase the latency of the foreground I/O as perceived by the applications.

- ▶ Read I/O overload of primary storage

If the remote copy partnership background copy rate is set too high, the added read I/Os that are associated with background copy writes can overload the storage at the primary site and delay foreground (read and write) I/Os.

- ▶ Write I/O overload of auxiliary storage

If the remote copy partnership background copy rate is set too high for the storage at the secondary site, the background copy writes overload the auxiliary storage. Again, they delay the synchronous and asynchronous mirrored foreground write I/Os.

Important: An increase in the peak foreground workload can have a detrimental effect on foreground I/O. It does so by pushing more mirrored foreground write traffic along the intersystem network, which might not have the bandwidth to sustain it. It can also overload the primary storage.

To set the background copy bandwidth optimally, consider all aspects of your environments, starting with the following biggest contributing resources:

- ▶ Primary storage
- ▶ Intersystem network bandwidth
- ▶ Auxiliary storage

Provision the most restrictive of these three resources between the background copy bandwidth and the peak foreground I/O workload. Perform this provisioning by calculation or by determining experimentally how much background copy can be allowed before the foreground I/O latency becomes unacceptable.

Then, reduce the background copy to accommodate peaks in workload. In cases where the available network bandwidth is not able to sustain an acceptable background copy rate, consider alternatives to the initial copy as described in “Initial synchronization options and Offline Synchronization” on page 194.

Changes in the environment, or loading of it, can affect the foreground I/O. IBM Spectrum Virtualize and Storwize technology provides a means to monitor, and a parameter to control, how foreground I/O is affected by running remote copy processes. IBM Spectrum Virtualize software monitors the delivery of the mirrored foreground writes. If latency or performance of these writes extends beyond a (predefined or client-defined) limit for a period, the remote copy relationship is suspended (see 5.3.5, “1920 error” on page 202).

Finally, note that with Global Mirror Change Volume, the cycling process that transfers the data from the local to the remote system is a background copy task. For this reason, the background copy rate, as well as the `relationship_bandwidth_limit`, setting affects the available bandwidth not only during the initial synchronization, but also during the normal cycling process.

Background copy bandwidth allocation: As already mentioned in “Volume placement considerations” on page 191, the available bandwidth of a remote copy partnership is evenly divided among the cluster's I/O Groups. In a case of unbalanced distribution of the remote copies among the I/O groups, the partnership bandwidth should be adjusted accordingly to reach the desired background copy rate.

Consider, for example, a 4-I/O groups cluster that has a partnership bandwidth of 4,000 Mbps and a background copy percentage of 50. The expected maximum background copy rate for this partnership is then 250MB/s. Having the available bandwidth evenly divided among the I/O groups, every I/O group in this cluster can theoretically synchronize data at a maximum rate of about 62 MBps (50% of 1,000 Mbps). Now in an edge case where only volumes from one I/O group are being replicated, in order to reach the full background copy rate (250 MBps) the partnership bandwidth should be adjusted to 16000 Mbps.

Initial synchronization options and Offline Synchronization

When creating a remote copy relationship, two options regarding the initial synchronization process are available:

- ▶ The not synchronized option is the default. With this option, when a remote copy relationship is started, a full data synchronization at the background copy rate occurs between the source and target volumes. It is the simplest approach in that it requires no other administrative activity apart from issuing the necessary IBM Spectrum Virtualize commands. However, in some environments, the available bandwidth makes this option unsuitable.
- ▶ The already synchronized option does not force any data synchronization when the relationship is started. The administrator must ensure that the source and target volumes contain identical data before a relationship is created. The administrator can perform this check in one of the following ways:
 - Create both volumes with the security delete feature to change all data to zero.
 - Copy a complete tape image (or other method of moving data) from one disk to the other.

In either technique, no write I/O must take place to the source and target volume before the relationship is established. The administrator must then complete the following actions:

- Create the relationship with the already synchronized settings (**-sync** option).
- Start the relationship.

Attention: If you do not perform these steps correctly, the remote copy reports the relationship as being *consistent*, when it is not. This setting is likely to make any auxiliary volume useless.

By understanding the methods to start a Metro Mirror and Global Mirror relationship, you can use one of them as a means to implement the remote copy relationship, save bandwidth, and resize the Global Mirror volumes.

Consider a situation where you have a large source volume (or many source volumes) containing already active data and that you want to replicate to a remote site. Your planning shows that the mirror initial sync time takes too long (or is too costly if you pay for the traffic that you use). In this case, you can set up the sync by using another medium that is less expensive. This synchronization method is called *Offline Synchronization*.

Another reason that you might want to use this method is if you want to increase the size of the volume that is in a Metro Mirror relationship or in a Global Mirror relationship. Increasing the size of these volumes may require a deletion and redefinition of the current mirror relationships when the requirements described in “Remote Copy general restrictions” on page 189 are not met.

This example uses tape media as the source for the initial sync for the Metro Mirror relationship or the Global Mirror relationship target before it uses remote copy services to maintain the Metro Mirror or Global Mirror. This example does not require downtime for the hosts that use the source volumes.

Before you set up Global Mirror relationships, save bandwidth, and resize volumes, complete the following steps:

1. Ensure that the hosts are up and running and are using their volumes normally. No Metro Mirror relationship nor Global Mirror relationship is defined yet.

Identify all of the volumes that become the source volumes in a Metro Mirror relationship or in a Global Mirror relationship.

2. Establish the IBM Spectrum Virtualize system partnership with the target IBM Spectrum Virtualize system.

To set up Global Mirror relationships, save bandwidth, and resize volumes, complete the following steps:

1. Define a Metro Mirror relationship or a Global Mirror relationship for each source disk. When you define the relationship, ensure that you use the **-sync** option, which stops the system from performing an initial sync.

Attention: If you do not use the **-sync** option, all of these steps are redundant because the IBM Spectrum Virtualize/Storwize system performs a full initial synchronization anyway.

2. Stop each mirror relationship by using the **-access** option, which enables write access to the target volumes. You need this write access later.
3. Copy the source volume to the alternative media by using the **dd** command to copy the contents of the volume to tape. Another option is to use your backup tool (for example, IBM Spectrum Protect™) to make an image backup of the volume.

Change tracking: Although the source is being modified while you are copying the image, the IBM Spectrum Virtualize/Storwize system is tracking those changes. The image that you create might have some of the changes and is likely to also miss some of the changes.

When the relationship is restarted, the IBM Spectrum Virtualize/Storwize system applies all of the changes that occurred since the relationship stopped in step 2. After all the changes are applied, you have a consistent target image.

4. Ship your media to the remote site and apply the contents to the targets of the Metro Mirror or Global Mirror relationship. You can mount the Metro Mirror and Global Mirror target volumes to a UNIX server and use the **dd** command to copy the contents of the tape to the target volume.

If you used your backup tool to make an image of the volume, follow the instructions for your tool to restore the image to the target volume. Remember to remove the mount if the host is temporary.

Tip: It does not matter how long it takes to get your media to the remote site and perform this step. However, the faster you can get the media to the remote site and load it, the quicker IBM Spectrum Virtualize/Storwize system starts running and maintaining the Metro Mirror and Global Mirror.

5. Unmount the target volumes from your host. When you start the Metro Mirror and Global Mirror relationship later, the IBM Spectrum Virtualize/Storwize system stops write access to the volume while the mirror relationship is running.
6. Start your Metro Mirror and Global Mirror relationships. The relationships must be started with the **-clean** parameter. In this way, any changes that are made on the secondary volume are ignored, and only changes made on the clean primary volume are considered when synchronizing the primary and secondary volumes.
7. While the mirror relationship catches up, the target volume is not usable at all. When it reaches `ConsistentSynchnonized` status, your remote volume is ready for use in a disaster.

Back-end storage considerations

To reduce the overall solution costs, it is a common practice to provide the remote systems with lower performance characteristics compared to the local system, especially when using asynchronous remote copy technologies. This attitude can be risky especially when using the Global Mirror technology where the application performances at the primary system can indeed be limited by the performance of the remote system.

The preferred practice is to perform an accurate back-end resource sizing for the remote system to fulfill the following capabilities:

- ▶ The peak application workload to the Global Mirror or Metro Mirror volumes
- ▶ The defined level of background copy
- ▶ Any other I/O that is performed at the remote site

Remote Copy tunable parameters

Several commands and parameters help to control remote copy and its default settings. You can display the properties and features of the systems by using the **lssystem** command. Also, you can change the features of systems by using the **chssystem** command.

relationshipbandwidthlimit

The **relationshipbandwidthlimit** is an optional parameter that specifies the new background copy bandwidth in the range 1 - 1000 MBps. The default is 25 MBps. This parameter operates system-wide, and defines the maximum background copy bandwidth that any relationship can adopt. The existing background copy bandwidth settings that are defined on a partnership continue to operate, with the lower of the partnership and volume rates attempted.

Important: Do not set this value higher than the default without establishing that the higher bandwidth can be sustained.

The **relationshipbandwidthlimit** also applies to Metro Mirror relationships.

gmlinktolerance and gmmaxhostdelay

The **gmlinktolerance** and **gmmaxhostdelay** parameters are critical in the system for deciding internally whether to terminate a relationship due to a performance problem. In most cases, these two parameters need to be considered in tandem. The defaults would not normally be changed unless you had a specific reason to do so.

The **gmlinktolerance** parameter can be thought of as how long you allow the host delay to go on being significant before you decide to terminate a Global Mirror volume relationship. This parameter accepts values of 20 - 86,400 seconds in increments of 10 seconds. The default is 300 seconds. You can disable the link tolerance by entering a value of zero for this parameter.

The **gmmaxhostdelay** parameter can be thought of as the maximum host I/O impact that is due to Global Mirror. That is, how long would that local I/O take with Global Mirror turned off, and how long does it take with Global Mirror turned on. The difference is the host delay due to Global Mirror tag and forward processing.

Although the default settings are adequate for most situations, increasing one parameter while reducing another might deliver a tuned performance environment for a particular circumstance.

Example 5-1 shows how to change **gmlinktolerance** and the **gmmaxhostdelay** parameters using the **chsystem** command.

Example 5-1 Changing gmlinktolerance to 30 and gmmaxhostdelay to 100

```
chsystem -gmlinktolerance 30
chsystem -gmmaxhostdelay 100
```

Test and monitor: To reiterate, thoroughly test and carefully monitor the host impact of any changes like these before putting them into a live production environment.

A detailed description and settings considerations about the **gmlinktolerance** and the **gmmaxhostdelay** parameters are described in 5.3.5, “1920 error” on page 202.

rcbuffersize

rcbuffersize was introduced with the version 6.2 code level so that systems with intense and bursty write I/O would not fill the internal buffer while Global Mirror writes were undergoing sequence tagging.

Important: Do not change the **rcbuffersize** parameter except under the direction of IBM Support.

Example 5-2 shows how to change **rcbuffersize** to 64 MB by using the **chsystem** command. The default value for **rcbuffersize** is 48 MB and the maximum is 512 MB.

Example 5-2 Changing rcbuffersize to 64 MB

```
chsystem -rcbuffersize 64
```

Remember that any additional buffers you allocate are taken away from the general cache.

maxreplicationdelay and partnershipexclusionthreshold

IBM Spectrum Virtualize version 7.6 introduced two new parameters, **maxreplicationdelay** and **partnershipexclusionthreshold**, for remote copy advanced tuning.

maxreplicationdelay is a system-wide parameter that defines a maximum latency (in seconds) for any individual write passing through the Global Mirror logic. If a write is hung for that time, for example due to a rebuilding array on the secondary system, Global Mirror stops the relationship (and any containing consistency group), triggering a 1920 error.

The **partnershipexclusionthreshold** parameter was introduced to allow users to set the timeout for an I/O that triggers a temporarily dropping of the link to the remote cluster. The value must be a number from 30 - 315.

Important: Do not change the **partnershipexclusionthreshold** parameter except under the direction of IBM Support.

A detailed description and settings considerations about the **maxreplicationdelay** parameter are described in 5.3.5, “1920 error” on page 202.

Link delay simulation parameters

Even though Global Mirror is an asynchronous replication method, there can be an impact to server applications due to Global Mirror managing transactions and maintaining write order consistency over a network. To mitigate this impact, as a testing and planning feature, Global Mirror allows you to simulate the effect of the round-trip delay between sites by using the following parameters:

- ▶ The **gminterclusterdelaysimulation** parameter
This optional parameter specifies the intersystem delay simulation, which simulates the Global Mirror round-trip delay between two systems in milliseconds. The default is 0. The valid range is 0 - 100 milliseconds.
- ▶ The **gmintraclusterdelaysimulation** parameter
This optional parameter specifies the intrasystem delay simulation, which simulates the Global Mirror round-trip delay in milliseconds. The default is 0. The valid range is 0 - 100 milliseconds.

5.3.4 Remote Copy use cases

The most common use cases for the remote copy functions are obviously Disaster Recovery solutions. A complete discussion about the Disaster Recovery solutions based on IBM Spectrum Virtualize technology is beyond the intended scope for this book. For an overview of the Disaster Recovery solutions with the IBM Spectrum Virtualize copy services see *IBM System Storage SAN Volume Controller and Storwize V7000 Replication Family Services*, SG24-7574.

Another typical remote copy use case is the data movement among distant locations as required, for instance, for data center relocation and consolidation projects. In these scenarios, the IBM Spectrum Virtualize remote copy technology is particularly effective when combined with the image copy feature that allows data movement among storage systems of different technology or vendor.

Remote copy services can also be combined with Volume Mirroring to implement three site solutions, as shown in Figure 5-23.

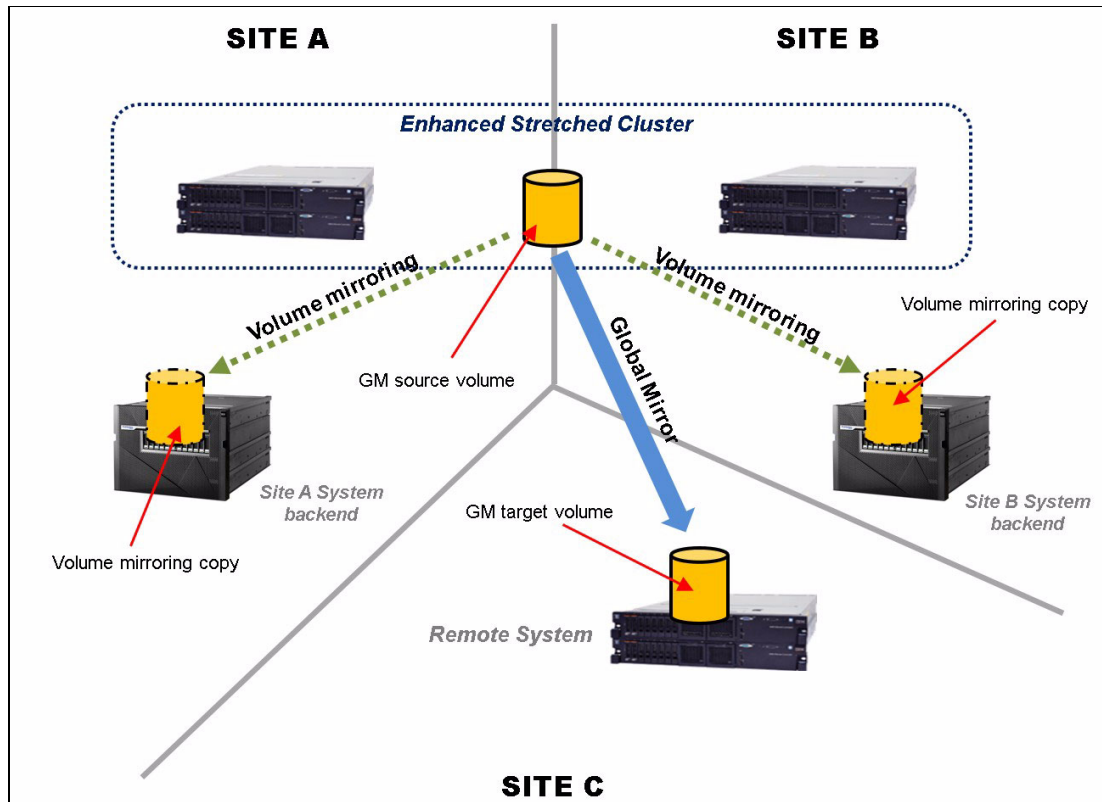


Figure 5-23 Three site configuration with Enhanced Stretched Cluster

Three site configurations can also be implemented using special cascading configurations, as described in the following sections.

Performing cascading copy service functions

Cascading copy service functions that use IBM Spectrum Virtualize/Storwize are not directly supported. However, you might require a three-way (or more) replication by using copy service functions (synchronous or asynchronous mirroring). You can address this requirement both by using IBM Spectrum Virtualize/Storwize copy services and by combining IBM Spectrum Virtualize/Storwize copy services (with image mode cache-disabled volumes) and native storage controller copy services.

Cascading with native storage controller copy services

Figure 5-24 describes the configuration for three-site cascading by using the native storage controller copy services in combination with IBM Spectrum Virtualize/Storwize remote copy functions.

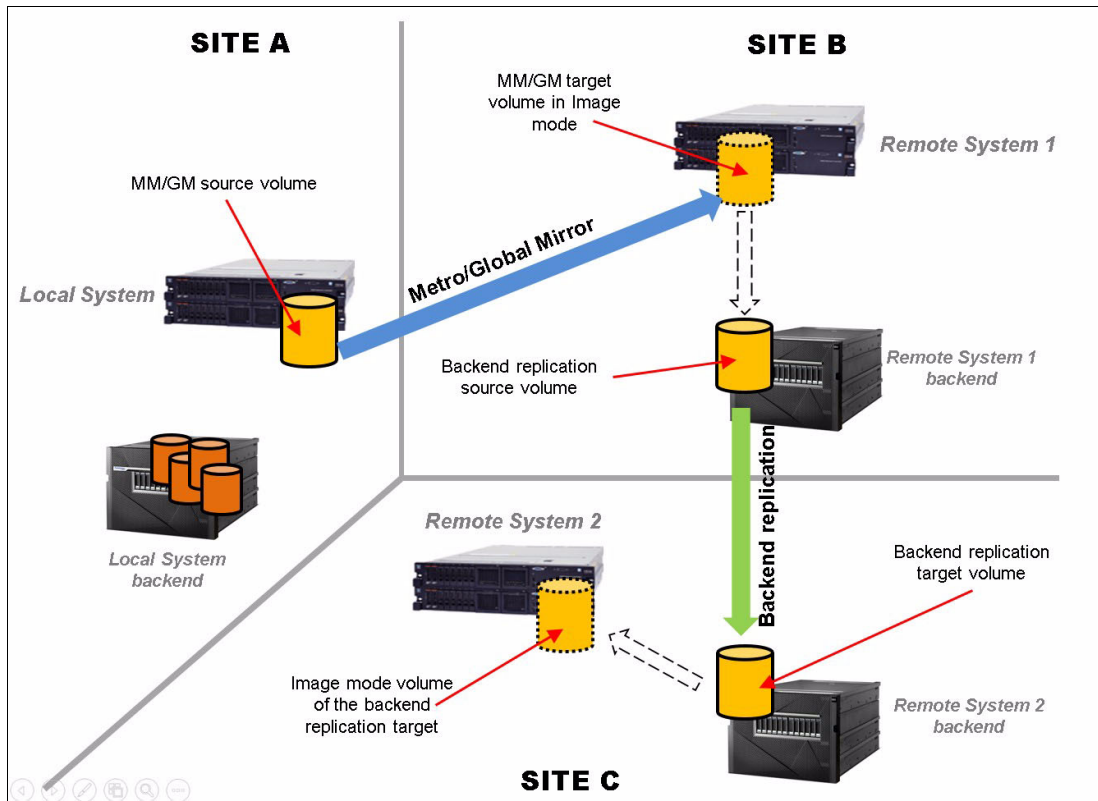


Figure 5-24 Using three-way copy services

In Figure 5-24, the primary site uses IBM Spectrum Virtualize/Storwize remote copy functions (Global Mirror or Metro Mirror) at the secondary site. Therefore, if a disaster occurs at the primary site, the storage administrator enables access to the target volume (from the secondary site) and the business application continues processing.

While the business continues processing at the secondary site, the storage controller copy services replicate to the third site. This configuration is allowed under the following conditions:

- ▶ The backend controller native copy services must be supported by Spectrum Virtualize (see “Native backend controller copy functions considerations” on page 190).
- ▶ The source and target volumes used by the backend controller native copy services must be imported to the IBM Spectrum Virtualize system as image-mode volumes with the cache disabled.

Cascading with IBM Spectrum Virtualize and Storwize systems copy services

A cascading-like solution is also possible by combining the IBM Spectrum Virtualize/Storwize copy services. These remote copy services implementations are useful in three site disaster recovery solutions and data center moving scenarios.

In the configuration described in Figure 5-25, a Global Mirror (Metro Mirror can also be used) solution is implemented between the Local System in Site A, the production site, and the Remote System 1 located in Site B, the primary disaster recover site. A third system, Remote System 2, is located in Site C, the secondary disaster recover site. Connectivity is provided between Site A and Site B, between Site B and Site C, and optionally between Site A and Site C.

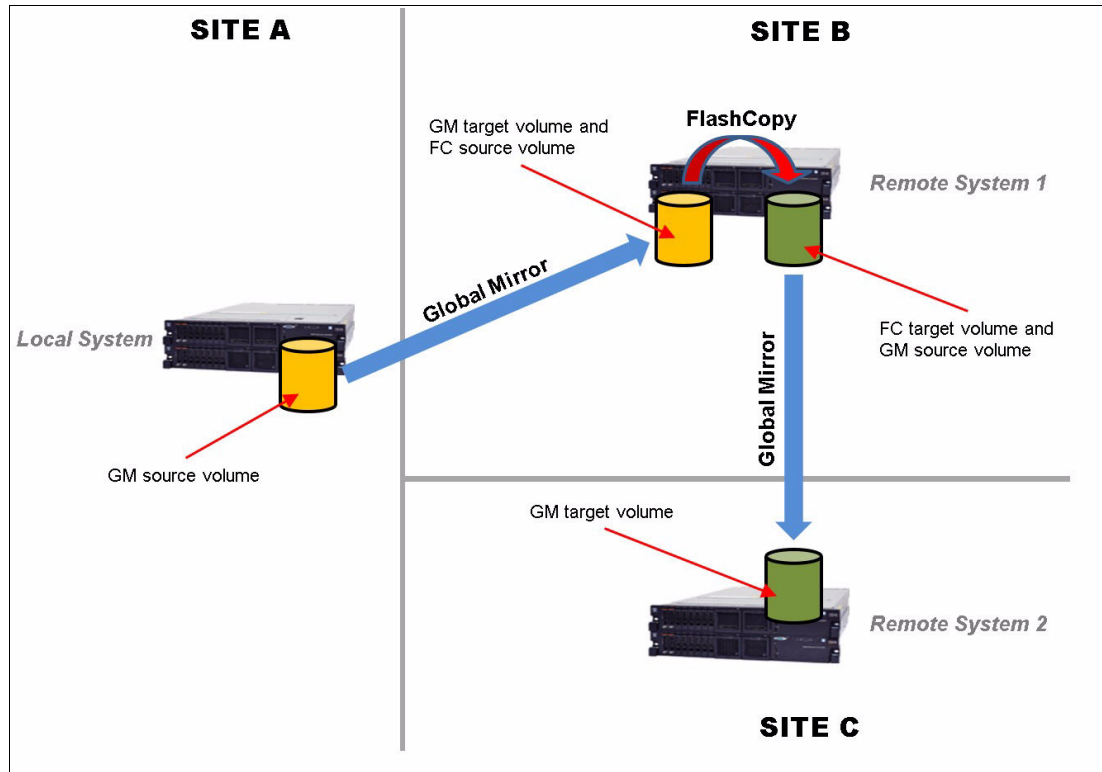


Figure 5-25 Cascading-like infrastructure

To implement a cascading-like solution, the following steps must be completed:

1. Set up phase. Perform the following actions to initially set up the environment:
 - a. Create the Global Mirror relationships between the Local System and Remote System 1.
 - b. Create the FlashCopy mappings in the Remote System 1 using the target Global Mirror volumes as FlashCopy source volumes. The FlashCopy must be incremental.
 - c. Create the Global Mirror relationships between Remote System 1 and Remote System 2 using the FlashCopy target volumes as Global Mirror source volumes.
 - d. Start the Global Mirror from Local System to Remote System 1.

After the Global Mirror is in ConsistentSynchronized state, you are ready to create the cascading.
2. Consistency point creation phase. The following actions must be performed every time a consistency point creation in the Site C is required.
 - a. Check whether the Global Mirror between Remote System 1 and Remote System 2 is in stopped or idle status, if it is not, stop the Global Mirror.
 - b. Stop the Global Mirror between the Local System to Remote System 1.
 - c. Start the FlashCopy in Remote Site 1.

- d. Resume the Global Mirror between the Local System and Remote System 1.
- e. Start/resume the Global Mirror between Remote System 1 and Remote System 2.

The first time that these operations are performed, a full copy between Remote System 1 and Remote System 2 occurs. Later executions of these operations perform incremental resynchronizations instead. After the Global Mirror between Remote System 1 and Remote System 2 is in ConsistentSynchronized state, the consistency point in Site C is created. The Global Mirror between Remote System 1 and Remote System 2 can now be stopped to be ready for the next consistency point creation.

5.3.5 1920 error

An IBM Spectrum Virtualize/Storwize system generates a 1920 error message whenever a Metro Mirror or Global Mirror relationship stops because of adverse conditions. The adverse conditions, if left unresolved, might affect performance of foreground I/O.

A 1920 error can result for many reasons. The condition might be the result of a temporary failure, such as maintenance on the intersystem connectivity, unexpectedly higher foreground host I/O workload, or a permanent error because of a hardware failure. It is also possible that not all relationships are affected and that multiple 1920 errors can be posted.

The 1920 error could be triggered both for Metro Mirror and Global Mirror relationships. However, in Metro Mirror configurations the 1920 error is associated only with a permanent I/O error condition. For this reason, the main focus of this section is 1920 errors in a Global Mirror configuration.

Internal Global Mirror control policy and raising 1920 errors

Although Global Mirror is an asynchronous remote copy service, the local and remote sites have some interplay. When data comes into a local volume, work must be done to ensure that the remote copies are consistent. This work can add a delay to the local write. Normally, this delay is low. The IBM Spectrum Virtualize code implements many control mechanisms that mitigate the impacts of the Global Mirror to the foreground I/Os.

gmmxhostdelay and gmlinktolerance

The ***gmlinktolerance*** parameter helps to ensure that hosts do not perceive the latency of the long-distance link, regardless of the bandwidth of the hardware that maintains the link or the storage at the secondary site. The hardware and storage must be provisioned so that, when combined, they can support the maximum throughput that is delivered by the applications at the primary that is using Global Mirror.

If the capabilities of this hardware are exceeded, the system becomes backlogged and the hosts receive higher latencies on their write I/O. Remote copy in Global Mirror implements a protection mechanism to detect this condition and halts mirrored foreground write and background copy I/O. Suspension of this type of I/O traffic ensures that misconfiguration or hardware problems (or both) do not affect host application availability.

Global Mirror attempts to detect and differentiate between backlogs that occur because of the operation of the Global Mirror protocol. It does not examine the general delays in the system when it is heavily loaded, where a host might see high latency even if Global Mirror were disabled.

To detect these specific scenarios, Global Mirror measures the time that is taken to perform the messaging to assign and record the sequence number for a write I/O. If this process exceeds the expected value over a period of 10 seconds, this period is treated as being overloaded (*bad period*).

Global Mirror uses the **gmmaxhostdelay** and **gm1inktolerance** parameters to monitor Global Mirror protocol backlogs in the following ways:

- ▶ Users set the **gmmaxhostdelay** and **gm1inktolerance** parameters to control how software responds to these delays. The **gmmaxhostdelay** parameter is a value in milliseconds that can go up to 100.
- ▶ Every 10 seconds, Global Mirror samples all of the Global Mirror writes and determines how much of a delay it added. If at least a third of these writes are greater than the **gmmaxhostdelay** setting, that sample period is marked as *bad*.
- ▶ Software keeps a running count of *bad periods*. Each time that a bad period occurs, this count goes up by one. Each time a good period occurs, this count goes down by 1, to a minimum value of 0.

The **gm1inktolerance** parameter is defined in seconds. Bad periods are assessed at intervals of 10 seconds. The maximum bad period count is the **gm1inktolerance** parameter value that is divided by 10. For instance, with a **gm1inktolerance** value of 300, the maximum bad period count is 30. When maximum bad period count is reached, a 1920 error is reported.

Bad periods do not need to be consecutive, and the bad period count increments or decrements at intervals of 10. That is, 10 bad periods, followed by five good periods, followed by 10 bad periods, results in a bad period count of 15.

Within each sample period, Global Mirror writes are assessed. If in a write operation, the delay added by the Global Mirror protocol exceeds the **gmmaxhostdelay** value, the operation is counted as a bad write. Otherwise, a good write is counted. The proportion of bad writes to good writes is calculated. If at least one third of writes are identified as bad, the sample period is defined as a bad period. A consequence is that, under a light I/O load, a single bad write can become significant. For example, if only one write I/O is performed for every 10 and this write is considered slow, the bad period count increments.

An edge case is achieved by setting the **gmmaxhostdelay** and **gm1inktolerance** parameters to their minimum settings (1 ms and 20 s). With these settings, you need only two consecutive bad sample periods before a 1920 error condition is reported. Consider a foreground write I/O that has a light I/O load. For example, a single I/O happens in the 20 s. With unlucky timing, a single bad I/O results (that is, a write I/O that took over 1 ms in remote copy), and it spans the boundary of two, 10-second sample periods. This single bad I/O theoretically can be counted as 2 x the bad periods and trigger a 1920 error.

A higher **gm1inktolerance** value, **gmmaxhostdelay** setting, or I/O load might reduce the risk of encountering this edge case.

maxreplicationdelay and partnershipexclusionthreshold

IBM Spectrum Virtualize version 7.6 has introduced the **maxreplicationdelay** and **partnershipexclusionthreshold** parameters to provide further performance protection mechanisms when remote copy services (Metro Mirror and Global Mirror) are used.

maxreplicationdelay is a system-wide attribute that configures how long a single write can be outstanding from the host before the relationship is stopped, triggering a 1920 error. It can protect the hosts from seeing timeouts due to secondary hung I/Os.

This parameter is mainly intended to protect from secondary system issues. It does not help with ongoing performance issues, but can be used to limit the exposure of hosts to long write response times that can cause application errors. For instance, setting **maxreplicationdelay** to 30 means that if a write operation for a volume in a remote copy relationship does not complete within 30 seconds, the relationship is stopped, triggering a 1920 error.

In addition to the 1920 error, the specific event ID 985004 is generated with the text “Maximum replication delay exceeded”.

The **maxreplicationdelay** values can be 0 - 360 seconds. Setting **maxreplicationdelay** to 0 disables the feature.

The **partnershipexclusionthreshold** is a system-wide parameter that sets the timeout for an I/O that triggers a temporarily dropping of the link to the remote system. Similar to **maxreplicationdelay**, the **partnershipexclusionthreshold** attribute provides some flexibility in a part of replication that tries to shield a production system from hung I/Os on a secondary system.

In an IBM Spectrum Virtualize/Storwize system, a node assert (restart with a 2030 error) occurs if any individual I/O takes longer than 6 minutes. To avoid this situation, some actions are attempted to clean up anything that might be hanging I/O before the I/O gets to 6 minutes.

One of these actions is temporarily dropping (for 15 minutes) the link between systems if any I/O takes longer than 5 minutes 15 seconds (315 seconds). This action often removes hang conditions caused by replication problems. The **partnershipexclusionthreshold** parameter introduced the ability to set this value to a time lower than 315 seconds to respond to hung I/O more swiftly. The **partnershipexclusionthreshold** value must be a number in the range 30 - 315.

If an I/O takes longer the **partnershipexclusionthreshold** value, a 1720 error is triggered (with an event ID 987301) and any regular Global Mirror or Metro Mirror relationships stop on the next write to the primary volume.

Important: Do not change the **partnershipexclusionthreshold** parameter except under the direction of IBM Support.

To set the **maxreplicationdelay** and **partnershipexclusionthreshold** parameters, the **chsystem** command must be used, as shown in Example 5-3.

Example 5-3 maxreplicationdelay and partnershipexclusionthreshold setting

```
IBM_2145:SVC_ESC:superuser>chsystem -maxreplicationdelay 30
IBM_2145:SVC_ESC:superuser>chsystem -partnershipexclusionthreshold 180
```

The **maxreplicationdelay** and **partnershipexclusionthreshold** parameters do not interact with the **gmlinktolerance** and **gmmaxhostdelay** parameters.

Troubleshooting 1920 errors

When you are troubleshooting 1920 errors that are posted across multiple relationships, you must diagnose the cause of the earliest error first. You must also consider whether other higher priority system errors exist and fix these errors because they might be the underlying cause of the 1920 error.

The diagnosis of a 1920 error is assisted by SAN performance statistics. To gather this information, you can use IBM Spectrum Control with a statistics monitoring interval of 1 or 5 minutes. Also, turn on the internal statistics gathering function, **I0stats**, in IBM Spectrum Virtualize. Although not as powerful as IBM Spectrum Control, **I0stats** can provide valuable debug information if the **snap** command gathers system configuration data close to the time of failure.

The following are the main performance statistics to investigate for the 1920 error:

► *Write I/O Rate and Write Data Rate*

For volumes that are primary volumes in relationships, these statistics are the total amount of write operations submitted per second by hosts on average over the sample period, and the bandwidth of those writes. For secondary volumes in relationships, this is the average number of replicated writes that are received per second, and the bandwidth that these writes consume. Summing the rate over the volumes you intend to replicate gives a coarse estimate of the replication link bandwidth required.

► *Write Response Time and Peak Write Response Time*

On primary volumes, these are the average time (in milliseconds) and peak time between a write request being received from a host, and the completion message being returned. The write response time is the best way to show what kind of write performance that the host is seeing.

If a user complains that an application is slow, and the stats show the write response time leap from 1 ms to 20 ms, the two are most likely linked. However, some applications with high queue depths and low to moderate workloads will not be affected by increased response times. Note that this being high is an effect of some other problem. The peak is less useful, as it is very sensitive to individual glitches in performance, but it can show more detail of the distribution of write response times.

On secondary volumes, these statistics describe the time for the write to be submitted from the replication feature into the system cache, and should normally be of a similar magnitude to those on the primary volume. Generally, the write response time should be below 1 ms for a fast-performing system.

► *Global Mirror Write I/O Rate*

This statistic shows the number of writes per second, the (regular) replication feature is processing for this volume. It applies to both types of Global Mirror and to Metro Mirror, but in each case only for the secondary volume. Because writes are always separated into 32 KB or smaller tracks before replication, this setting might be different from the Write I/O Rate on the primary volume (magnified further because the samples on the two systems will not be aligned, so they will capture a different set of writes).

► *Global Mirror Overlapping Write I/O Rate*

This statistic monitors the amount of overlapping I/O that the Global Mirror feature is handling for regular Global Mirror relationships. That is where an LBA is written again after the primary volume has been updated, but before the secondary volume has been updated for an earlier write to that LBA. To mitigate the effects of the overlapping I/Os, a journaling feature has been implemented, as discussed in “Colliding writes” on page 173.

► *Global Mirror secondary write lag*

This statistic is valid for regular Global Mirror primary and secondary volumes. For primary volumes, it tracks the length of time in milliseconds that replication writes are outstanding from the primary system. This amount includes the time to send the data to the remote system, consistently apply it to the secondary non-volatile cache, and send an acknowledgment back to the primary system.

For secondary volumes, this statistic records only the time that is taken to consistently apply it to the system cache, which is normally up to 20 ms. Most of that time is spent coordinating consistency across many nodes and volumes. Primary and secondary volumes for a relationship tend to record times that differ by the round-trip time between systems. If this statistic is high on the secondary system, look for congestion on the secondary system’s fabrics, saturated auxiliary storage, or high CPU utilization on the secondary system.

► *Write-cache Delay I/O Rate*

These statistics show how many writes could not be instantly accepted into the system cache because cache was full. It is a good indication that the write rate is faster than the storage can cope with. If this amount starts to increase on auxiliary storage while primary volumes suffer from increased Write Response Time, it is possible that the auxiliary storage is not fast enough for the replicated workload.

► *Port to Local Node Send Response Time*

The time in milliseconds that it takes this node to send a message to other nodes in the same system (which will mainly be the other node in the same I/O group) and get an acknowledgment back. This amount should be well below 1 ms, with values below 0.3 ms being essential for regular Global Mirror to provide a Write Response Time below 1 ms.

This requirement is necessary because up to three round-trip messages within the local system will happen before a write completes to the host. If this number is higher than you want, look at fabric congestion (Zero Buffer Credit Percentage) and CPU Utilization of all nodes in the system.

► *Port to Remote Node Send Response Time*

This value is the time in milliseconds that it takes to send a message to nodes in other systems and get an acknowledgment back. This amount is not separated out by remote system, but for environments that have replication to only one remote system. This amount should be very close to the low-level ping time between your sites. If this starts going significantly higher, it is likely that the link between your systems is saturated, which usually causes high Zero Buffer Credit Percentage as well.

► *Sum of Port-to-local node send response time and Port-to-local node send queue time*

The time must be less than 1 ms for the primary system. A number in excess of 1 ms might indicate that an I/O group is reaching its I/O throughput limit, which can limit performance.

► *System CPU Utilization (Core 1-8)*

These values show how heavily loaded the nodes in the system are. If any core has high utilization (say, over 90%) and there is an increase in write response time, it is possible that the workload is being CPU limited. You can resolve this by upgrading to faster hardware, or spreading out some of the workload to other nodes and systems.

► *Zero Buffer Credit Percentage*

This is the fraction of messages that this node attempted to send through Fibre Channel ports that had to be delayed because the port ran out of buffer credits. If you have a long link from the node to the switch it is attached to, there might be benefit in getting the switch to grant more buffer credits on its port.

It is more likely to be the result of congestion on the fabric, because running out of buffer credits is how Fibre Channel performs flow control. Normally, this value is well under 1%. From 1 - 10% is a concerning level of congestion, but you might find the performance acceptable. Over 10% indicates severe congestion. This amount is also called out on a port-by-port basis in the port-level statistics, which gives finer granularity about where any congestion might be.

When looking at the port-level statistics, high values on ports used for messages to nodes in the same system are much more concerning than those on ports that are used for messages to nodes in other systems.

► *Back-end Write Response Time*

This value is the average response time in milliseconds for write operations to the back-end storage. This time might include several physical I/O operations, depending on the type of RAID architecture.

Poor back-end performances on secondary system is a frequent cause of 1920 errors, while it is not so common for primary systems. Exact values to watch out for depend on the storage technology, but usually the response time should be less than 50 ms. A longer response time can indicate that the storage controller is overloaded. If the response time for a specific storage controller is outside of its specified operating range, investigate for the same reason.

Focus areas for 1920 errors

The causes of 1920 errors might be numerous. To fully understand the underlying reasons for posting this error, consider the following components that are related to the remote copy relationship:

- The intersystem connectivity network
- Primary storage and remote storage
- IBM Spectrum Virtualize nodes and Storwize node canisters
- Storage area network

Data collection for diagnostic purposes

A successful diagnosis depends on the collection of the following data at both systems:

- The **snap** command with **livedump** (triggered at the point of failure)
- I/O Stats running (if possible)
- IBM Spectrum Control performance statistics data (if possible)
- The following information and logs from other components:
 - Intersystem network and switch details:
 - Technology
 - Bandwidth
 - Typical measured latency on the Intersystem network
 - Distance on all links (which can take multiple paths for redundancy)
 - Whether trunking is enabled
 - How the link interfaces with the two SANs
 - Whether compression is enabled on the link
 - Whether the link dedicated or shared; if so, the resource and amount of those resources they use
 - Switch Write Acceleration to check with IBM for compatibility or known limitations
 - Switch Compression, which should be transparent but complicates the ability to predict bandwidth
 - Storage and application:
 - Specific workloads at the time of 1920 errors, which might not be relevant, depending upon the occurrence of the 1920 errors and the volumes that are involved
 - RAID rebuilds
 - Whether 1920 errors are associated with Workload Peaks or Scheduled Backup

Intersystem network

For diagnostic purposes, ask the following questions about the intersystem network:

- ▶ Was network maintenance being performed?

Consider the hardware or software maintenance that is associated with intersystem network, such as updating firmware or adding more capacity.

- ▶ Is the intersystem network overloaded?

You can find indications of this situation by using statistical analysis with the help of I/O stats, IBM Spectrum Control, or both. Examine the internode communications, storage controller performance, or both. By using IBM Spectrum Control, you can check the storage metrics for the Global Mirror relationships were stopped, which can be tens of minutes depending on the **gmLinktolerance** and **maxreplicationdelay** parameters.

Diagnose the overloaded link by using the following methods:

- Look at the statistics generated by the routers or switches near your most bandwidth-constrained link between the systems

Exactly what is provided, and how to analyze it varies depending on the equipment used.

- Look at the port statistics for high response time in the internode communication

An overloaded long-distance link causes high response times in the internode messages (the *Port to remote node send response time* statistic) that are sent by IBM Spectrum Virtualize. If delays persist, the messaging protocols exhaust their tolerance elasticity and the Global Mirror protocol is forced to delay handling new foreground writes while waiting for resources to free up.

- Look at the port statistics for buffer credit starvation

The *Zero Buffer Credit Percentage* statistic can be useful here too, because you normally have a high value here as the link saturates. Only look at ports that are replicating to the remote system.

- Look at the volume statistics (before the 1920 error is posted):

- Target volume write throughput approaches the link bandwidth.

If the write throughput on the target volume is equal to your link bandwidth, your link is likely overloaded. Check what is driving this situation. For example, does peak foreground write activity exceed the bandwidth, or does a combination of this peak I/O and the background copy exceed the link capacity?

- Source volume write throughput approaches the link bandwidth.

This write throughput represents only the I/O that is performed by the application hosts. If this number approaches the link bandwidth, you might need to upgrade the link's bandwidth. Alternatively, reduce the foreground write I/O that the application is attempting to perform, or reduce the number of remote copy relationships.

- Target volume write throughput is greater than the source volume write throughput.

If this condition exists, the situation suggests a high level of background copy and mirrored foreground write I/O. In these circumstances, decrease the background copy rate parameter of the Global Mirror partnership to bring the combined mirrored foreground I/O and background copy I/O rates back within the remote links bandwidth.

- Look at the volume statistics (after the 1920 error is posted):
 - Source volume write throughput after the Global Mirror relationships were stopped.

If write throughput increases greatly (by 30% or more) after the Global Mirror relationships are stopped, the application host was attempting to perform more I/O than the remote link can sustain.

When the Global Mirror relationships are active, the overloaded remote link causes higher response times to the application host. This overload, in turn, decreases the throughput of application host I/O at the source volume. After the Global Mirror relationships stop, the application host I/O sees a lower response time, and the true write throughput returns.

To resolve this issue, increase the remote link bandwidth, reduce the application host I/O, or reduce the number of Global Mirror relationships.

Storage controllers

Investigate the primary and remote storage controllers, starting at the remote site. If the back-end storage at the secondary system is overloaded, or another problem is affecting the cache there, the Global Mirror protocol fails to keep up. Similarly, the problem exhausts the (**gmlinktolerance**) elasticity and has a similar effect at the primary system.

In this situation, ask the following questions:

- ▶ Are the storage controllers at the remote system overloaded (performing slowly)?

Use IBM Spectrum Control to obtain the back-end write response time for each MDisk at the remote system. A response time for any individual MDisk that exhibits a sudden increase of 50 ms or more, or that is higher than 100 ms, generally indicates a problem with the back end.

However, if you followed the specified back-end storage controller requirements and were running without problems until recently, the error is most likely caused by a decrease in controller performance because of maintenance actions or a hardware failure of the controller. Check whether an error condition is on the storage controller, for example, media errors, a failed physical disk, or a recovery activity, such as RAID array rebuilding that uses more bandwidth.

If an error occurs, fix the problem and then restart the Global Mirror relationships.

If no error occurs, consider whether the secondary controller can process the required level of application host I/O. You might improve the performance of the controller in the following ways:

- Adding more or faster physical disks to a RAID array.
- Changing the RAID level of the array.
- Changing the cache settings of the controller and checking that the cache batteries are healthy, if applicable.
- Changing other controller-specific configuration parameter.

- ▶ Are the storage controllers at the primary site overloaded?

Analyze the performance of the primary back-end storage by using the same steps that you use for the remote back-end storage. The main effect of bad performance is to limit the amount of I/O that can be performed by application hosts. Therefore, you must monitor back-end storage at the primary site regardless of Global Mirror.

However, if bad performance continues for a prolonged period, a false 1920 error might be flagged.

Node and canister

For IBM Spectrum Virtualize node and Storwize node canister hardware, the possible cause of the 1920 errors might be from a heavily loaded secondary or primary system. If this condition persists, a 1920 error might be posted.

Global Mirror needs to synchronize its I/O processing across all nodes in the system to ensure data consistency. If any node is running out of CPU, it can affect all relationships. So check the CPU cores usage statistic. If it looks higher when there is a performance problem, then running out of CPU bandwidth might be causing the problem. Of course, CPU usage goes up when the IOPS going through a node goes up, so if the workload increases, you would expect to see CPU usage increase.

If there is an increase in CPU usage on the secondary system but no increase in IOPS, and volume write latency increases too, it is likely that the increase in CPU usage has caused the increased volume write latency. In that case, try to work out what might have caused the increase in CPU usage (for example, starting many FlashCopy mappings). Consider moving that activity to a time with less workload. If there is an increase in both CPU usage and IOPS, and the CPU usage is close to 100%, then that node might be overloaded. A *Port-to-local node send queue time* value higher than 0.2 ms often denotes CPU cores overloading.

In a primary system, if it is sufficiently busy, the write ordering detection in Global Mirror can delay writes enough to reach a latency of `gmmxhostdelay` and cause a 1920 error. Stopping replication potentially lowers CPU usage, and also lowers the opportunities for each I/O to be delayed by slow scheduling on a busy system.

Solve overloaded nodes by upgrading them to newer, faster hardware if possible, or by adding more I/O groups/control enclosures (or systems) to spread the workload over more resources.

Storage area network

Issues and congestions both in local and remote SANs can lead to 1920 errors. The *Port to local node send response time* is the key statistic to investigate on. It captures the round-trip time between nodes in the same system. Anything over 1.0 ms is surprisingly high, and will cause high secondary volume write response time. Values greater than 1 ms on primary system will cause an impact on write latency to Global Mirror primary volumes of 3 ms or more.

If you have checked CPU utilization on all the nodes, and it has not gotten near 100%, a high *Port to local node send response time* means that there is fabric congestion or a slow-draining Fibre Channel device.

A good indicator of SAN congestion is the *Zero Buffer Credit Percentage* on the port statistics (see “Buffer credits” on page 184 for more information on Buffer Credit). If any port is seeing over 10% zero buffer credits, that is definitely going to cause a problem for all I/O, not just Global Mirror writes. Values from 1 - 10% are moderately high and might contribute to performance issues.

For both primary and secondary systems, congestion on the fabric from other slow-draining devices becomes much less of an issue when only dedicated ports are used for node-to-node traffic within the system. However, this only really becomes an option on systems with more than four ports per node. Use port masking to segment your ports.

FlashCopy considerations

Check that FlashCopy mappings are in the *prepared* state. Check whether the Global Mirror target volumes are the sources of a FlashCopy mapping and whether that mapping was in the *prepared* state for an extended time.

Volumes in the prepared state are cache disabled, so their performance is impacted. To resolve this problem, start the FlashCopy mapping, which reenables the cache and improves the performance of the volume and of the Global Mirror relationship.

Consider also that FlashCopy can add significant workload to the back-end storage, especially when the background copy is active (see “Background copy considerations” on page 161). In cases where the remote system is used to create golden or practice copies for Disaster Recovery testing, the workload added by the FlashCopy background processes can overload the system. This overload can lead to poor remote copy performances and then to a 1920 error.

Careful planning of the back-end resources is particularly important with these kinds of scenarios. Reducing the FlashCopy background copy rate can also help to mitigate this situation. Furthermore, note that the FlashCopy copy-on-write process adds some latency by delaying the write operations on the primary volumes until the data is written to the FlashCopy target.

This process doesn't affect directly the remote copy operations since it is logically placed below the remote copy processing in the I/O stack, as shown in Figure 5-6 on page 151. Nevertheless, in some circumstances, especially with write intensive environments, the copy-on-write process tends to stress some systems's internal resources, like CPU and memory, and this condition can also affect the remote copy, that competes for the same resources, leading eventually to 1920 errors.

FCIP considerations

When you get a 1920 error, always check the latency first. The FCIP routing layer can introduce latency if it is not properly configured. If your network provider reports a much lower latency, you might have a problem at your FCIP routing layer. Most FCIP routing devices have built-in tools to enable you to check the RTT. When you are checking latency, remember that TCP/IP routing devices (including FCIP routers) report RTT by using standard 64-byte ping packets.

In Figure 5-26 on page 212, you can see why the effective transit time must be measured only by using packets that are large enough to hold an FC frame, or 2148 bytes (2112 bytes of payload and 36 bytes of header). Allow estimated resource requirements to be a safe amount because various switch vendors have optional features that might increase this size. After you verify your latency by using the proper packet size, proceed with normal hardware troubleshooting.

Look at the second largest component of your RTT, which is *serialization delay*. Serialization delay is the amount of time that is required to move a packet of data of a specific size across a network link of a certain bandwidth. The required time to move a specific amount of data decreases as the data transmission rate increases.

Figure 5-26 shows the orders of magnitude of difference between the link bandwidths. It is easy to see how 1920 errors can arise when your bandwidth is insufficient. Never use a TCP/IP ping to measure RTT for FCIP traffic.

Packet Size	Link Size	Serialization Delay (Time Required to Send Data)	Unit
64	256 Kbps	2.0E+03	microseconds
64	1.5 Mbps	3.4E+02	microseconds
64	100 Mbps	5.1E+00	microseconds
64	155 Mbps	3.3E+00	microseconds
64	622 Mbps	8.2E-01	microseconds
64	1 Gbps	5.1E-04	microseconds
64	10 Gbps	5.1E-05	microseconds
1500	256 Kbps	4.7E+04	microseconds
1500	1.5 Mbps	8.0E+03	microseconds
1500	100 Mbps	1.2E+02	microseconds
1500	155 Mbps	7.7E+01	microseconds
1500	622 Mbps	1.9E+01	microseconds
1500	1 Gbps	1.2E+01	microseconds
1500	10 Gbps	1.2E+00	microseconds
2148	256 Kbps	6.7E+04	microseconds
2148	1.5 Mbps	1.1E+04	microseconds
2148	100 Mbps	1.7E+02	microseconds
2148	155 Mbps	1.1E+02	microseconds
2148	622 Mbps	2.8E+01	microseconds
2148	1 Gbps	1.7E+01	microseconds
2148	10 Gbps	1.7E-03	microseconds

Figure 5-26 Effect of packet size (in bytes) versus the link size

In Figure 5-26, the amount of time in microseconds that is required to transmit a packet across network links of varying bandwidth capacity is compared. The following packet sizes are used:

- ▶ 64 bytes: The size of the common ping packet
- ▶ 1500 bytes: The size of the standard TCP/IP packet
- ▶ 2148 bytes: The size of an FC frame

Finally, your path maximum transmission unit (MTU) affects the delay that is incurred to get a packet from one location to another location. An MTU might cause fragmentation, or be too large and cause too many retransmits when a packet is lost.

Recovery after 1920 errors

After a 1920 error occurs, the Global Mirror auxiliary volumes are no longer in a Consistent Synchronized state. You must establish the cause of the problem and fix it before you restart the relationship.

When the relationship is restarted, you must resynchronize it. During this period, the data on the Metro Mirror or Global Mirror auxiliary volumes on the secondary system is inconsistent, and your applications cannot use the volumes as backup disks. To address this data consistency exposure on the secondary system, a FlashCopy of the auxiliary volumes can be created to maintain a consistent image until the Global Mirror (or the Metro Mirror) relationships are synchronized again and back in a consistent state.

IBM Spectrum Virtualize V7.8.1 introduced the Remote Copy *Consistency Protection* feature that automates this process. When Consistency Protection is configured, the relationship between the primary and secondary volumes does not go in to the Inconsistent copying status once restarted. Instead, the system uses a secondary *change volume* to automatically copy the previous consistent state of the secondary volume.

The relationship automatically moves to the `Consistent copying` status as the system resynchronizes and protects the consistency of the data. The relationship status changes to `Consistent synchronized` when the resynchronization process completes. For further details about the Consistency Protection feature, see *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.1*, SG24-7933.

To ensure that the system can handle the background copy load, delay restarting the Metro Mirror or Global Mirror relationship until a quiet period occurs. If the required link capacity is unavailable, you might experience another 1920 error, and the Metro Mirror or Global Mirror relationship may stop in an inconsistent state.

Copy services tools, like IBM Copy Services Manager (CSM), or manual scripts can be used to automatize the relationships to restart after a 1920 error. CSM implements a logic to avoid recurring restart operations in case of a persistent problem. CSM attempts an automatic restart for every occurrence of 1720/1920 error a certain number of times (determined by the `gminktolerance` value) within a 30 minute time period.

If the number of allowable automatic restarts is exceeded within the time period, CSM will not automatically restart GM on the next 1720/1920 error. Furthermore, with CSM it is possible to specify the amount of time, in seconds, in which the tool will wait after an 1720/1920 error before automatically restarting the GM. Further details about IBM Copy Services Manager can be found at the following website:

<https://ibm.biz/BdjVCe>

Tip: When implementing automatic restart functions, it is advised to preserve the data consistency on GM target volumes during the resynchronization using features like Flashcopy or Consistency Protection.

Adjusting the Global Mirror settings

Although the default values are valid in most configurations, the settings of the `gminktolerance` and `gmmaxhostdelay` can be adjusted to accommodate particular environment or workload conditions.

For example, Global Mirror is designed to look at average delays. However, some hosts such as VMware ESX might not tolerate a single I/O getting old, for example, 45 seconds, before it decides to reboot. Given that it is better to terminate a Global Mirror relationship than it is to reboot a host, you might want to set `gminktolerance` to something like 30 seconds and then compensate so that you do not get too many relationship terminations by setting `gmmaxhostdelay` to something larger, such as 100 ms.

If you compare the two approaches, the default (`gminktolerance 300`, `gmmaxhostdelay 5`) is a rule that “If more than one third of the I/Os are slow and that happens repeatedly for 5 minutes, then terminate the busiest relationship in that stream.” In contrast, the example of `gminktolerance 30`, `gmmaxhostdelay 100` is a rule that “If more than one third of the I/Os are extremely slow and that happens repeatedly for 30 seconds, then terminate the busiest relationship in the stream.”

So one approach is designed to pick up general slowness, and the other approach is designed to pick up shorter bursts of extreme slowness that might disrupt your server environment. The general recommendation is to change the `gminktolerance` and `gmmaxhostdelay` values progressively and evaluate the overall impact to find an acceptable compromise between performances and Global Mirror stability.

You can even disable the **gmLinktolerance** feature by setting the **gmLinktolerance** value to 0. However, the **gmLinktolerance** parameter cannot protect applications from extended response times if it is disabled. You might consider disabling the **gmLinktolerance** feature in the following circumstances:

- ▶ During SAN maintenance windows, where degraded performance is expected from SAN components and application hosts can withstand extended response times from Global Mirror volumes.
- ▶ During periods when application hosts can tolerate extended response times and it is expected that the **gmLinktolerance** feature might stop the Global Mirror relationships. For example, you are testing usage of an I/O generator that is configured to stress the back-end storage. Then, the **gmLinktolerance** feature might detect high latency and stop the Global Mirror relationships. Disabling the **gmLinktolerance** parameter stops the Global Mirror relationships at the risk of exposing the test host to extended response times.

Another tunable parameter that interacts with the GM is the **maxreplicationdelay**. Note that the **maxreplicationdelay** settings do not mitigate the 1920 error occurrence because it actually adds a trigger to the 1920 error itself. However, the **maxreplicationdelay** provides users with a fine granularity mechanism to manage the hung I/Os condition and it can be used in combination with **gmLinktolerance** and **gmmaxhostdelay** settings to better address particular environment conditions.

In the above VMware example, an alternative option is to set the **maxreplicationdelay** to 30 seconds and leave the **gmLinktolerance** and **gmmaxhostdelay** settings to their default. With these settings, the **maxreplicationdelay** timeout effectively handles the hung I/Os conditions, while the **gmLinktolerance** and **gmmaxhostdelay** settings still provide an adequate mechanism to protect from ongoing performance issues.

5.4 Native IP replication

The native IP replication feature enables replication between any IBM Spectrum Virtualize and Storwize family products running code V7.2 or higher. It does so by using the built-in networking ports or optional 1/10 Gb adapter.

Following a recent partnership with IBM, native IP replication uses SANslide technology developed by Bridgeworks Limited of Christchurch, UK. They specialize in products that can bridge storage protocols and accelerate data transfer over long distances. Adding this technology at each end of a wide area network (WAN) TCP/IP link significantly improves the utilization of the link.

It does this by applying patented artificial intelligence (AI) to hide latency that is normally associated with WANs. Doing so can greatly improve the performance of mirroring services, in particular Global Mirror with Change Volumes (GM/CV) over long distances.

5.4.1 Native IP replication technology

Remote Mirroring over IP communication is supported on the IBM Spectrum Virtualize and Storwize Family systems by using Ethernet communication links. The IBM Spectrum Virtualize Software IP replication uses innovative *Bridgeworks SANSlide* technology to optimize network bandwidth and utilization. This new function enables the use of a lower-speed and lower-cost networking infrastructure for data replication.

Bridgeworks' SANSlide technology, which is integrated into the IBM Spectrum Virtualize Software, uses artificial intelligence to help optimize network bandwidth use and adapt to changing workload and network conditions. This technology can improve remote mirroring network bandwidth usage up to three times. It can enable clients to deploy a less costly network infrastructure, or speed up remote replication cycles to enhance disaster recovery effectiveness.

With an Ethernet network data flow, the data transfer can slow down over time. This condition occurs because of the latency that is caused by waiting for the acknowledgment of each set of packets that are sent. The next packet set cannot be sent until the previous packet is acknowledged, as shown in Figure 5-27.

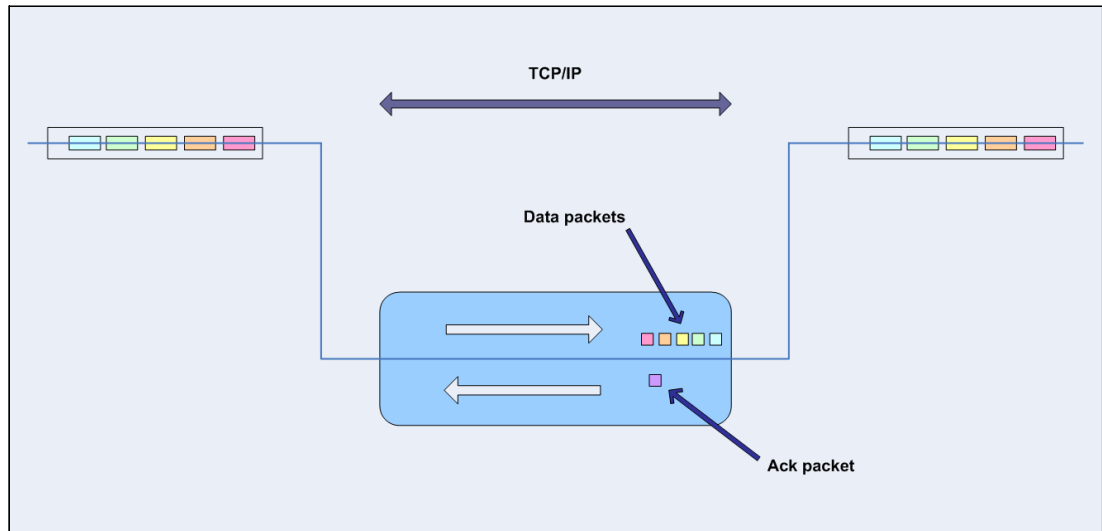


Figure 5-27 Typical Ethernet network data flow

However, by using the embedded IP replication, this behavior can be eliminated with the enhanced parallelism of the data flow. This parallelism uses multiple virtual connections (VCs) that share IP links and addresses.

The artificial intelligence engine can dynamically adjust the number of VCs, receive window size, and packet size as appropriate to maintain optimum performance. While the engine is waiting for one VC's ACK, it sends more packets across other VCs. If packets are lost from any VC, data is automatically retransmitted, as shown in Figure 5-28.

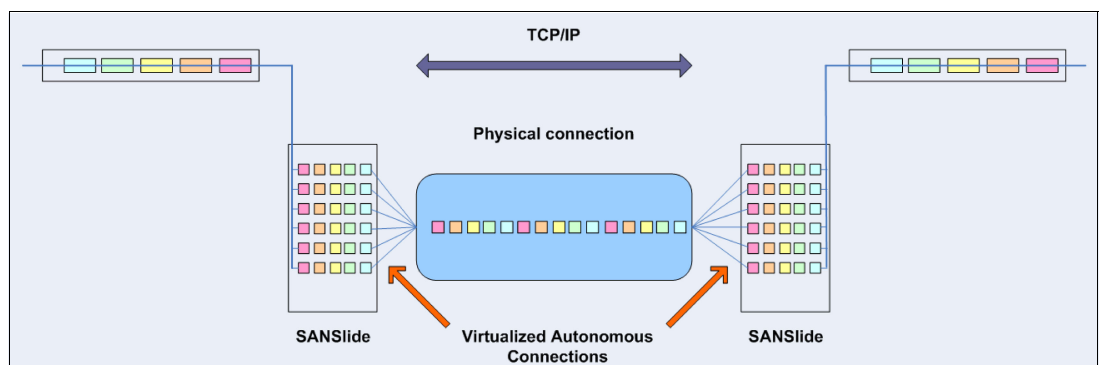


Figure 5-28 Optimized network data flow by using Bridgeworks SANSlide technology

For more information about this technology, see *IBM SAN Volume Controller and Storwize Family Native IP Replication*, REDP-5103.

Metro Mirror, Global Mirror, and Global Mirror Change Volume are supported with native IP partnership.

5.4.2 IP partnership limitations

The following prerequisites and assumptions must be considered before IP partnership between two IBM Spectrum Virtualize or Storwize family systems can be established:

- ▶ The systems are successfully installed with V7.2 or later code levels.
- ▶ The systems have the necessary licenses that enable remote copy partnerships to be configured between two systems. No separate license is required to enable IP partnership.
- ▶ The storage SANs are configured correctly and the correct infrastructure to support the systems in remote copy partnerships over IP links is in place.
- ▶ The two systems must be able to ping each other and perform the discovery.
- ▶ The maximum number of partnerships between the local and remote systems, including both IP and Fibre Channel (FC) partnerships, is limited to the current maximum that is supported, which is three partnerships (four systems total).
- ▶ Only a single partnership over IP is supported.
- ▶ A system can have simultaneous partnerships over FC and IP, but with separate systems. The FC zones between two systems must be removed before an IP partnership is configured.
- ▶ IP partnerships are supported on both 10 gigabits per second (Gbps) links and 1 Gbps links. However, the intermix of both on a single link is not supported.
- ▶ The maximum supported round-trip time is 80 milliseconds (ms) for 1 Gbps links.
- ▶ The maximum supported round-trip time is 10 ms for 10 Gbps links.
- ▶ The minimum supported link bandwidth is 10 Mbps.
- ▶ The inter-cluster heartbeat traffic uses 1 Mbps per link.
- ▶ Only nodes from two I/O Groups can have ports that are configured for an IP partnership.
- ▶ Migrations of remote copy relationships directly from FC-based partnerships to IP partnerships are not supported.
- ▶ IP partnerships between the two systems can be over IPv4 or IPv6 only, but not both.
- ▶ Virtual LAN (VLAN) tagging of the IP addresses that are configured for remote copy is supported starting with V7.4.
- ▶ Management IP and Internet SCSI (iSCSI) IP on the same port can be in a different network starting with V7.4.
- ▶ An added layer of security is provided by using Challenge Handshake Authentication Protocol (CHAP) authentication.
- ▶ Direct attached systems configurations are supported with the following restrictions:
 - Only two direct attach link are allowed.
 - The direct attach links must be on the same I/O group.
 - Use two port groups, where a port group contains only the two ports that are directly linked.

- ▶ Transmission Control Protocol (TCP) ports 3260 and 3265 are used for IP partnership communications. Therefore, these ports must be open in firewalls between the systems.
- ▶ Network address translation (NAT) between systems that are being configured in an IP Partnership group is not supported.
- ▶ Only a single Remote Copy data session per physical link can be established. It is intended that only one connection (for sending/receiving Remote Copy data) is made for each independent physical link between the systems.

Note: A physical link is the physical IP link between the two sites, A (local) and B (remote). Multiple IP addresses on local system A can be connected (by Ethernet switches) to this physical link. Similarly, multiple IP addresses on remote system B can be connected (by Ethernet switches) to the same physical link. At any point, only a single IP address on cluster A can form an RC data session with an IP address on cluster B.

- ▶ The maximum throughput is restricted based on the use of 1 Gbps or 10 Gbps Ethernet ports. The output varies based on distance (for example, round-trip latency) and quality of communication link (for example, packet loss):
 - One 1 Gbps port can transfer up to 110 megabytes per second (MBps) unidirectional, 190 MBps bidirectional
 - Two 1 Gbps ports can transfer up to 220 MBps unidirectional, 325 MBps bidirectional
 - One 10 Gbps port can transfer up to 240 MBps unidirectional, 350 MBps bidirectional
 - Two 10 Gbps port can transfer up to 440 MBps unidirectional, 600 MBps bidirectional

Note: The Bandwidth setting definition when the IP partnerships are created has changed. Previously, the bandwidth setting defaulted to 50 MB, and was the maximum transfer rate from the primary site to the secondary site for initial sync/resyncs of volumes.

The Link Bandwidth setting is now configured by using megabits (Mb), not MB. You set the Link Bandwidth setting to a value that the communication link can sustain, or to what is allocated for replication. The Background Copy Rate setting is now a percentage of the Link Bandwidth. The Background Copy Rate setting determines the available bandwidth for the initial sync and resyncs or for GM with Change Volumes.

5.4.3 VLAN support

VLAN tagging is supported for both iSCSI host attachment and IP replication. Hosts and remote-copy operations can connect to the system through Ethernet ports. Each traffic type has different bandwidth requirements, which can interfere with each other if they share IP connections. VLAN tagging creates two separate connections on the same IP network for different types of traffic. The system supports VLAN configuration on both IPv4 and IPv6 connections.

When the VLAN ID is configured for the IP addresses that are used for either iSCSI host attach or IP replication, the appropriate VLAN settings on the Ethernet network and servers must be configured correctly to avoid connectivity issues. After the VLANs are configured, changes to the VLAN settings disrupt iSCSI and IP replication traffic to and from the partnerships.

During the VLAN configuration for each IP address, the VLAN settings for the local and failover ports on two nodes of an I/O Group can differ. To avoid any service disruption, switches must be configured so the failover VLANs are configured on the local switch ports and the failover of IP addresses from a failing node to a surviving node succeeds. If failover VLANs are not configured on the local switch ports, there are no paths to Storwize V7000 system during a node failure and the replication fails.

Consider the following requirements and procedures when implementing VLAN tagging:

- ▶ VLAN tagging is supported for IP partnership traffic between two systems.
- ▶ VLAN provides network traffic separation at the layer 2 level for Ethernet transport.
- ▶ VLAN tagging by default is disabled for any IP address of a node port. You can use the CLI or GUI to set the VLAN ID for port IPs on both systems in the IP partnership.
- ▶ When a VLAN ID is configured for the port IP addresses that are used in remote copy port groups, appropriate VLAN settings on the Ethernet network must also be properly configured to prevent connectivity issues.

Setting VLAN tags for a port is disruptive. Therefore, VLAN tagging requires that you stop the partnership first before you configure VLAN tags. Then, restart again when the configuration is complete.

5.4.4 IP Compression

IBM Spectrum Virtualize version 7.7 introduced the IP compression capability that speed up replication cycles or that can allow use of less bandwidth. This feature reduces the volume of data that must be transmitted during remote copy operations by using compression capabilities similar to those experienced with existing Real-time Compression implementations.

No License: IP compression feature does not require an RtC software license.

The data compression is made within the IP replication component of the IBM Spectrum Virtualize code. It can be used with all the remote copy technology (Metro Mirror, Global Mirror, and Global Mirror Change Volume). The IP compression is supported, with some restrictions, in the following systems:

- ▶ SAN Volume controller with CF8 nodes
- ▶ SAN Volume controller with CG8 nodes
- ▶ SAN Volume controller with DH8 nodes
- ▶ SAN Volume controller with SV1 nodes
- ▶ FlashSystem V9000
- ▶ Storwize V7000 Gen1
- ▶ Storwize V7000 Gen2 and Gen2+
- ▶ Storwize V5000 Gen2

The IP compression can be enabled on hardware configurations that support RtC only. The IP compression feature provides two kinds of compression mechanisms: The hardware compression and software compression. The hardware compression is active when compression accelerator cards are available, otherwise software compression is used.

Hardware compression makes use of currently underused cards. The internal resources are shared between RACE and IP compression. Software compression uses the system CPU and might have an impact on heavily used systems.

To evaluate the benefits of the IP compression, the Comprestimator tool can be used to estimate the compression ratio of the data to be replicated. The IP compression can be enabled and disabled without stopping the remote copy relationship by using the `mkpartnership` and `chpartnership` commands with the `-compress` parameter. Furthermore, in systems with replication enabled in both directions, the IP compression can be enabled in only one direction. IP compression is supported for IPv4 and IPv6 partnerships.

Figure 5-29 reports the current compression limits by system type and compression mechanism.

Supported System	Max IP replication throughput per node	
	Software	Hardware Acceleration
SVC CF8	70 MB/s	N/A
SVC CG8	70 MB/s	N/A
SVC DH8/SV1 with RtC cards only	N/A	500 MB/s
FlashSystem V9000	N/A	500 MB/s
Storwize V5000 Gen2	140 MB/s	N/A
Storwize V7000 Gen1	70 MB/s	N/A
Storwize V7000 Gen2/Gen2+	N/A	500 MB/s

Figure 5-29 IP compression limits by systems and compression types

5.4.5 Remote copy groups

This section describes remote copy groups (or remote copy port groups) and different ways to configure the links between the two remote systems. The two systems can be connected to each other over one link or, at most, two links. To address the requirement to enable the systems to know about the physical links between the two sites, the concept of remote copy port groups was introduced.

Remote copy port group ID is a numerical tag that is associated with an IP port of system to indicate which physical IP link it is connected to. Multiple IBM Spectrum Virtualize nodes can be connected to the same physical long-distance link, and must therefore share a remote copy port group ID.

In scenarios with two physical links between the local and remote clusters, two remote copy port group IDs must be used to designate which IP addresses are connected to which physical link. This configuration must be done by the system administrator by using the GUI or the `cfgportip` CLI command.

Remember: IP ports on both partners must have been configured with identical remote copy port group IDs for the partnership to be established correctly.

The system IP addresses that are connected to the same physical link are designated with identical remote copy port groups. The IBM Spectrum Virtualize and Storwize family systems supports three remote copy groups: 0, 1, and 2.

The IP addresses are, by default, in remote copy port group 0. Ports in port group 0 are not considered for creating remote copy data paths between two systems. For partnerships to be established over IP links directly, IP ports must be configured in remote copy group 1 if a single inter-site link exists, or in remote copy groups 1 and 2 if two inter-site links exist.

You can assign one IPv4 address and one IPv6 address to each Ethernet port on the IBM Spectrum Virtualize and Storwize family systems. Each of these IP addresses can be shared between iSCSI host attach and the IP partnership. The user must configure the required IP address (IPv4 or IPv6) on an Ethernet port with a remote copy port group.

The administrator might want to use IPv6 addresses for remote copy operations and use IPv4 addresses on that same port for iSCSI host attach. This configuration also implies that for two systems to establish an IP partnership, both systems must have IPv6 addresses that are configured.

Administrators can choose to dedicate an Ethernet port for IP partnership only. In that case, host access must be explicitly disabled for that IP address and any other IP address that is configured on that Ethernet port.

Note: To establish an IP partnership, each Storwize V7000 canister must have only a single remote copy port group that is configured 1 or 2. The remaining IP addresses must be in remote copy port group 0.

Failover operations within and between port groups

Within one remote-copy port group, only one port from each system is selected for sending and receiving remote copy data at any one time. Therefore, on each system, at most one port for each remote-copy port group is reported as used.

If the IP partnership becomes unable to continue over an IP port, the system fails over to another port within that remote-copy port group. Some reasons this might occur are the switch to which it is connected fails, the node goes offline, or the cable that is connected to the port is unplugged.

For the IP partnership to continue during a failover, multiple ports must be configured within the remote-copy port group. If only one link is configured between the two systems, configure two ports (one per node) within the remote-copy port group. You can configure these two ports on two nodes within the same I/O group or within separate I/O groups. Configurations 4, 5, and 6 in IP partnership requirements are the supported dual-link configurations.

While failover is in progress, no connections in that remote-copy port group exist between the two systems in the IP partnership for a short time. Typically, failover completes within 30 seconds to 1 minute. If the systems are configured with two remote-copy port groups, the failover process within each port group continues independently of each other.

The disadvantage of configuring only one link between two systems is that, during a failover, a discovery is initiated. When the discovery succeeds, the IP partnership is reestablished. As a result, the relationships might stop, in which case a manual restart is required. To configure two intersystem links, you must configure two remote-copy port groups.

When a node fails in this scenario, the IP partnership can continue over the other link until the node failure is rectified. Failback then happens when both links are again active and available to the IP partnership. The discovery is triggered so that the active IP partnership data path is made available from the new IP address.

In a two-node system, or if there is more than one I/O Group and the node in the other I/O group has IP ports pre-configured within the remote-copy port group, the discovery is triggered. The discovery makes the active IP partnership data path available from the new IP address.

5.4.6 Supported configurations

Multiple IP partnership configurations are available depending on the number of physical links and the number of nodes. In the following sections, some example configurations are described.

Single inter-site link configurations

Consider two 2-node systems in IP partnership over a single inter-site link (with failover ports configured), as shown in Figure 5-30.

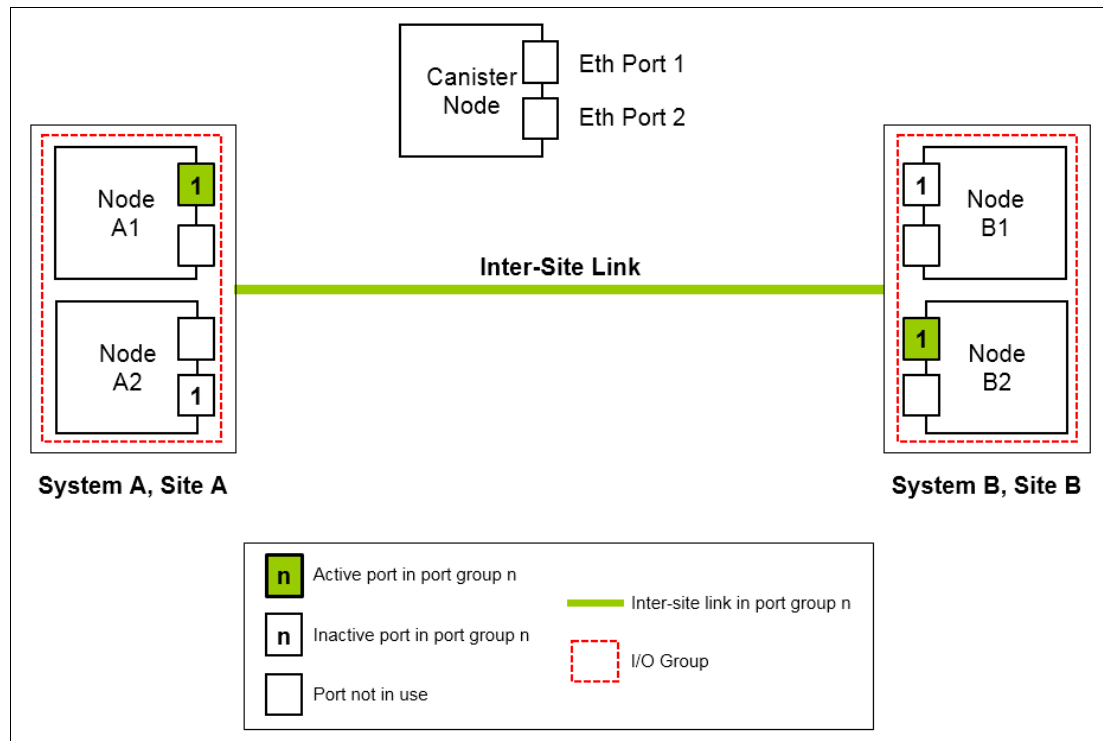


Figure 5-30 Only one remote copy group on each system and nodes with failover ports configured

Figure 5-30 shows two systems: System A and System B. A single remote copy port group 1 is configured on two Ethernet ports, one each on Node A1 and Node A2 on System A. Similarly, a single remote copy port group is configured on two Ethernet ports on Node B1 and Node B2 on System B.

Although two ports on each system are configured for remote copy port group 1, only one Ethernet port in each system actively participates in the IP partnership process. This selection is determined by a path configuration algorithm that is designed to choose data paths between the two systems to optimize performance.

The other port on the partner node in the I/O Group behaves as a standby port that is used during a node failure. If Node A1 fails in System A, IP partnership continues servicing replication I/O from Ethernet Port 2 because a failover port is configured on Node A2 on Ethernet Port 2.

However, it might take some time for discovery and path configuration logic to reestablish paths post failover. This delay can cause partnerships to change to Not_Present for that time. The details of the particular IP port that is actively participating in IP partnership is provided in the `1 sport ip` output (reported as used).

This configuration has the following characteristics:

- ▶ Each node in the I/O group has the same remote copy port group that is configured. However, only one port in that remote copy port group is active at any time at each system.
- ▶ If Node A1 in System A or Node B2 in System B fails in the respective systems, IP partnerships rediscovery is triggered and continues servicing the I/O from the failover port.
- ▶ The discovery mechanism that is triggered because of failover might introduce a delay where the partnerships momentarily change to the Not_Present state and recover.

Figure 5-31 shows a configuration with two 4-node systems in IP partnership over a single inter-site link (with failover ports configured).

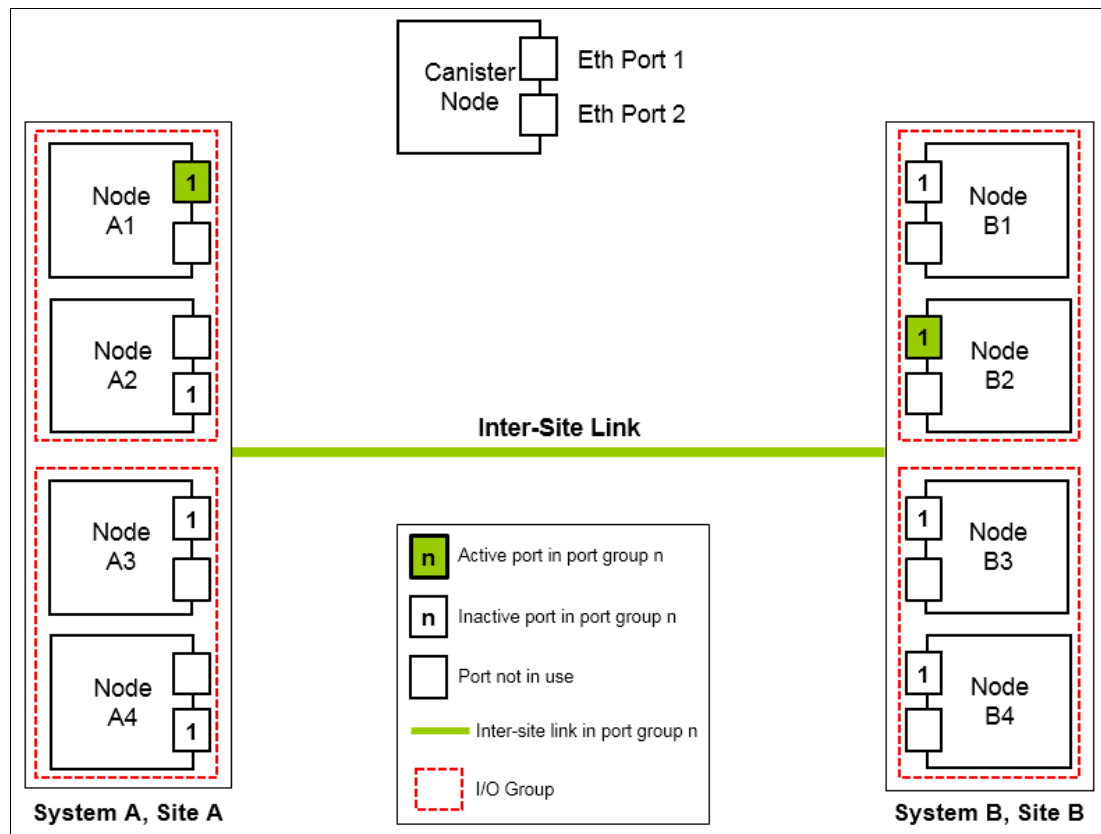


Figure 5-31 Multinode systems single inter-site link with only one remote copy port group

Figure 5-31 shows two 4-node systems: System A and System B. A single remote copy port group 1 is configured on nodes A1, A2, A3, and A4 on System A, Site A, and on nodes B1, B2, B3, and B4 on System B, Site B.

Although four ports are configured for remote copy group 1, only one Ethernet port in each remote copy port group on each system actively participates in the IP partnership process. Port selection is determined by a path configuration algorithm. The other ports play the role of standby ports.

If Node A1 fails in System A, the IP partnership selects one of the remaining ports that is configured with remote copy port group 1 from any of the nodes from either of the two I/O groups in System A. However, it might take some time (generally seconds) for discovery and path configuration logic to reestablish paths post failover. This process can cause partnerships to change to the `Not_Present` state.

This result causes remote copy relationships to stop. The administrator might need to manually verify the issues in the event log and start the relationships or remote copy consistency groups, if they do not automatically recover. The details of the particular IP port actively participating in the IP partnership process is provided in the `lspport ip` view (reported as used). This configuration has the following characteristics:

- ▶ Each node has the remote copy port group that is configured in both I/O groups. However, only one port in that remote copy port group remains active and participates in IP partnership on each system.
- ▶ If Node A1 in System A or Node B2 in System B encounter some failure in the system, IP partnerships discovery is triggered and continues servicing the I/O from the failover port.
- ▶ The discovery mechanism that is triggered because of failover might introduce a delay where the partnerships momentarily change to the `Not_Present` state and then recover.
- ▶ The bandwidth of the single link is used completely.

An eight-node system in IP partnership with four-node system over single inter-site link is shown in Figure 5-32.

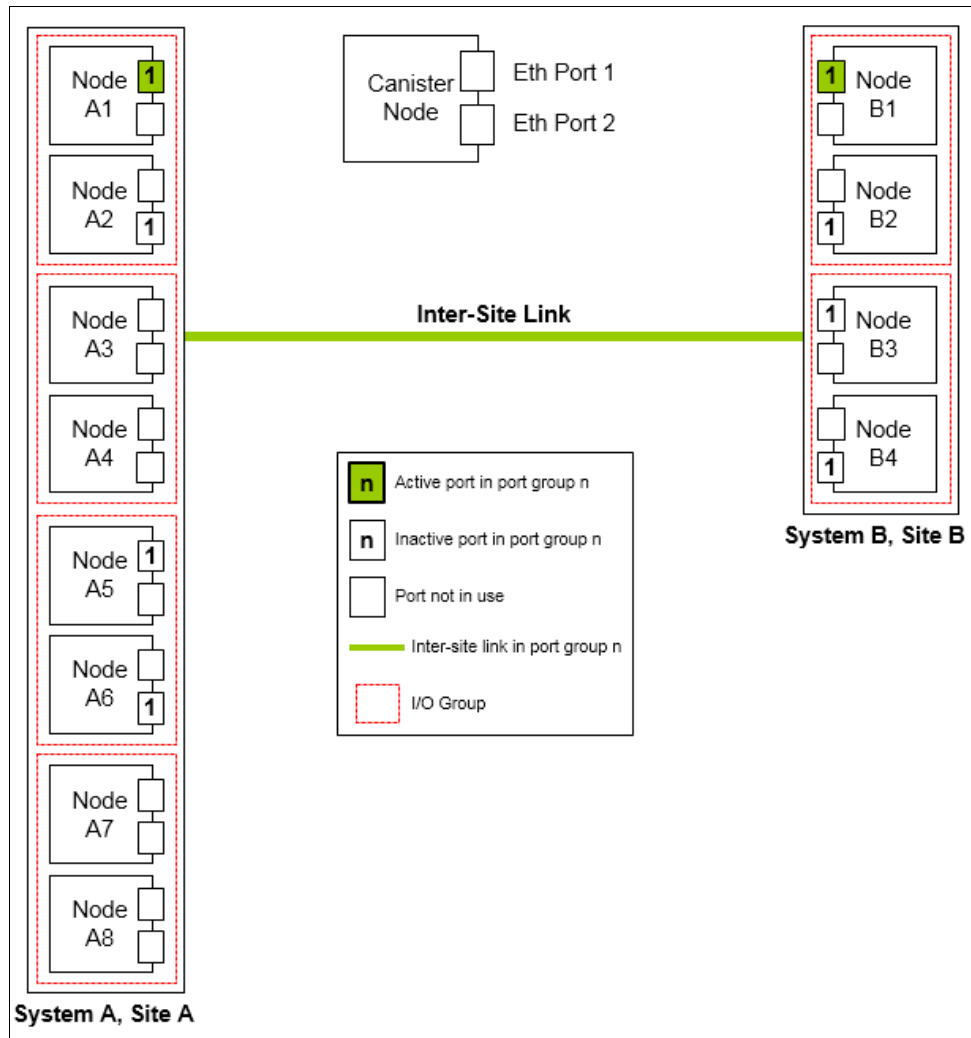


Figure 5-32 Multinode systems single inter-site link with only one remote copy port group

Figure 5-32 shows an eight-node system (System A in Site A) and a four-node system (System B in Site B). A single remote copy port group 1 is configured on nodes A1, A2, A5, and A6 on System A at Site A. Similarly, a single remote copy port group 1 is configured on nodes B1, B2, B3, and B4 on System B.

Although there are four I/O groups (eight nodes) in System A, any two I/O groups at maximum are supported to be configured for IP partnerships. If Node A1 fails in System A, IP partnership continues using one of the ports that is configured in remote copy port group from any of the nodes from either of the two I/O groups in System A.

However, it might take some time for discovery and path configuration logic to reestablish paths post-failover. This delay might cause partnerships to change to the Not_Present state.

This process can lead to remote copy relationships stopping. The administrator must manually start them if the relationships do not auto-recover. The details of which particular IP port is actively participating in IP partnership process is provided in `1sport ip` output (reported as used).

This configuration has the following characteristics:

- ▶ Each node has the remote copy port group that is configured in both the I/O groups that are identified for participating in IP Replication. However, only one port in that remote copy port group remains active on each system and participates in IP Replication.
- ▶ If the Node A1 in System A or the Node B2 in System B fails in the system, the IP partnerships trigger discovery and continue servicing the I/O from the failover ports.
- ▶ The discovery mechanism that is triggered because of failover might introduce a delay where the partnerships momentarily change to the Not_Present state and then recover.
- ▶ The bandwidth of the single link is used completely.

Two inter-site link configurations

A two 2-node systems with two inter-site links configuration is depicted in Figure 5-33.

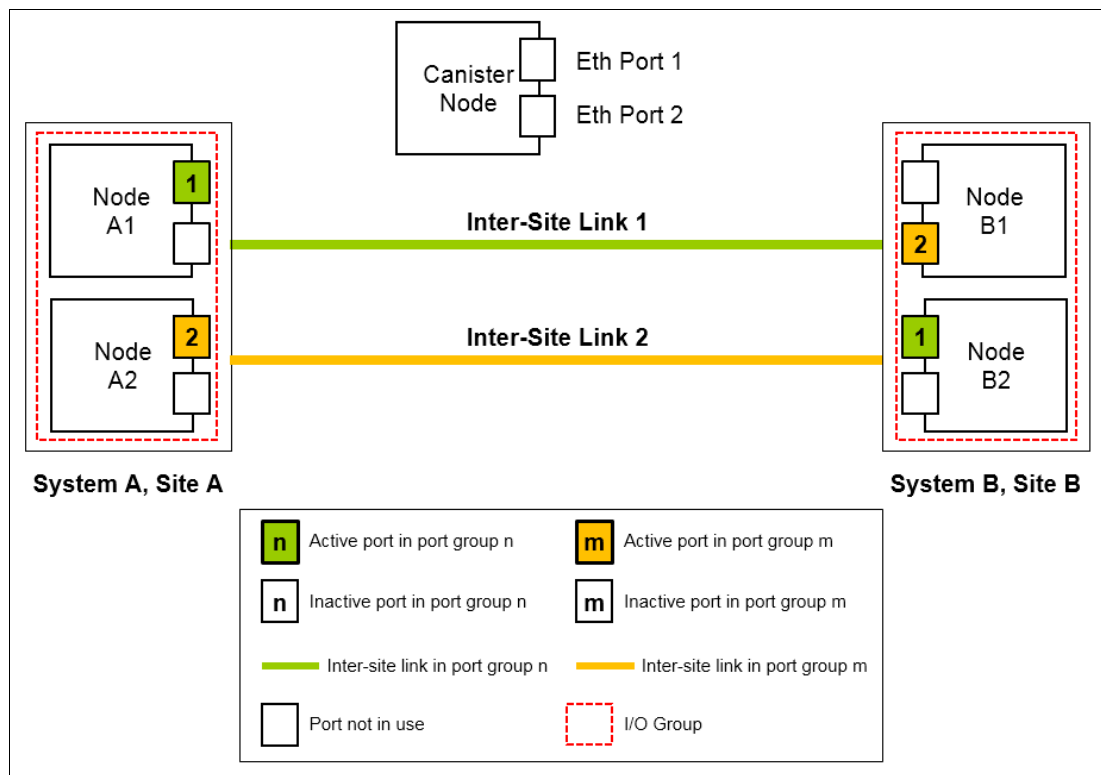


Figure 5-33 Dual links with two remote copy groups on each system configured

As shown in Figure 5-33, remote copy port groups 1 and 2 are configured on the nodes in System A and System B because two inter-site links are available. In this configuration, the failover ports are not configured on partner nodes in the I/O group. Rather, the ports are maintained in different remote copy port groups on both of the links. They can remain active and participate in IP partnership by using both of the links.

However, if either of the nodes in the I/O group fail (that is, if Node A1 on System A fails), the IP partnership continues only from the available IP port that is configured in remote copy port group 2. Therefore, the effective bandwidth of the two links is reduced to 50% because only the bandwidth of a single link is available until the failure is resolved.

This configuration has the following characteristics:

- ▶ There are two inter-site links, and two remote copy port groups are configured.
- ▶ Each node has only one IP port in remote copy port group 1 or 2.
- ▶ Both the IP ports in the two remote copy port groups participate simultaneously in IP partnerships. Therefore, both of the links are used.
- ▶ During node failure or link failure, the IP partnership traffic continues from the other available link and the port group. Therefore, if two links of 10 Mbps each are available and you have 20 Mbps of effective link bandwidth, bandwidth is reduced to 10 Mbps only during a failure.
- ▶ After the node failure or link failure is resolved and failback happens, the entire bandwidth of both of the links is available as before.

A configuration with two 4-node systems in IP partnership with dual inter-site links is shown in Figure 5-34.

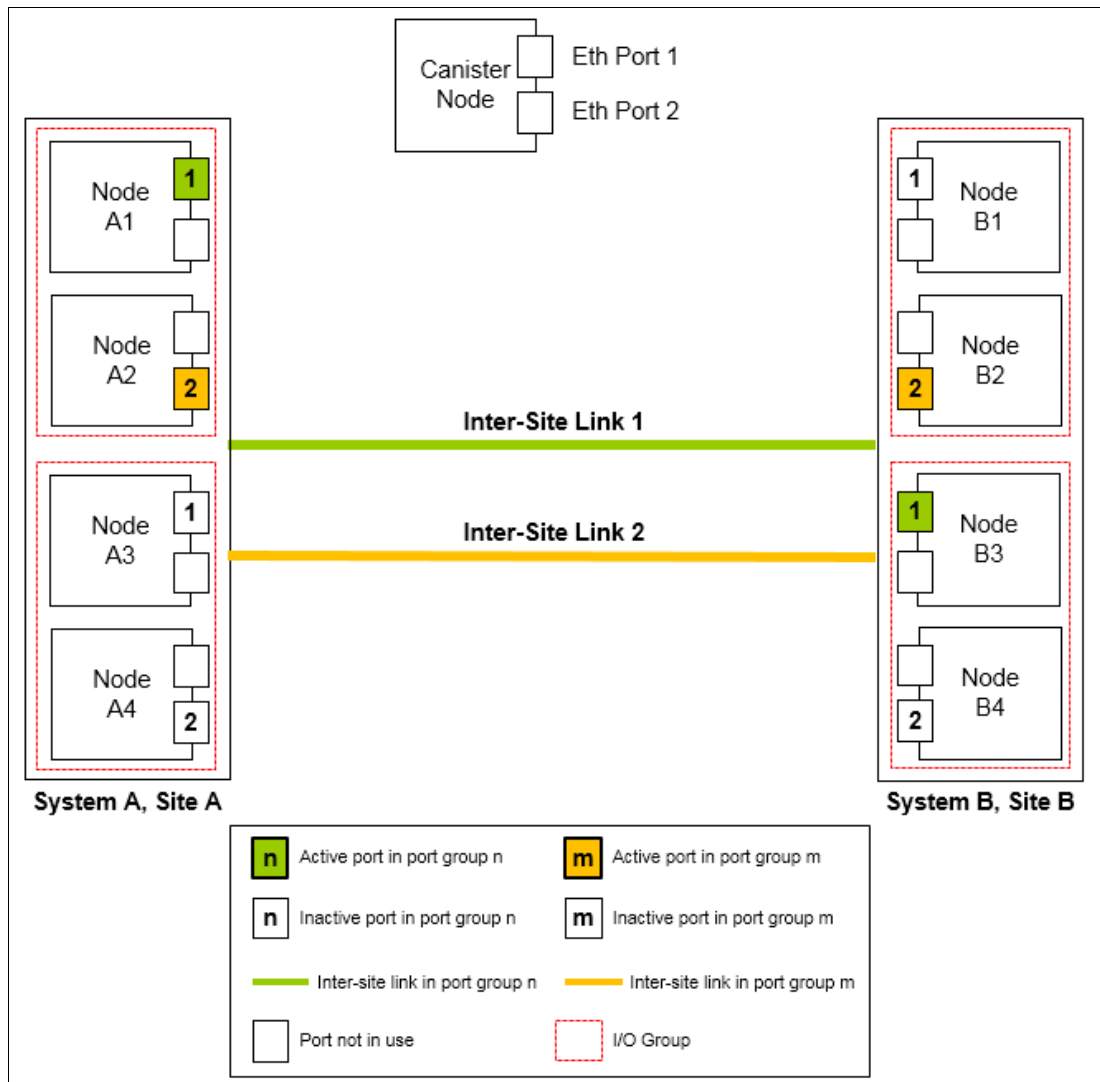


Figure 5-34 Multinode systems with dual inter-site links between the two systems

Figure 5-34 shows two 4-node systems: System A and System B. This configuration is an extension of Configuration 5 to a multinode multi-I/O group environment.

As seen in this configuration, there are two I/O groups. Each node in the I/O group has a single port that is configured in remote copy port groups 1 or 2.

Although two ports are configured in remote copy port groups 1 and 2 on each system, only one IP port in each remote copy port group on each system actively participates in IP partnership. The other ports that are configured in the same remote copy port group act as standby ports during a failure. Which port in a configured remote copy port group participates in IP partnership at any moment is determined by a path configuration algorithm.

In this configuration, if Node A1 fails in System A, IP partnership traffic continues from Node A2 (that is, remote copy port group 2). At the same time, the failover also causes discovery in remote copy port group 1. Therefore, the IP partnership traffic continues from Node A3 on which remote copy port group 1 is configured. The details of the particular IP port that is actively participating in IP partnership process is provided in the `lsport ip` output (reported as used).

This configuration has the following characteristics:

- ▶ Each node has the remote copy port group that is configured in the I/O groups 1 or 2. However, only one port per system in both remote copy port groups remains active and participates in IP partnership.
- ▶ Only a single port per system from each configured remote copy port group participates simultaneously in IP partnership. Therefore, both of the links are used.
- ▶ During node failure or port failure of a node that is actively participating in IP partnership, IP partnership continues from the alternative port because another port is in the system in the same remote copy port group, but in a different I/O Group.
- ▶ The pathing algorithm can start discovery of available port in the affected remote copy port group in the second I/O group and pathing is reestablished. This process restores the total bandwidth, so both of the links are available to support IP partnership.

Finally, an eight-node system in IP partnership with a four-node system over dual inter-site links is depicted in Figure 5-35.

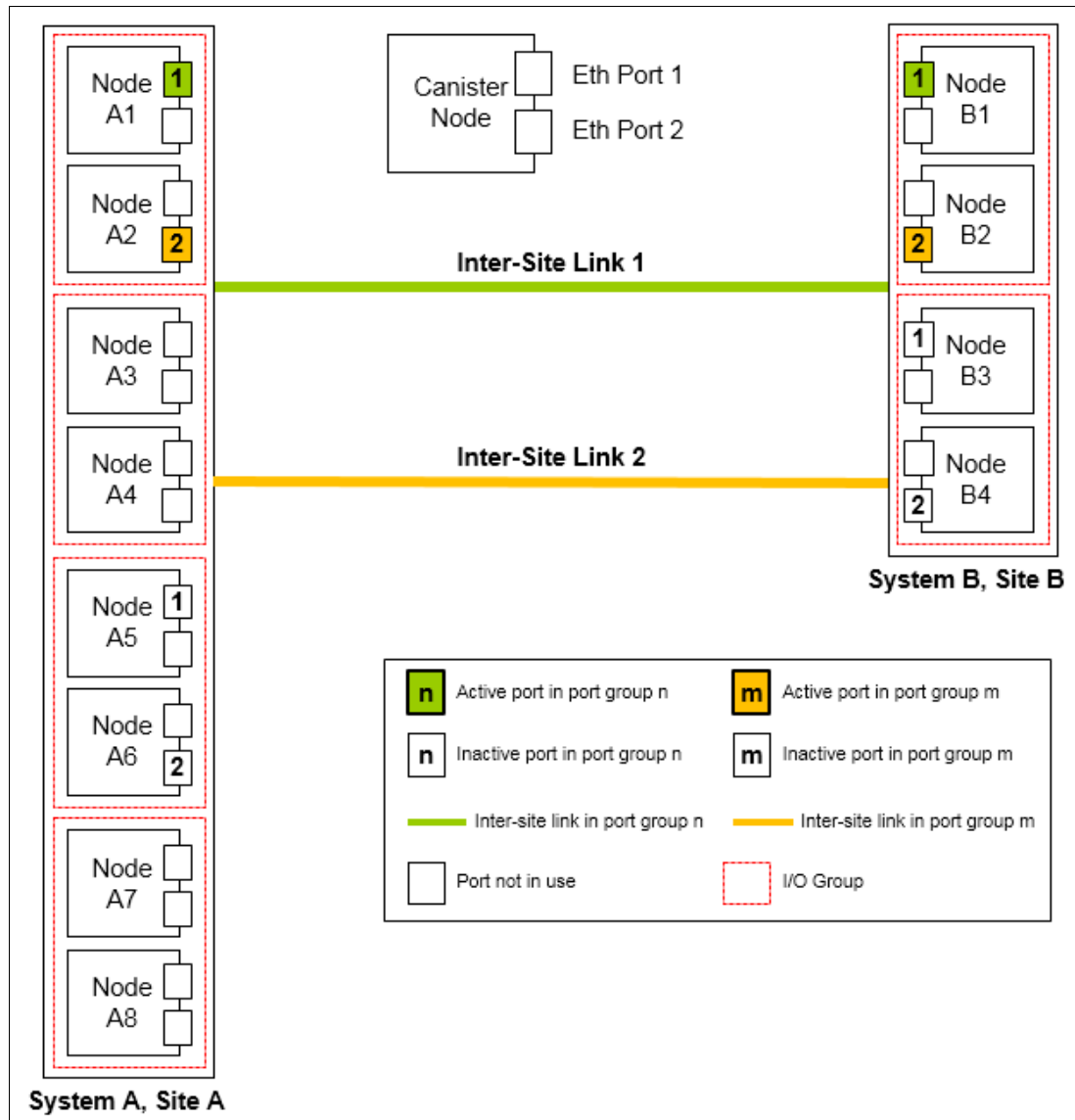


Figure 5-35 Multinode systems with dual inter-site links between the two systems

Figure 5-35 shows an eight-node System A in Site A and a four-node System B in Site B. Because a maximum of two I/O groups in IP partnership is supported in a system, although there are four I/O groups (eight nodes), nodes from only two I/O groups' are configured with remote copy port groups in System A. The remaining or all of the I/O groups can be configured to be remote copy partnerships over FC.

In this configuration, there are two links and two I/O groups that are configured with remote copy port groups 1 and 2. However, path selection logic is managed by an internal algorithm. Therefore, this configuration depends on the pathing algorithm to decide which of the nodes actively participate in IP partnership. Even if Node A5 and Node A6 are configured with remote copy port groups properly, active IP partnership traffic on both of the links can be driven from Node A1 and Node A2 only.

If Node A1 fails in System A, IP partnership traffic continues from Node A2 (that is, remote copy port group 2). The failover also causes IP partnership traffic to continue from Node A5 on which remote copy port group 1 is configured. The details of the particular IP port actively participating in IP partnership process is provided in the `l sport ip` output (reported as used).

This configuration has the following characteristics:

- ▶ There are two I/O Groups with nodes in those I/O groups that are configured in two remote copy port groups because there are two inter-site links for participating in IP partnership. However, only one port per system in a particular remote copy port group remains active and participates in IP partnership.
- ▶ One port per system from each remote copy port group participates in IP partnership simultaneously. Therefore, both of the links are used.
- ▶ If a node or port on the node that is actively participating in IP partnership fails, the remote copy (RC) data path is established from that port because another port is available on an alternative node in the system with the same remote copy port group.
- ▶ The path selection algorithm starts discovery of available ports in the affected remote copy port group in the alternative I/O groups and paths are reestablished. This process restores the total bandwidth across both links.
- ▶ The remaining or all of the I/O groups can be in remote copy partnerships with other systems.

5.4.7 Native IP replication performance consideration

A number of factors affect the performance of an IP partnership. Some of these factors are latency, link speed, number of intersite links, host I/O, MDisk latency, and hardware. Since the introduction with version 7.2, many improvements have been made to make the IP replication better performing and more reliable.

With V7.7, a new workload distribution algorithm was introduced that optimize the usage of the 10 Gbps ports. Nevertheless, in presence of poor quality networks that have significant packet loss and high latency, the actual usable bandwidth might decrease considerably.

Figure 5-36 shows the throughput trend for a 1 Gbps port in respect of the packet loss ratio and the latency.

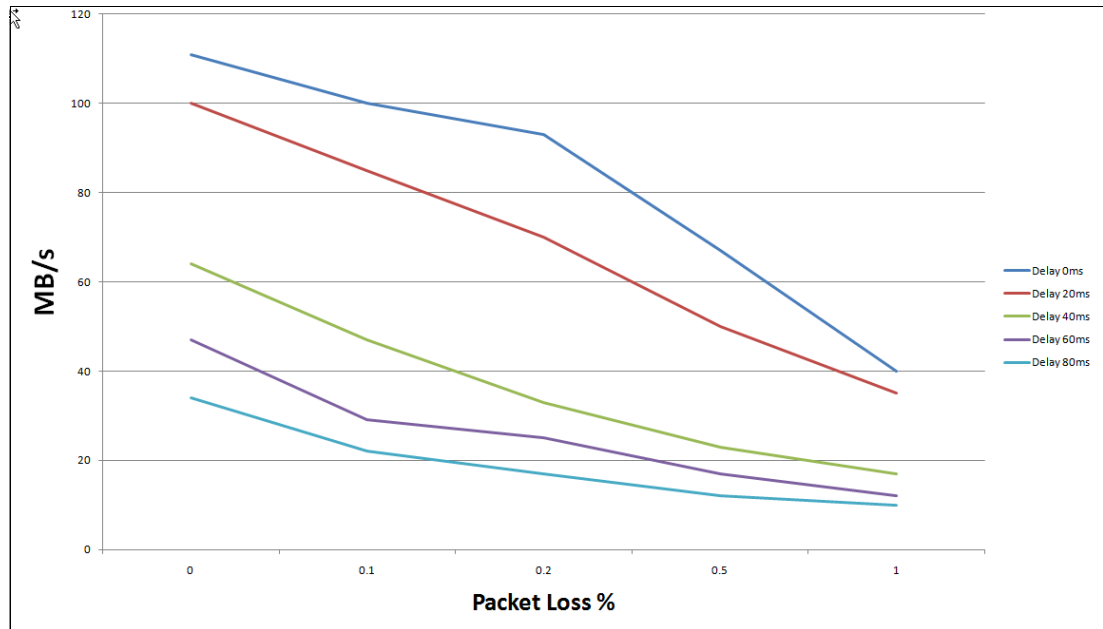


Figure 5-36 1 Gbps port throughput trend

The chart shows how the combined effect of the packet loss and the latency can lead to a throughput reduction of more than 85%. For these reasons, the IP replication option should only be considered for the replication configuration not requiring high quality and performing networks. Due to its characteristic of low-bandwidth requirement, the Global Mirror Change Volume is the preferred solution with the IP replication.

The following recommendations might help improve this performance when using compression and IP partnership in the same system:

- ▶ Using nodes older than SAN Volume Controller CG8 with IP partnership, or Global Mirror and compression in the same I/O group is not recommended.
- ▶ To use the IP partnership on a multiple I/O group system that has nodes older than SAN Volume Controller 2145-CG8 and compressed volumes, configure ports for the IP partnership in I/O groups that do not contain compressed volumes.
- ▶ To use the IP partnership on Storwize Family product that has compressed volumes, configure ports for the IP partnership in I/O groups that do not contain compressed volumes.
- ▶ For SAN Volume Controller CG8 nodes using IP partnership, or Global Mirror and compression, update your hardware to an “RPQ 8S1296 hardware update for 2145-CG8”.
- ▶ If you require more than a 100 MBps throughput per intersite link with IP partnership on a node that uses compression, consider virtualizing the system with SAN Volume Controller 2145-SV1.
- ▶ Use a different port for iSCSI host I/O and IP partnership traffic. Also, use a different VLAN ID for iSCSI host I/O and IP partnership traffic.

5.5 Volume Mirroring

By using Volume Mirroring, you can have two physical copies of a volume that provide a basic RAID-1 function. These copies can be in the same storage pool or in different storage pools, with different extent sizes of the storage pool. Typically the two copies are allocated in different storage pools.

The first storage pool contains the original (primary volume copy). If one storage controller or storage pool fails, a volume copy is not affected if it has been placed on a different storage controller or in a different storage pool.

If a volume is created with two copies, both copies use the same virtualization policy. However, you can have two copies of a volume with different virtualization policies. In combination with *thin-provisioning*, each mirror of a volume can be thin-provisioned, compressed or fully allocated, and in striped, sequential, or image mode.

A mirrored (secondary) volume has all of the capabilities of the primary volume copy. It also has the same restrictions (for example, a mirrored volume is owned by an I/O Group, just as any other volume). This feature also provides a *point-in-time copy* function that is achieved by “splitting” a copy from the volume. However, the mirrored volume does not address other forms of mirroring based on Remote Copy (Global or Metro Mirror functions), which mirrors volumes across I/O Groups or clustered systems.

One copy is the primary copy, and the other copy is the secondary copy. Initially, the first volume copy is the primary copy. You can change the primary copy to the secondary copy if required.

Figure 5-37 provides an overview of Volume Mirroring.

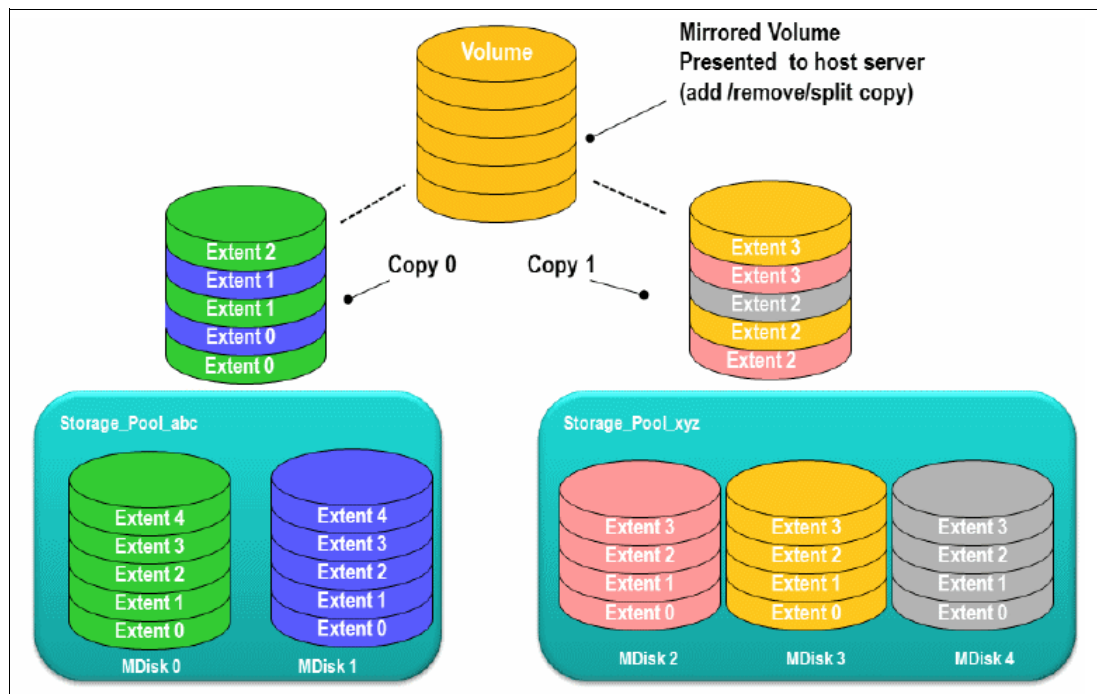


Figure 5-37 Volume Mirroring overview

5.5.1 Read and write operations

Read and write operations behavior depends on the status of the copies and on other environment settings. During the initial synchronization or a resynchronization, only one of the copies is in synchronized status, and all the reads are directed to this copy. The write operations are directed to both copies.

When both copies are synchronized, the write operations are again directed to both copies. The read operations usually are directed to the primary copy, unless the system is configured in Enhanced Stretched Cluster topology. With this system topology and the enablement of site awareness capability, the concept of primary copy still exists, but is not more relevant. The read operation follows the site affinity.

For example, consider an Enhanced Stretched Cluster configuration with mirrored volumes with one copy in Site A and the other in Site B. If a host I/O read is attempted to a mirrored disk through a Spectrum Virtualize Node in Site A, then the I/O read is directed to the copy in Site A, if available. Similarly, a host I/O read attempted through a node in Site B goes to the Site B copy.

Important: For best performance, keep consistency between Hosts, Nodes, and Storage Controller site affinity as long as possible.

During back-end storage failure, note the following points:

- ▶ If one of the mirrored volume copies is temporarily unavailable, the volume remains accessible to servers.
- ▶ The system remembers which areas of the volume are written and resynchronizes these areas when both copies are available.
- ▶ The remaining copy can service read I/O when the failing one is offline, without user intervention.

5.5.2 Volume mirroring use cases

Volume Mirroring offers the capability to provide extra copies of the data that can be used for High Availability solutions and data migration scenarios. You can convert a non-mirrored volume into a mirrored volume by adding a copy. When a copy is added using this method, the cluster system synchronizes the new copy so that it is the same as the existing volume. You can convert a mirrored volume into a non-mirrored volume by deleting one copy or by splitting one copy to create a new non-mirrored volume.

Access: Servers can access the volume during the synchronization processes described.

You can use mirrored volumes to provide extra protection for your environment or to perform a migration. This solution offers several options:

- ▶ Stretched Cluster configurations
Standard and Enhanced Stretched Cluster configuration uses the Volume Mirroring feature to implement the data availability across the sites.
- ▶ Export to Image mode
This option allows you to move storage from *managed mode* to *image mode*. This option is useful if you are using IBM Spectrum Virtualize or Storwize V7000 as a migration device. For example, suppose vendor A's product cannot communicate with vendor B's product, but you need to migrate existing data from vendor A to vendor B.

Using “Export to image mode” allows you to migrate data by using the Copy Services functions and then return control to the native array, while maintaining access to the hosts.

► Import to Image mode

This option allows you to import an existing storage MDisk or logical unit number (LUN) with its existing data from an external storage system, without putting metadata on it. The existing data remains intact. After you import it, the volume mirroring function can be used to migrate the storage to the other locations, while the data remains accessible to your hosts.

► Volume cloning using Volume Mirroring and then using the Split into New Volume option

This option allows any volume to be cloned without any interruption to the host access. You have to create two mirrored copies of data and then break the mirroring with the split option to make two independent copies of data. This option doesn't apply to already mirrored volumes.

► Volume pool migration using the volume mirroring option

This option allows any volume to be moved between storage pools without any interruption to the host access. You might use this option to move volumes as an alternative to Migrate to Another Pool function. Compared to the Migrate to Another Pool function, volume mirroring provides more manageability because it can be suspended and resumed anytime and also it allows you to move volumes among pools with different extent sizes. This option doesn't apply to already mirrored volumes.

► Volume capacity saving change

This option allows you to modify the capacity saving characteristics of any volume from standard to thin provisioned or compressed and vice versa, without any interruption to host access. This option works the same as the volume pool migration but specifying a different capacity saving for the newly created copy. This option doesn't apply to already mirrored volumes.

When you use Volume Mirroring, consider how quorum candidate disks are allocated. Volume Mirroring maintains some state data on the quorum disks. If a quorum disk is not accessible and Volume Mirroring is unable to update the state information, a mirrored volume might need to be taken offline to maintain data integrity. To ensure the high availability of the system, ensure that multiple quorum candidate disks, which are allocated on different storage systems, are configured.

Quorum disk consideration: Mirrored volumes can be taken offline if there is no quorum disk available. This behavior occurs because synchronization status for mirrored volumes is recorded on the quorum disk. To protect against mirrored volumes being taken offline, follow the guidelines for setting up quorum disks.

The following are other Volume Mirroring usage cases and characteristics:

► Creating a mirrored volume:

– The maximum number of copies is two.

– Both copies are created with the same virtualization policy.

To have a volume mirrored using different policies, you need to add a volume copy with a different policy to a volume that has only one copy.

– Both copies can be located in different storage pools. The first storage pool that is specified contains the primary copy.

– It is not possible to create a volume with two copies when specifying a set of MDisk.

- ▶ Add a volume copy to an existing volume:
 - The volume copy to be added can have a different space allocation policy.
 - Two existing volumes with one copy each cannot be merged into a single mirrored volume with two copies.
- ▶ Remove a volume copy from a mirrored volume:
 - The volume remains with only one copy.
 - It is not possible to remove the last copy from a volume.
- ▶ Split a volume copy from a mirrored volume and create a new volume with the split copy:
 - This function is only allowed when the volume copies are synchronized. Otherwise, use the **-force** command.
 - It is not possible to recombine the two volumes after they have been split.
 - Adding and splitting in one workflow enables migrations that are not currently allowed.
 - The split volume copy can be used as a means for creating a point-in-time copy (clone).
- ▶ Repair/validate in three ways. This compares volume copies and performs these functions:
 - Reports the first difference found. It can iterate by starting at a specific LBA by using the **-startlba** parameter.
 - Creates virtual medium errors where there are differences.
 - Corrects the differences that are found (reads from primary copy and writes to secondary copy).
- ▶ View to list volumes affected by a back-end disk subsystem being offline:
 - Assumes that a standard use is for mirror between disk subsystems.
 - Verifies that mirrored volumes remain accessible if a disk system is being shut down.
 - Reports an error in case a quorum disk is on the back-end disk subsystem.
- ▶ Expand or shrink a volume:
 - This function works on both of the volume copies at once.
 - All volume copies always have the same size.
 - All copies must be synchronized before expanding or shrinking them.
- ▶ Delete a volume. When a volume gets deleted, all copies get deleted.
- ▶ Migration commands apply to a specific volume copy.
- ▶ Out-of-sync bitmaps share the bitmap space with FlashCopy and Metro Mirror/Global Mirror. Creating, expanding, and changing I/O groups might fail if there is insufficient memory.
- ▶ GUI views contain volume copy identifiers.

5.5.3 Mirrored volume components

Note the following points regarding mirrored volume components:

- ▶ A mirrored volume is always composed of two copies (copy 0 and copy1).
- ▶ A volume that is not mirrored consists of a single copy (which for reference might be copy 0 or copy 1).

A mirrored volume looks the same to upper-layer clients as a non-mirrored volume. That is, upper layers within the cluster software, such as FlashCopy and Metro Mirror/Global Mirror, and storage clients, do not know whether a volume is mirrored. They all continue to handle the volume as they did before without being aware of whether the volume is mirrored.

Figure 5-38 shows the attributes of a volume and Volume Mirroring.

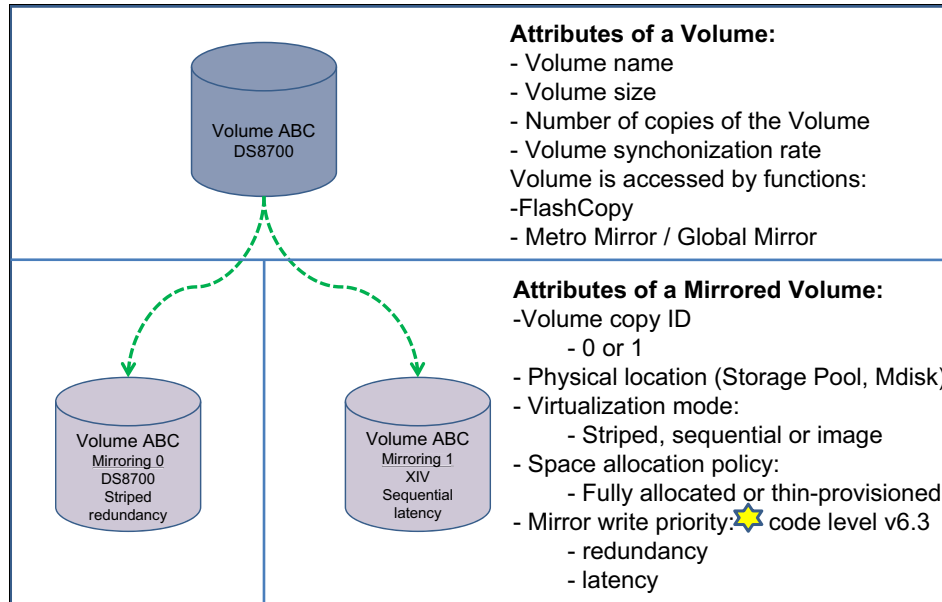


Figure 5-38 Attributes of a volume and Volume Mirroring

In Figure 5-38, XIV and IBM DS8700 illustrate that a mirrored volume can use different storage devices.

5.5.4 Volume Mirroring synchronization options

As soon as a volume is created with two copies, copies are in the *out-of-sync* state. The primary volume copy (located in the first specified storage pool) is defined as in sync and the secondary volume copy as out of sync. The secondary copy is synchronized through the synchronization process.

This process runs at the default synchronization rate of 50 (Table 5-9 on page 236), or at the defined rate while creating or modifying the volume (see 5.5.5, “Volume Mirroring performance considerations” on page 236 for the effect of the copy rate setting). Once the synchronization process is completed, the volume mirroring copies are *in-sync* state.

Prior to IBM Spectrum Virtualize V7.5, the `-fmt disk` parameter added to the `mkvdisk` command ensured that both copies are overwritten with zeros. After this process, the volume came online and they could be considered as *in-sync*. Starting with V7.5, the format process is initiated by default at the time of the volume creation without keeping the volume offline. The format processing overwrite with zeros only the Copy 0 and then synchronize the Copy 1.

You can specify that a volume is synchronized (`-createsync` parameter), even if it is not. Using this parameter can cause data corruption if the primary copy fails and leaves an unsynchronized secondary copy to provide data. Using this parameter can cause loss of read stability in unwritten areas if the primary copy fails, data is read from the primary copy, and then different data is read from the secondary copy. To avoid data loss or read stability loss, use this parameter only for a primary copy that has been formatted and not written to.

Another example use case for `-createsync` is for a newly created mirrored volume where both copies are thin provisioned or compressed because no data has been written to disk and unwritten areas return zeros (0). If the synchronization between the volume copies has been lost, the resynchronization process is incremental. This term means that only grains that have been written to need to be copied, and then get synchronized volume copies again.

The progress of the volume mirror synchronization can be obtained from the GUI or by using the `1svdisksyncprogress` command.

5.5.5 Volume Mirroring performance considerations

Because the writes of mirrored volumes always occur to both copies, mirrored volumes put more workload on the cluster, the back-end disk subsystems, and the connectivity infrastructure. The mirroring is symmetrical, and writes are only acknowledged when the write to the last copy completes. The result is that if the volumes copies are on storage pools with different performance characteristics, the slowest storage pool determines the performance of writes to the volume. This performance applies when writes must be destaged to disk.

Tip: Locate volume copies of one volume on storage pools of the same or similar characteristics. Usually, if only good read performance is required, you can place the primary copy of a volume in a storage pool with better performance. Because the data is always only read from one volume copy, reads are not faster than without Volume Mirroring.

However, be aware that this is only true when both copies are synchronized. If the primary is out of sync, then reads are submitted to the other copy. Finally, note that these considerations do not apply to IBM Spectrum Virtualize systems in Enhanced Stretched Cluster configuration where the primary copy attribute is irrelevant.

Synchronization between volume copies has a similar impact on the cluster and the back-end disk subsystems as FlashCopy or data migration. The synchronization rate is a property of a volume that is expressed as a value of 0 - 100. A value of 0 disables synchronization. Table 5-9 shows the relationship between the *rate value* and the *data copied per second*.

Table 5-9 Relationship between the rate value and the data copied per second

User-specified rate attribute value per volume	Data copied/sec
0	Synchronization is disabled
1 - 10	128 KB
11 - 20	256 KB
21 - 30	512 KB
31 - 40	1 MB
41 - 50	2 MB ** 50% is the default value
51 - 60	4 MB
61 - 70	8 MB
71 - 80	16 MB
81 - 90	32 MB
91 - 100	64 MB

Rate attribute value: The rate attribute is configured on each volume that you want to mirror. The default value of a new volume mirror is 50%.

In large, IBM Spectrum Virtualize or Storwize system configurations, the settings of the copy rate can considerably affect the performance in scenarios where a back-end storage failure occurs. For instance, consider a scenario where a failure of a back-end storage controller is affecting one copy of 300 mirrored volumes. The host continues the operations by using the remaining copy.

When the failed controller comes back online, the resynchronization process for all the 300 mirrored volumes starts at the same time. With a copy rate of 100 for each volume, this process would add a theoretical workload of 18.75 GB/s, which will drastically overload the system.

The general suggestion for the copy rate settings is then to evaluate the impact of massive resynchronization and set the parameter accordingly. Consider setting the copy rate to high values for initial synchronization only, and with a limited number of volumes at a time. Alternatively, consider defining a volume provisioning process that allows the safe creation of already synchronized mirrored volumes, as described in 5.5.4, “Volume Mirroring synchronization options” on page 235.

Volume mirroring I/O Time-out configuration

A mirrored volume has pointers to the two copies of data, usually in different storage pools, and each write completes on both copies before the host receives I/O completion status. For a synchronized mirrored volume, if a write I/O to a copy has failed or a long timeout has expired, then system has completed all available controller level Error Recovery Procedures (ERPs). In this case, that copy is taken offline and goes out of sync. The volume remains online and continues to service I/O requests from the remaining copy.

The *Fast Failover* feature isolates hosts from temporarily poorly-performing back-end storage of one Copy at the expense of a short interruption to redundancy. The fast failover feature behavior is that during normal processing of host write I/O, the system submits writes to both copies with a timeout of 10 seconds (20 seconds for stretched volumes). If one write succeeds and the other write takes longer than 5 seconds, then the slow write is stopped. The Fibre Channel abort sequence can take around 25 seconds.

When the stop is completed, one copy is marked as out of sync and the host write I/O completed. The overall fast failover ERP aims to complete the host I/O in around 30 seconds (40 seconds for stretched volumes).

In v6.3.x and later, the fast failover can be set for *each* mirrored volume by using the `chvdisk` command and the `mirror_write_priority` attribute settings:

- ▶ *Latency* (default value): A short timeout prioritizing low host latency. This option enables the fast failover feature.
- ▶ *Redundancy*: A long timeout prioritizing redundancy. This option indicates a copy that is slow to respond to a write I/O can use the full ERP time. The response to the I/O is delayed until it completes to keep the copy in sync if possible. This option disables the fast failover feature.

Volume Mirroring ceases to use the slow copy for 4 - 6 minutes, and subsequent I/O data is not affected by a slow copy. Synchronization is suspended during this period. After the copy suspension completes, Volume Mirroring resumes, which allows I/O data and synchronization operations to the slow copy that will, typically, quickly complete the synchronization.

If another I/O times out during the synchronization, then the system stops using that copy again for 4 - 6 minutes. If one copy is always slow, then the system tries it every 4 - 6 minutes and the copy gets progressively more out of sync as more grains are written. If fast failovers are occurring regularly, there is probably an underlying performance problem with the copy's back-end storage.

The preferred `mirror_write_priority` setting for the Enhanced Stretched Cluster configurations is *latency*.

5.5.6 Bitmap space for out-of-sync volume copies

The grain size for the synchronization of volume copies is 256 KB. One grain takes up one bit of bitmap space. 20 MB of bitmap space supports 40 TB of mirrored volumes. This relationship is the same as the relationship for copy services (Global and Metro Mirror) and standard FlashCopy with a grain size of 256 KB (Table 5-10).

Table 5-10 Relationship of bitmap space to Volume Mirroring address space

Function	Grain size in KB	1 byte of bitmap space gives a total of	4 KB of bitmap space gives a total of	1 MB of bitmap space gives a total of	20 MB of bitmap space gives a total of	512 MB of bitmap space gives a total of
Volume Mirroring	256	2 MB of volume capacity	8 GB of volume capacity	2 TB of volume capacity	40 TB of volume capacity	1024 TB of volume capacity

Shared bitmap space: This bitmap space on one I/O group is shared between Metro Mirror, Global Mirror, FlashCopy, and Volume Mirroring.

The command to create Mirrored Volumes can fail if there is not enough space to allocate bitmaps in the target I/O Group. To verify and change the space allocated and available on each I/O Group with the CLI, see the Example 5-4.

Example 5-4 A `lsiogrp` and `chiogrp` command example

```

IBM_2145:SVC_ESC:superuser>lsiogrp
id name          node_count vdisk_count host_count site_id site_name
0 io_grp0        2          9           0          0
1 io_grp1        0          0           0          0
2 io_grp2        0          0           0          0
3 io_grp3        0          0           0          0
4 recovery_io_grp 0          0           0          0
IBM_2145:SVC_ESC:superuser>lsiogrp io_grp0
id 0
name io_grp0
node_count 2
vdisk_count 9
host_count 0
flash_copy_total_memory 20.0MB
flash_copy_free_memory 19.9MB
remote_copy_total_memory 20.0MB
remote_copy_free_memory 19.9MB
mirroring_total_memory 20.0MB

```

```

mirroring_free_memory 20.0MB
raid_total_memory 40.0MB
raid_free_memory 40.0MB
.
lines removed for brevity
.
IBM_2145:SVC_ESC:superuser>chiogrp -feature mirror -size 64 io_grp0
IBM_2145:SVC_ESC:superuser>lsiogrp io_grp0
id 0
name io_grp0
node_count 2
vdisk_count 9
host_count 0
flash_copy_total_memory 20.0MB
flash_copy_free_memory 19.9MB
remote_copy_total_memory 20.0MB
remote_copy_free_memory 19.9MB
mirroring_total_memory 64.0MB
mirroring_free_memory 64.0MB
.
lines removed for brevity
.

```

To verify and change the space allocated and available on each I/O Group with the GUI, see Figure 5-39.

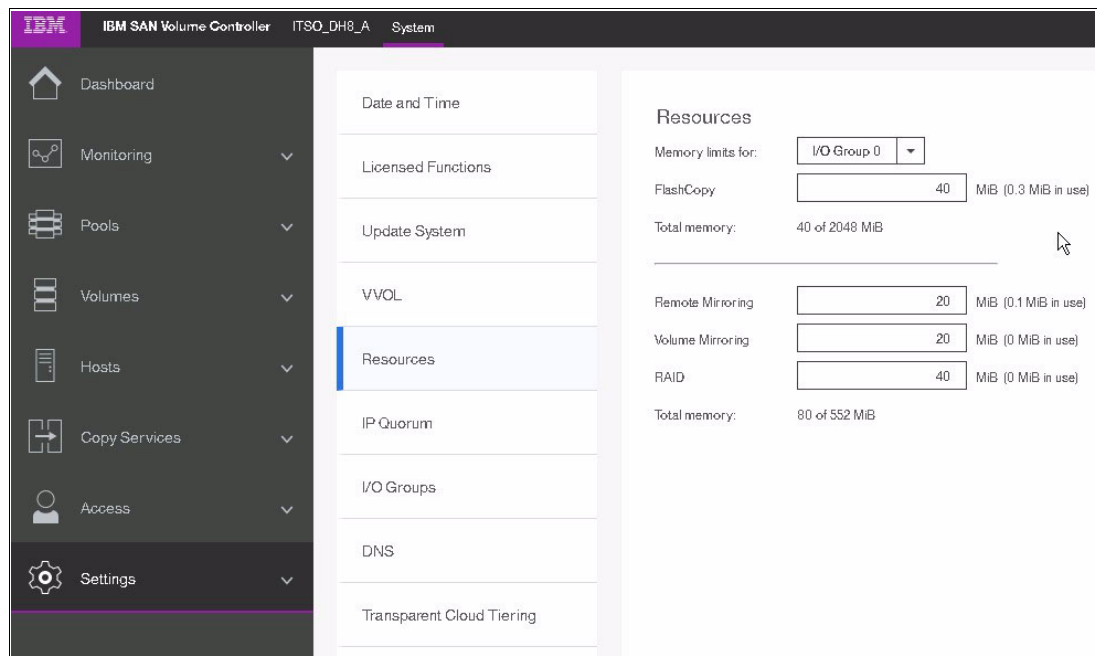


Figure 5-39 IOgrp feature example



Hosts

This chapter describes the guidelines on how to configure host systems on IBM Spectrum Virtualize by following several preferred practices. A *host system* is an Open Systems computer that is connected to the switch through a Fibre Channel (FC) interface.

One of the most important parts of tuning, troubleshooting, and performance is the host that is attached to IBM Spectrum Virtualize. You must consider the following areas for performance:

- ▶ The use of multipathing and bandwidth (physical capability of SAN and back-end storage)
- ▶ Understanding how your host performs I/O and the types of I/O
- ▶ The use of measurement and test tools to determine host performance and for tuning

This chapter supplements the IBM System Storage Spectrum Virtualize V8.1 documentation at Knowledge Center, which is available at:

<https://ibm.biz/BdjV9E>

This chapter includes the following sections:

- ▶ Configuration guidelines
- ▶ N-Port ID Virtualization
- ▶ Host pathing
- ▶ I/O queues
- ▶ Host clustering and reserves
- ▶ AIX hosts
- ▶ Virtual I/O Server
- ▶ Windows hosts
- ▶ Linux hosts
- ▶ Solaris hosts
- ▶ VMware server
- ▶ Monitoring

6.1 Configuration guidelines

When IBM Spectrum Virtualize is used to manage storage that is connected to any host, you must follow basic configuration guidelines. These guidelines pertain to these considerations:

- ▶ The number of paths through the fabric that are allocated to the host
- ▶ The number of host ports to use
- ▶ The approach for spreading the hosts across I/O groups
- ▶ Logical unit number (LUN) mapping
- ▶ The correct size of virtual disks (volumes) to use

6.1.1 Host levels and host object name

When a new host is configured to IBM Spectrum Virtualize, determine first the preferred operating system, driver, firmware, and supported host bus adapters (HBAs) to prevent unanticipated problems because of untested levels. Before you provision a new host into IBM Spectrum Virtualize at the preferred levels, see *IBM System Storage Interoperation Center (SSIC)*, available at the following website:

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

When you are creating the host, use the host name from the host as the host object name in IBM Spectrum Virtualize to aid in configuration updates or problem determination in the future.

6.1.2 Host cluster

IBM Spectrum Virtualize software supports host clusters starting with V7.7.1. On V8.1.0 the host cluster feature became available using the GUI. The host cluster allows a user to create a group of hosts to form a cluster, which is treated as one single entity. This technique allows multiple hosts to have access to the same set of volumes.

Volumes that are mapped to that host cluster are assigned to all members of the host cluster with the same SCSI ID.

A typical use-case is to define a host cluster that contains all the WWPNs belonging to the hosts participating in a host operating system based cluster, such as IBM PowerHA® or Microsoft Cluster Server (MSCS).

The following commands have been added to deal with host clusters:

- ▶ **lshostcluster**
- ▶ **lshostclustermember**
- ▶ **lshostclustervolumemap**
- ▶ **addhostclustermember**
- ▶ **chostcluster**
- ▶ **mkhost** (with parameter **-hostcluster** to create the host in one existing cluster)
- ▶ **rmhostclustermember**
- ▶ **rmhostcluster**
- ▶ **rmvolumehostclustermap**

On V8.1, host clusters can be added by using the GUI. For more information, see *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.1*, SG24-7933.

For GUI users, when you use the *Map to Host or Host Cluster* function, it now allows you to let the system assign the SCSI ID for the volume or to manually assign the SCSI ID. For ease of management purposes, it is suggested to use separate ranges of SCSI IDs for hosts and host clusters, for example, you can use SCSI IDs 0 - 99 to non-cluster host volumes, and above 100 for the cluster host volumes. When you choose the option **System Assign** the system automatically assigns the SCSI IDs starting from the first available in the sequence.

If you choose **Self Assign**, the system enables you to select the SCSI IDs manually for each volume, and on the right part of the screen it shows the SCSI IDs that are already used by the selected host/host cluster, as shown in Figure 6-1.

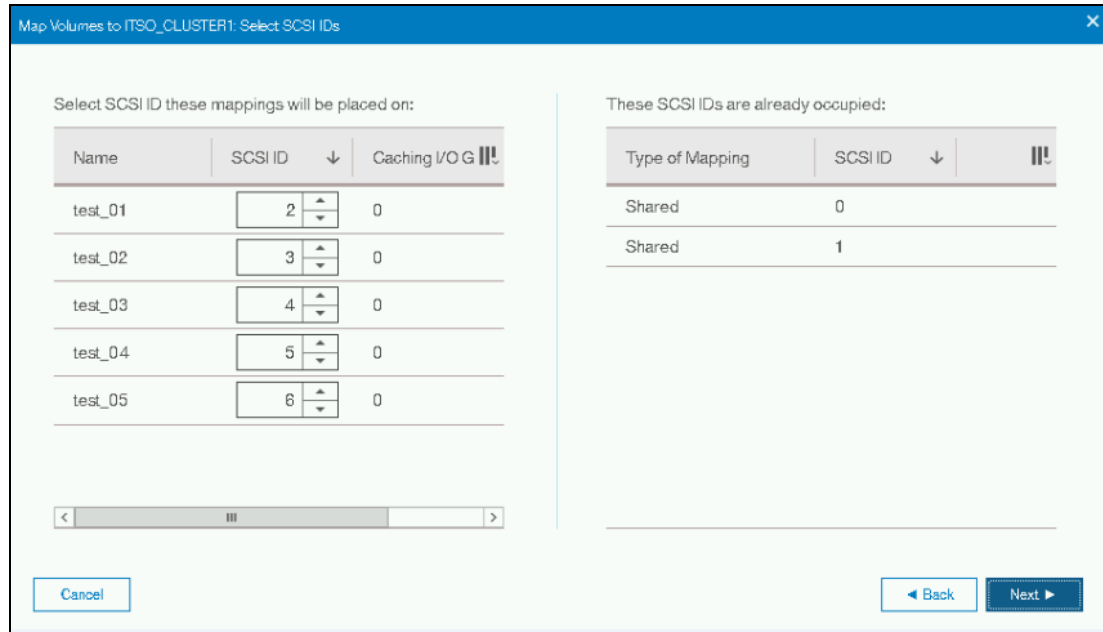


Figure 6-1 SCSI ID assignment on volume mappings

Note: Although extra care is always recommended when dealing with hosts, IBM Spectrum Virtualize does not allow you to join a host into a host cluster if it already has a volume mapping with a SCSI ID that also exists in the host cluster, as shown below:

```
IBM_2145:ITS0_DH8_LAB:superuser>addhostclustermember -host ITS0_HOST3
ITS0_CLUSTER1

CMMVC9068E Hosts in the host cluster have conflicting SCSI ID's for their
private mappings.

IBM_2145:ITS0_DH8_LAB:superuser>
```

6.1.3 The number of paths

Based on our general experience, it is best to limit the total number of paths from any host to IBM Spectrum Virtualize. Limit the total number of paths that the multipathing software on each host is managing to four paths, even though the maximum supported is eight paths. Following these rules solves many issues with high port fan-outs, fabric state changes, and host memory management, and it improves performance.

For more information about maximum host configurations and restrictions, see *V8.1.0 Configuration Limits and Restrictions* which are available at:

- ▶ IBM Spectrum Virtualize
<http://www.ibm.com/support/docview.wss?uid=ssg1S1010644>
- ▶ Storwize V7000
<http://www.ibm.com/support/docview.wss?uid=ssg1S1010643>

The most important reason to limit the number of paths that are available to a host from IBM Spectrum Virtualize is for error recovery, failover, and failback purposes. The overall time for handling errors by a host is reduced. In addition, resources within the host are greatly reduced when you remove a path from the multipathing management.

Two path configurations have only one path to each node, which is a supported configuration but not preferred for most configurations. In previous IBM Spectrum Virtualize releases, host configuration information is available by using the *IBM System Storage Spectrum Virtualize V5.1.0 - Host Attachment Guide*, SC26-7905, which is available at this website:

ftp://ftp.software.ibm.com/storage/san/sanvc/V5.1.0/pubs/English/SVC_Host_Attach_Guide.pdf

For V8.1, this information is now consolidated into the IBM System Storage Spectrum Virtualize Information Center, which is available at:

- ▶ IBM Spectrum Virtualize
<https://ibm.biz/BdjV9E>
- ▶ Storwize V7000
<https://ibm.biz/BdjVCE>

6.1.4 Host ports

When you are using host ports that are connected to IBM Spectrum Virtualize, limit the number of physical ports to two ports on two different physical adapters. Each port is zoned to one target port in each IBM Spectrum Virtualize node, which limits the number of total paths to four, preferably on separate redundant SAN fabrics.

If four host ports are preferred for maximum redundant paths, the requirement is to zone each host adapter to one IBM Spectrum Virtualize target port on each node (for a maximum of eight paths). The benefits of path redundancy are outweighed by the host memory resource utilization that is required for more paths.

When working with clustered hosts, the preferred practice is to use the host cluster object on IBM Spectrum Virtualize. Previously, the advice was to use a single host and register all the initiators under it. A single host object can have a maximum of 32 initiators, while a host cluster object can have 128 hosts, therefore, our suggestion is to use the host cluster, instead of a single host object.

Preferred practice: Keep Fibre Channel tape (including Virtual Tape Libraries) and Fibre Channel disks on separate HBAs. These devices have two different data patterns when operating in their optimum mode. The switching between them can cause unwanted processor usage and performance slowdown for the applications.

6.1.5 Port masking

You can use a port mask to control the node target ports that a host can access. Using local FC port masking, you can set which ports can be used for node-to-node/intracluster communication. By using remote FC port masking, you can set which ports can be used for replication communication.

Using port masking is preferable because mixed traffic of host, back-end, intracluster, and replication might cause congestion and buffer-to-buffer credit exhaustion. This kind of traffic could otherwise result in heavy degradation of performance in your IBM Spectrum Virtualize environment.

The port mask is a 64-bit field that applies to all nodes in the cluster. In the local FC port masking, you can set a port to be dedicated to node-to-node/intracluster traffic by setting a 1 to that port. Also, by using remote FC port masking, you can set which ports can be used for replication traffic by setting 1 to that port.

If a port has a 0 in the mask, it means no traffic of that type is allowed. So, in a local FC port map, a 0 means no node-to-node traffic happens, and a 0 on the remote FC port masking means no replication traffic happens on that port. Therefore, if a port has a 0 on both local and remote FC port masking, only host/back-end traffic is allowed on it. The port mask can vary depending on the number of ports that your IBM Spectrum Virtualize HBA cards have. For an example of portmask on nodes 2145-DH8 and 2145-SV1, see Figure 6-2.

Slot/Port	Port #	SAN	4-port Nodes	8-port Nodes with 2 port cards	8-port Nodes with 4 port cards	12-port Nodes	16-port Nodes
S1P1	1	A / 1	Host/Storage/Inter-node	Host/Storage	Host/Storage	Host/Storage	Host/Storage
S1P2	2	B / 2	Host/Storage/Inter-node	Host/Storage	Host/Storage	Host/Storage	Host/Storage
S1P3	3	A / 1	Host/Storage/Replication*	--	Inter-node	Host/Storage	Host/Storage
S1P4	4	B / 2	Host/Storage/Replication*	--	Host/Storage or Replication**	Host/Storage	Host/Storage
S2P1	5	A / 1		Host/Storage	Host/Storage	Inter-node	Inter-node
S2P2	6	B / 2		Host/Storage	Host/Storage	Inter-node	Inter-node
S2P3	7	A / 1		--	Host/Storage or Replication**	Host/Storage or Replication**	Host/Storage or Replication**
S2P4	8	B / 2		--	Inter-node	Host/Storage	Host/Storage
S3P1	9	A / 1		Inter-node		Host/Storage	Host/Storage
S3P2	10	B / 2		Host/Storage or Replication**		Host/Storage or Replication**	Host/Storage or Replication**
S3P3	11	A / 1		--		Inter-node or Host/Storage	Inter-node or Host/Storage
S3P4	12	B / 2		--		Inter-node or Host/Storage	Inter-node or Host/Storage
S5P1	13	A / 1		Host/Storage or Replication**			Host/Storage
S5P2	14	B / 2		Inter-node			Host/Storage
S5P3	15	A / 1		--			Host/Storage
S5P4	16	B / 2		--			Host/Storage
localfcportmask			With Rep 0011 / No Rep 1111	10010000	10000100	110000110000	0000110000110000
remotefcportmask			1100	01100000	01001000	001001000000	0000001001000000

* Inter-node if no replication planned
 ** Use for Host/Storage in case no replication is in place.

Figure 6-2 Portmask on nodes DH8 and SV1

How to set a port mask using the CLI and GUI

The command to apply a local FC port mask on CLI is `chsystem -localfcportmask mask`. The command to apply a remote FC port mask is `chsystem -partnerfcportmask mask`.

If you are using the GUI, click **Settings** → **Network** → **Fibre Channel Ports**. Then you can select the use of a port from these options:

- ▶ Setting none means no node-to-node and no replication traffic is allowed. Only host and storage traffic is allowed.
- ▶ Setting local means only node-to-node traffic is allowed.
- ▶ Setting remote means that only replication traffic is allowed.

Figure 6-3 shows the port mask in the GUI.

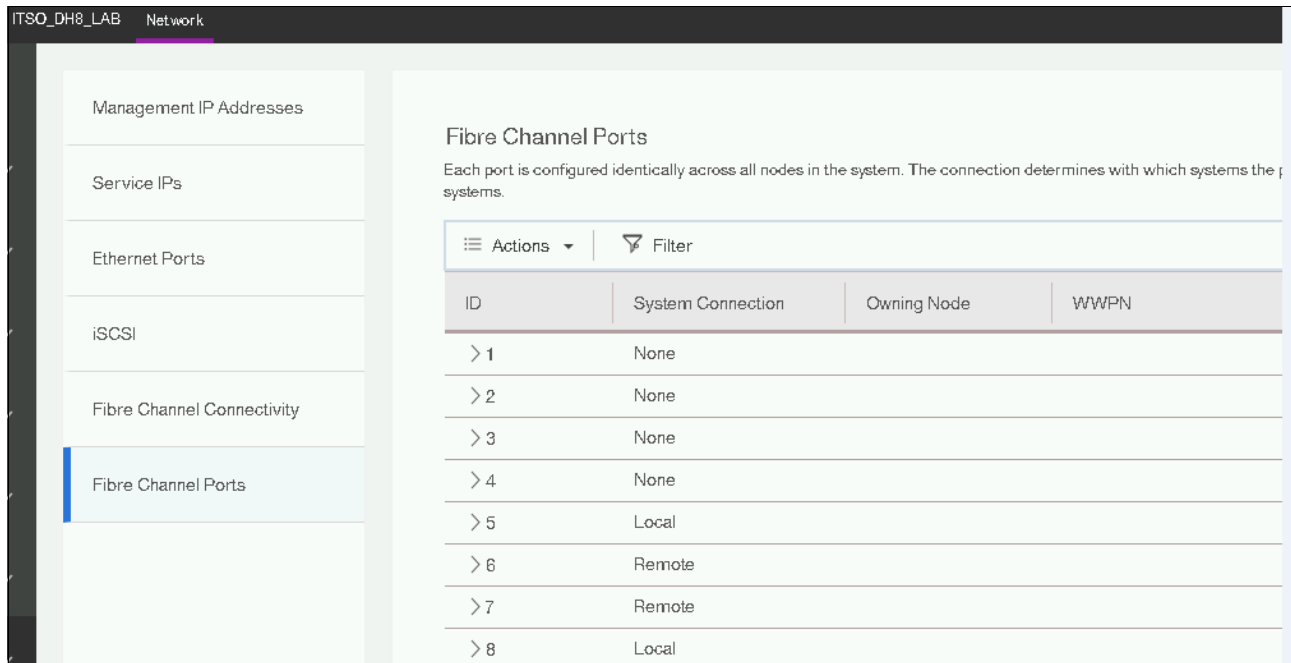


Figure 6-3 Fiber Channel Ports menu

6.1.6 Host to I/O group mapping

An *I/O grouping* consists of two IBM Spectrum Virtualize nodes that share management of volumes within a cluster. Use a single I/O group (iogrp) for all volumes that are allocated to a particular host. This guideline has the following benefits:

- ▶ Minimizes port fan-outs within the SAN fabric
- ▶ Maximizes the potential host attachments to IBM Spectrum Virtualize because maximums are based on I/O groups
- ▶ Fewer target ports to manage within the host

The number of host ports and host objects that are allowed per I/O group depends on the switch fabric type. For more information about the maximum configurations, see V8.1 Configuration Limits and Restrictions:

- ▶ IBM Spectrum Virtualize
<http://www.ibm.com/support/docview.wss?uid=ssg1S1010644>
- ▶ Storwize V7000
<http://www.ibm.com/support/docview.wss?uid=ssg1S1010643>

6.1.7 Volume size as opposed to quantity

In general, host resources, such as memory and processing time, are used up by each storage LUN that is mapped to the host. For each extra path, more memory can be used, and a portion of more processing time is also required.

The user can control this effect by using fewer larger LUNs rather than many small LUNs. However, you might need to tune queue depths and I/O buffers to support controlling the memory and processing time efficiently. If a host does not have tunable parameters, such as on the Windows operating system, the host does not benefit from large volume sizes. AIX greatly benefits from larger volumes with a smaller number of volumes and paths that are presented to it.

6.1.8 Host volume mapping

When you create a host mapping, the host ports that are associated with the host object can detect the LUN that represents the volume up to eight FC ports (the four ports on each node in an I/O group). Nodes always present the logical unit (LU) that represents a specific volume with the same LUN on all ports in an I/O group.

This LUN mapping is called the Small Computer System Interface ID (SCSI ID). The IBM Spectrum Virtualize software automatically assigns the next available ID if none is specified. In addition, a unique identifier, called the *LUN serial number*, is on each volume.

You can allocate the operating system volume of the SAN boot as the lowest SCSI ID (zero for most hosts), and then allocate the various data disks. If you share a volume among multiple hosts, consider controlling the SCSI ID so that the IDs are identical across the hosts. This consistency ensures ease of management at the host level and prevents potential issues during IBM Spectrum Virtualize updates and even node reboots, mostly for ESX operating systems.

If you are using image mode to migrate a host to IBM Spectrum Virtualize, allocate the volumes in the same order that they were originally assigned on the host from the back-end storage.

The `lshostvdiskmap` command displays a list of VDisk (volumes) that are mapped to a host. These volumes are recognized by the specified host. Example 6-1 shows the syntax of the `lshostvdiskmap` command that is used to determine the SCSI ID and the UID of volumes.

Example 6-1 The lshostvdiskmap command

```
svcinfolshostvdiskmap -delim
```

Example 6-2 shows the results of using the `lshostvdiskmap` command.

Example 6-2 Output of using the lshostvdiskmap command

```
svcinfolsvdiskhostmap -delim : EEXCLS_HBin01
id:name:SCSI_id:host_id:host_name:wwpn:vdisk_UID
950:EEXCLS_HBin01:14:109:HDMCENTEX1N1:10000000C938CFDF:600507680191011D480000000000466
950:EEXCLS_HBin01:14:109:HDMCENTEX1N1:10000000C938D01F:600507680191011D480000000000466
950:EEXCLS_HBin01:13:110:HDMCENTEX1N2:10000000C938D65B:600507680191011D480000000000466
950:EEXCLS_HBin01:13:110:HDMCENTEX1N2:10000000C938D3D3:600507680191011D480000000000466
950:EEXCLS_HBin01:14:111:HDMCENTEX1N3:10000000C938D615:600507680191011D480000000000466
950:EEXCLS_HBin01:14:111:HDMCENTEX1N3:10000000C938D612:600507680191011D480000000000466
950:EEXCLS_HBin01:14:112:HDMCENTEX1N4:10000000C938CFBD:600507680191011D480000000000466
950:EEXCLS_HBin01:14:112:HDMCENTEX1N4:10000000C938CE29:600507680191011D480000000000466
950:EEXCLS_HBin01:14:113:HDMCENTEX1N5:10000000C92EE1D8:600507680191011D480000000000466
950:EEXCLS_HBin01:14:113:HDMCENTEX1N5:10000000C92EDFFE:600507680191011D480000000000466
```

Note: Example 6-2 shows the same volume mapped to five different hosts, but host 110 has a different SCSI ID from the other four hosts. This is an example of a non-recommended practice which can lead to loss of access in some situations due to SCSI ID mismatch.

In this example, vDisk 10 has a unique device identifier (UID, which is represented by the UID field) of 6005076801958001500000000000000A (see Example 6-3), but the SCSI_id that host2 uses for access is 0.

Example 6-3 VDisk 10 with a UID

```
id:name:SCSI_id:vdisk_id:vdisk_name:wwpn:vdisk_UID
2:host2:0:10:vdisk10:0000000000000000ACA:6005076801958001500000000000000A
2:host2:1:11:vdisk11:0000000000000000ACA:6005076801958001500000000000000B
2:host2:2:12:vdisk12:0000000000000000ACA:6005076801958001500000000000000C
2:host2:3:13:vdisk13:0000000000000000ACA:6005076801958001500000000000000D
2:host2:4:14:vdisk14:0000000000000000ACA:6005076801958001500000000000000E
```

If you are using IBM multipathing software (Subsystem Device Driver Device Specific Module (SDDSM)), the **datapath query device** command shows the vdisk_UID (unique identifier), which enables easier management of volumes. The equivalent command for Subsystem Device Driver Path Control Module (SDDPCM) is the **pcmpath query device** command.

Host mapping from more than one I/O group

The SCSI ID field in the host mapping might not be unique for a volume for a host because it does not completely define the uniqueness of the LUN. The target port is also used as part of the identification. If two I/O groups of volumes are assigned to a host port, one set starts with SCSI ID 0 and then increments (by default). The SCSI ID for the second I/O group also starts at zero and then increments by default.

Example 6-4 shows this type of hostmap. Volume s-0-6-4 and volume s-1-8-2 both have a SCSI ID of ONE, yet they have different LUN serial numbers.

Example 6-4 Host mapping for one host from two I/O groups

```
IBM_2145:ITS0CL1:admin>svcinfolshostvdiskmap senegal
wwpn                vdisk_UID
0 senegal           1          60          s-0-6-4      210000E08B89CCC2 60050768018101BF28000000000000A8
0 senegal           2          58          s-0-6-5      210000E08B89CCC2 60050768018101BF28000000000000A9
0 senegal           3          57          s-0-5-1      210000E08B89CCC2 60050768018101BF28000000000000AA
0 senegal           4          56          s-0-5-2      210000E08B89CCC2 60050768018101BF28000000000000AB
0 senegal           5          61          s-0-6-3      210000E08B89CCC2 60050768018101BF28000000000000A7
0 senegal           6          36          big-0-1      210000E08B89CCC2 60050768018101BF28000000000000B9
0 senegal           7          34          big-0-2      210000E08B89CCC2 60050768018101BF28000000000000BA
0 senegal           1          40          s-1-8-2      210000E08B89CCC2 60050768018101BF28000000000000B5
0 senegal           2          50          s-1-4-3      210000E08B89CCC2 60050768018101BF28000000000000B1
0 senegal           3          49          s-1-4-4      210000E08B89CCC2 60050768018101BF28000000000000B2
0 senegal           4          42          s-1-4-5      210000E08B89CCC2 60050768018101BF28000000000000B3
0 senegal           5          41          s-1-8-1      210000E08B89CCC2 60050768018101BF28000000000000B4
```

Example 6-5 on page 249 shows the **datapath query device** output of this Windows host. The order of the volumes of the two I/O groups is reversed from the hostmap. Volume s-1-8-2 is first, followed by the rest of the LUNs from the second I/O group, then volume s-0-6-4, and the rest of the LUNs from the first I/O group. Most likely, Windows discovered the second set of LUNs first. However, the relative order within an I/O group is maintained.

Example 6-5 Using datapath query device for the hostmap

C:\Program Files\IBM\Subsystem Device Driver>datapath query device

Total Devices : 12

DEV#: 0 DEVICE NAME: Disk1 Part0 TYPE: 2145 POLICY: OPTIMIZED
SERIAL: 60050768018101BF2800000000000B5

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	Scsi Port2 Bus0/Disk1 Part0	OPEN	NORMAL	0	0
1	Scsi Port2 Bus0/Disk1 Part0	OPEN	NORMAL	1342	0
2	Scsi Port3 Bus0/Disk1 Part0	OPEN	NORMAL	0	0
3	Scsi Port3 Bus0/Disk1 Part0	OPEN	NORMAL	1444	0

DEV#: 1 DEVICE NAME: Disk2 Part0 TYPE: 2145 POLICY: OPTIMIZED
SERIAL: 60050768018101BF2800000000000B1

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	Scsi Port2 Bus0/Disk2 Part0	OPEN	NORMAL	1405	0
1	Scsi Port2 Bus0/Disk2 Part0	OPEN	NORMAL	0	0
2	Scsi Port3 Bus0/Disk2 Part0	OPEN	NORMAL	1387	0
3	Scsi Port3 Bus0/Disk2 Part0	OPEN	NORMAL	0	0

DEV#: 2 DEVICE NAME: Disk3 Part0 TYPE: 2145 POLICY: OPTIMIZED
SERIAL: 60050768018101BF2800000000000B2

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	Scsi Port2 Bus0/Disk3 Part0	OPEN	NORMAL	1398	0
1	Scsi Port2 Bus0/Disk3 Part0	OPEN	NORMAL	0	0
2	Scsi Port3 Bus0/Disk3 Part0	OPEN	NORMAL	1407	0
3	Scsi Port3 Bus0/Disk3 Part0	OPEN	NORMAL	0	0

DEV#: 3 DEVICE NAME: Disk4 Part0 TYPE: 2145 POLICY: OPTIMIZED
SERIAL: 60050768018101BF2800000000000B3

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	Scsi Port2 Bus0/Disk4 Part0	OPEN	NORMAL	1504	0
1	Scsi Port2 Bus0/Disk4 Part0	OPEN	NORMAL	0	0
2	Scsi Port3 Bus0/Disk4 Part0	OPEN	NORMAL	1281	0
3	Scsi Port3 Bus0/Disk4 Part0	OPEN	NORMAL	0	0

DEV#: 4 DEVICE NAME: Disk5 Part0 TYPE: 2145 POLICY: OPTIMIZED
SERIAL: 60050768018101BF2800000000000B4

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	Scsi Port2 Bus0/Disk5 Part0	OPEN	NORMAL	0	0
1	Scsi Port2 Bus0/Disk5 Part0	OPEN	NORMAL	1399	0
2	Scsi Port3 Bus0/Disk5 Part0	OPEN	NORMAL	0	0
3	Scsi Port3 Bus0/Disk5 Part0	OPEN	NORMAL	1391	0

DEV#: 5 DEVICE NAME: Disk6 Part0 TYPE: 2145 POLICY: OPTIMIZED
SERIAL: 60050768018101BF2800000000000A8

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	Scsi Port2 Bus0/Disk6 Part0	OPEN	NORMAL	1400	0

1	Scsi Port2 Bus0/Disk6 Part0	OPEN	NORMAL	0	0
2	Scsi Port3 Bus0/Disk6 Part0	OPEN	NORMAL	1390	0
3	Scsi Port3 Bus0/Disk6 Part0	OPEN	NORMAL	0	0

DEV#: 6 DEVICE NAME: Disk7 Part0 TYPE: 2145 POLICY: OPTIMIZED
SERIAL: 60050768018101BF2800000000000A9

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	Scsi Port2 Bus0/Disk7 Part0	OPEN	NORMAL	1379	0
1	Scsi Port2 Bus0/Disk7 Part0	OPEN	NORMAL	0	0
2	Scsi Port3 Bus0/Disk7 Part0	OPEN	NORMAL	1412	0
3	Scsi Port3 Bus0/Disk7 Part0	OPEN	NORMAL	0	0

DEV#: 7 DEVICE NAME: Disk8 Part0 TYPE: 2145 POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000AA

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	Scsi Port2 Bus0/Disk8 Part0	OPEN	NORMAL	0	0
1	Scsi Port2 Bus0/Disk8 Part0	OPEN	NORMAL	1417	0
2	Scsi Port3 Bus0/Disk8 Part0	OPEN	NORMAL	0	0
3	Scsi Port3 Bus0/Disk8 Part0	OPEN	NORMAL	1381	0

DEV#: 8 DEVICE NAME: Disk9 Part0 TYPE: 2145 POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000AB

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	Scsi Port2 Bus0/Disk9 Part0	OPEN	NORMAL	0	0
1	Scsi Port2 Bus0/Disk9 Part0	OPEN	NORMAL	1388	0
2	Scsi Port3 Bus0/Disk9 Part0	OPEN	NORMAL	0	0
3	Scsi Port3 Bus0/Disk9 Part0	OPEN	NORMAL	1413	0

DEV#: 9 DEVICE NAME: Disk10 Part0 TYPE: 2145 POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000A7

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	Scsi Port2 Bus0/Disk10 Part0	OPEN	NORMAL	1293	0
1	Scsi Port2 Bus0/Disk10 Part0	OPEN	NORMAL	0	0
2	Scsi Port3 Bus0/Disk10 Part0	OPEN	NORMAL	1477	0
3	Scsi Port3 Bus0/Disk10 Part0	OPEN	NORMAL	0	0

DEV#: 10 DEVICE NAME: Disk11 Part0 TYPE: 2145 POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000B9

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	Scsi Port2 Bus0/Disk11 Part0	OPEN	NORMAL	0	0
1	Scsi Port2 Bus0/Disk11 Part0	OPEN	NORMAL	59981	0
2	Scsi Port3 Bus0/Disk11 Part0	OPEN	NORMAL	0	0
3	Scsi Port3 Bus0/Disk11 Part0	OPEN	NORMAL	60179	0

DEV#: 11 DEVICE NAME: Disk12 Part0 TYPE: 2145 POLICY: OPTIMIZED
SERIAL: 60050768018101BF28000000000000BA

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
0	Scsi Port2 Bus0/Disk12 Part0	OPEN	NORMAL	28324	0
1	Scsi Port2 Bus0/Disk12 Part0	OPEN	NORMAL	0	0

2	Scsi	Port3	Bus0/Disk12	Part0	OPEN	NORMAL	27111	0
3	Scsi	Port3	Bus0/Disk12	Part0	OPEN	NORMAL	0	0

Sometimes, a host might discover everything correctly at the initial configuration, but it does not keep up with the dynamic changes in the configuration. Therefore, the SCSI ID is important.

6.1.9 Server adapter layout

If your host system has multiple internal I/O busses, place the two adapters that are used for IBM Spectrum Virtualize cluster access on two different I/O busses to maximize the availability and performance. When purchasing a server, always have two cards instead of one. For example, have two dual port HBA cards instead of one quad port HBA card because you can spread the I/O and keep the redundancy.

6.2 N-Port ID Virtualization

The usage model for all IBM Spectrum Virtualize products is based around two-way active/active node models. That is, a pair of distinct control modules that share active/active access for a volume. These nodes each have their own Fibre Channel WWNN, so all ports presented from each node have a set of WWPNS that are presented to the fabric.

Traditionally, if one node fails or is removed for some reason, the paths presented for volumes from that node go offline. It is up to the native OS multipathing software to fail over from using both sets of WWPN to just those that remain online. While this process is exactly what multipathing software is designed to do, sometimes it can be problematic, particularly if paths are not seen as coming back online for some reason.

N-Port ID Virtualization (NPIV) on IBM Spectrum Virtualize is a feature that was released in V7.7. The NPIV feature aims to provide an availability improvement for the hosts that are connected to the IBM Spectrum Virtualize nodes. It creates a virtual WWPN that is available only for host connection. During a node assert, failure, reboot, or service mode, the virtual WWPN from that node is transferred to the other node in the iogrp, to the same port.

That process ensures that, instead of having the host lose the connection to the IBM Spectrum Virtualize node, the connection remains active. The multipath software does not have to handle the path failures, mitigating in this case the occurrence of problems of hosts not recovering from path failure and an alerting storm from servers, for instance, in a code upgrade situation on IBM Spectrum Virtualize.

NPIV works in a symmetric way, which means the NPIV port from node 1 port 1 has the failover on node 2 port 1. For NPIV to work properly, you must have a symmetric cabling of IBM Spectrum Virtualize in your switch, such as odd ports on Fabric 1 and even Ports on Fabric 2, or vice versa. In short, you must have the ports that perform the failover in the same SAN Fabric.

NPIV is available only for the hosts. The back-end storage must still be zoned to the physical WWPN address. No intracluster or replication zone is allowed on the NPIV WWPN as well, because the NPIV ports are target only, as shown in Example 6-6.

Example 6-6 NPIV ports

```

ITS0_SAN_01:root> nodefind 50:05:07:68:0c:25:45:28
Local:
Type Pid   COS      PortName                               NodeName                               SCR
N   0a4c01;  2,3;50:05:07:68:0c:25:45:28;50:05:07:68:0c:00:45:28; 0x00000003
FC4s: FCP
Fabric Port Name: 20:4c:50:eb:1a:a9:8f:b8
Permanent Port Name: 50:05:07:68:0c:21:45:28
Device type: NPIV Target
Port Index: 76
Share Area: No
Device Shared in Other AD: No
Redirect: No
Partial: No
LSAN: No
Aliases: SVC_ITS0_DH8_LAB_N1_P3_HOST

```

NPIV is native on new deployments from V7.8 and above. You can disable the NPIV feature at installation, although generally we recommend that you do not do this, because it is a new deployment, so no extra effort needs to take place in order for your hosts to greatly benefit from this feature. If you have an existing IBM Spectrum Virtualize cluster, and want to use the NPIV feature, you must upgrade at least to V7.7.

When NPIV is enabled on IBM Spectrum Virtualize system nodes, each physical WWPN reports up to three virtual WWPNs, as shown in Table 6-1.

Table 6-1 IBM Spectrum Virtualize NPIV Ports

NPIV port	Port description
Primary NPIV Port	This is the WWPN that communicates with backend storage only.
Primary Host Attach Port	This is the WWPN that communicates with hosts. It is a target port only, and this is the primary port that represents this local nodes WWNN.
Failover Host Attach Port	This is a standby WWPN that communicates with hosts and is only brought online on this node if the partner node in this I/O Group goes offline. This is the same as the Primary Host Attach WWPN on the partner node.

Then, when NPIV effectively goes into action, you can see a situation such as that illustrated in Example 6-7.

Example 6-7 NPIV failover example

```

itso-sansw01:admin> portshow 4/12
portIndex: 60
portName: slot4 port12
portHealth: HEALTHY

Authentication: None

```



```
portDisableReason: None
portCFlags: 0x1
portFlags: 0x24b03 PRESENT ACTIVE F_PORT G_PORT U_PORT NPIV LOGICAL_ONLINE LOGIN
NOELP LED ACCEPT FLOGI
LocalSwcFlags: 0x0
portType: 24.0
portState: 1Online
Protocol: FC
portPhys: 6In_Sync portScn: 32F_Port
port generation number: 2164
state transition count: 18
```

```
portId: 0a3c00
portIfId: 4342080b
portWwn: 20:3c:50:eb:1a:a9:8f:b8
portWwn of device(s) connected:
    50:05:07:68:0c:15:45:28
    50:05:07:68:0c:11:45:28
Distance: normal
portSpeed: N16Gbps
```

```
itso-sansw01:admin> nodefind 50:05:07:68:0c:15:45:28
```

```
Local:
Type Pid COS PortName NodeName SCR
N 0a3c01; 2,3;50:05:07:68:0c:15:45:28;50:05:07:68:0c:00:45:28; 0x00000003
FC4s: FCP
Fabric Port Name: 20:3c:50:eb:1a:a9:8f:b8
Permanent Port Name: 50:05:07:68:0c:11:45:28
Device type: NPIV Target
Port Index: 60
Share Area: No
Device Shared in Other AD: No
Redirect: No
Partial: No
LSAN: No
Aliases: ITSO_SVCLAB01_NODE1_NP1
itso-sansw01:admin>
```

Then we took the node offline

```
itso-sansw01:admin> nodefind 50:05:07:68:0c:15:45:28
```

```
Local:
Type Pid COS PortName NodeName SCR
N 0a2502; 2,3;50:05:07:68:0c:15:45:28;50:05:07:68:0c:00:45:28; 0x00000003
FC4s: FCP
Fabric Port Name: 20:25:50:eb:1a:a9:8f:b8
Permanent Port Name: 50:05:07:68:0c:11:46:fc
Device type: NPIV Target
Port Index: 37
Share Area: No
Device Shared in Other AD: No
Redirect: No
Partial: No
LSAN: No
Aliases: ITSO_SVCLAB01_NODE1_NP1
```

```
itso-sansw01:admin>

itso-sansw01:admin> portshow 3/5
portIndex: 37
portName: slot3 port5
portHealth: HEALTHY

Authentication: None
portDisableReason: None
portCFlags: 0x1
portFlags: 0x24b03 PRESENT ACTIVE F_PORT G_PORT U_PORT NPIV LOGICAL_ONLINE LOGIN
NOELP LED ACCEPT FLOGI
LocalSwcFlags: 0x0
portType: 24.0
portState: 1Online
Protocol: FC
portPhys: 6In_Sync portScn: 32F_Port
port generation number: 2130
state transition count: 14

portId: 0a2500
portIfId: 4332001a
portWwn: 20:25:50:eb:1a:a9:8f:b8
portWwn of device(s) connected:
 50:05:07:68:0c:15:46:fc
 50:05:07:68:0c:11:46:fc
 50:05:07:68:0c:15:45:28
```

6.3 Host pathing

Each host mapping associates a volume with a host object and allows all HBA ports on the host object to access the volume. You can map a volume to multiple host objects.

When a mapping is created, multiple paths can exist across the SAN fabric from the hosts to the IBM Spectrum Virtualize nodes that present the volume. Most operating systems present each path to a volume as a separate storage device. Therefore IBM Spectrum Virtualize requires that multipathing software runs on the host. The multipathing software manages the many paths that are available to the volume and presents a single storage device to the operating system.

6.3.1 Multipathing Software

IBM Spectrum Virtualize requires the use of multipathing software on hosts that are connected. For the latest levels for each host operating system and multipathing software package, see *IBM System Storage Interoperation Center (SSIC)*:

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

6.3.2 Preferred path algorithm

I/O traffic for a particular volume is managed exclusively by the nodes in a single I/O group. The distributed cache in the SAN controller is two way. When a volume is created, a preferred node is chosen. This task is controllable at the time of volume creation. The owner node for a volume is the preferred node when both nodes are available.

When I/O is performed to a volume, the node that processes the I/O duplicates the data onto the partner node that is in the I/O group. A write from an IBM Spectrum Virtualize node to the back-end managed disk (MDisk) is only destaged by using the owner node (normally, the preferred node).

Therefore, when a new write or read comes in on the non-preferred node, it must send extra messages to the preferred node. The messages prompt the owner node to check whether it has the data in cache or if it is in the middle of destaging that data. Remember that performance is enhanced by accessing the volume through the preferred node.

IBM multipathing software (SDDPCM or SDDDSM) uses Asymmetric Logical Unit Access (ALUA) and checks the following preferred path settings during the initial configuration for each volume and manages path usage:

- ▶ Nonpreferred paths: Failover only
- ▶ Preferred path: Chosen multipath algorithm (default is load balance)

6.3.3 Path selection

IBM Spectrum Virtualize and Storwize family devices are storage subsystem ALUA compliant. That designation means that the host multipath driver must understand ALUA to achieve the best performance, resilience, and availability from IBM Spectrum Virtualize/Storwize. If the multipathing driver understands ALUA, it applies the load balance multipath policy only in the paths that belong to the preferred node. If the multipath driver does not understand ALUA, it spreads the data across all the paths, including the non-preferred node ones.

When a read or write I/O comes through a non-preferred node, IBM Spectrum Virtualize/Storwize sends the data by using the intracluster/node-to-node connection. This process enables the operation to be run by the preferred node. The specific behavior can take place when IBM Spectrum Virtualize Enhanced Stretched Cluster or HyperSwap is implemented. For more information, see the following IBM publications:

- ▶ *IBM Spectrum Virtualize and SAN Volume Controller Enhanced Stretched Cluster with VMware*, SG24-8211
- ▶ *IBM Storwize V7000, Spectrum Virtualize, HyperSwap, and VMware Implementation*, SG24-8317

Use caution when you are allocating volumes with the IBM Spectrum Virtualize console GUI to ensure adequate dispersion of the preferred node among the volumes. If the preferred node goes offline, all I/O goes through the nonpreferred node in write-through mode.

Table 6-2 shows the effect with 16 devices and read misses of the preferred node versus a nonpreferred node on performance. It also shows the significant effect on throughput.

Table 6-2 Sixteen device random 4 Kb read miss response time (lab nodes, in microseconds)

Preferred node (owner)	Nonpreferred node	Delta
18,227	21,256	3,029

Table 6-3 shows the change in throughput for 16 devices and a random 4 Kb read miss throughput by using the preferred node versus a nonpreferred node (as shown in Table 6-2 on page 255).

Table 6-3 Sixteen device random 4 Kb read miss throughput (input/output per second (IOPS))

Preferred node (owner)	Nonpreferred node	Delta
105,274.3	90,292.3	14,982

Table 6-4 shows the effect of the use of the nonpreferred paths versus the preferred paths on read performance.

Table 6-4 Random (1 TB) 4 Kb read response time (lab nodes, microseconds)

Preferred node (owner)	Nonpreferred node	Delta
5,074	5,147	73

Table 6-5 shows the effect of the use of nonpreferred nodes on write performance.

Table 6-5 Random (1 TB) 4 Kb write response time (lab nodes, microseconds)

Preferred node (owner)	Nonpreferred node	Delta
5,346	5,433	87

The IBM SDDDSM and SDDPCM software recognizes the preferred nodes and uses the preferred paths.

6.3.4 Path management

IBM Spectrum Virtualize design is based on multiple path access from the host to both IBM Spectrum Virtualize nodes. Multipathing software is expected to retry down multiple paths upon error detection.

Actively check the multipathing software display of paths that are available and currently in usage. Do this check periodically and immediately before any SAN maintenance or software upgrades. With IBM multipathing software (SDDPCM, native AIX MPIO and SDDDSM), this monitoring is done by using the `datapath query device` or `pcmpath query device` commands.

Fast node reset

IBM Spectrum Virtualize supports a major improvement in software error recovery. *Fast node reset* restarts a node after a software failure, but before the host fails I/O to applications. This node reset time has improved from several minutes to approximately 30 seconds for the standard node reset.

6.3.5 Non-disruptive volume migration between I/O groups

Attention: These migration tasks can be non-disruptive if they are performed correctly and hosts that are mapped to the volume support non-disruptive volume move. The cached data that is held within the system must first be written to disk before the allocation of the volume can be changed.

Modifying the I/O group that services the volume can be done concurrently with I/O operations if the host supports non-disruptive volume move. It also requires a rescan at the host level to ensure that the multipathing driver is notified that the allocation of the preferred node changed and the ports by which the volume is accessed changed. This process can be performed in the situation where one pair of nodes becomes over-used.

If there are any host mappings for the volume, the hosts must be members of the target I/O group or the migration fails. Make sure that you create paths to I/O groups on the host system. After the system successfully adds the new I/O group to the volume's access set and you move the selected volumes to another I/O group, detect the new paths to the volumes on the host.

The commands and actions on the host vary depending on the type of host and the connection method that is used. This process must be completed on all hosts to which the selected volumes are currently mapped.

You can also use the management GUI to move volumes between I/O groups non-disruptively. In the management GUI, click **Volumes** → **Volumes**. In the Volumes window, select the volume that you want to move and click **Actions** → **Move to Another I/O Group**. The wizard guides you through the steps for moving a volume to another I/O group, including any changes to hosts that are required. For more information, click **Need Help** in the associated management GUI windows.

In the following example, we move VDisk ndvm to another I/O group nondescriptively by using Red Hat Enterprise Linux 6.5 (Default Kernel). Example 6-8 shows the Red Hat Enterprise Linux 6.5 system before I/O group migration. For this example, the Storwize V7000/IBM Spectrum Virtualize caching I/O group is io_grp0.

Example 6-8 Native Linux multipath display before I/O group migration

```
[root@RHEL_65 ~]# multipath -ll
mpathb (360050768028100550000000000000fd) dm-2 IBM,2145
size=100G features='1 queue_if_no_path' hwhandler='0' wp=rw
|-+- policy='round-robin 0' prio=50 status=active
|  |- 0:0:0:0 sdb 8:16 active ready running
|  |- 1:0:0:0 sde 8:64 active ready running
|  |- 1:0:1:0 sdf 8:80 active ready running
|  `-- 0:0:7:0 sdi 8:128 active ready running
`-+- policy='round-robin 0' prio=10 status=enabled
|  |- 0:0:1:0 sdc 8:32 active ready running
|  |- 0:0:2:0 sdd 8:48 active ready running
|  |- 1:0:2:0 sdg 8:96 active ready running
|  `-- 1:0:3:0 sdh 8:112 active ready running
```

Complete the following steps:

1. Run the following commands to enable VDisk ndvm access for both I/O groups, io_grp0 and io_grp1:

```
svctask addvdiskaccess -iogrp io_grp1 ndvm
svctask movevdisk -iogrp io_grp1 ndvm
```
2. Detect the new paths to the volume in the destination I/O group, as shown in Example 6-9.

Example 6-9 SCSI rescan command on Red Hat Enterprise Linux 6.5

```
[root@RHEL_65 ~]# scsi-rescan -r
Host adapter 0 (qla2xxx) found.
Host adapter 1 (qla2xxx) found.
```

Scanning SCSI subsystem for new devices

.....

0 new device(s) found.

1 device(s) removed.

3. Validate that the new paths are detected by Red Hat Enterprise Linux 6.5, as shown in Example 6-10.

Example 6-10 Native Linux multipath display access to both I/O groups

```
[root@RHEL_65 ~]# multipath -ll
mpathb (360050768028100550000000000000fd) dm-2 IBM,2145
size=100G features='1 queue_if_no_path' hwhandler='0' wp=rw
|+- policy='round-robin 0' prio=50 status=active
|  |- 0:0:5:0 sd1 8:176 active ready running
|  |- 0:0:6:0 sdm 8:192 active ready running
|  |- 1:0:7:0 sdq 65:0 active ready running
|  `-- 1:0:6:0 sdp 8:240 active ready running
`+- policy='round-robin 0' prio=10 status=enabled
   |- 0:0:7:0 sdi 8:128 active ready running
   |- 1:0:1:0 sdf 8:80 active ready running
   |- 1:0:0:0 sde 8:64 active ready running
   |- 0:0:0:0 sdb 8:16 active ready running
   |- 0:0:1:0 sdc 8:32 active ready running
   |- 0:0:2:0 sdd 8:48 active ready running
   |- 1:0:2:0 sdg 8:96 active ready running
   |- 1:0:3:0 sdh 8:112 active ready running
   |- 0:0:3:0 sdj 8:144 active ready running
   |- 0:0:4:0 sdk 8:160 active ready running
   |- 1:0:4:0 sdn 8:208 active ready running
   `-- 1:0:5:0 sdo 8:224 active ready running
```

4. After you validate that the new paths are detected, you can safely remove access from the old I/O group by running the following command:

```
svctask rmdiskaccess -iogrp io_grp0 ndvm
```

5. Remove the path of the old I/O group by using the `scsi-rescan -r` command.
6. Validate that the old path was successfully removed, as shown in Example 6-11.

Example 6-11 Native Linux multipath display access to new I/O group

```
[root@RHEL_65 ~]# multipath -ll
mpathb (360050768028100550000000000000fd) dm-2 IBM,2145
size=100G features='1 queue_if_no_path' hwhandler='0' wp=rw
|+- policy='round-robin 0' prio=50 status=active
|  |- 0:0:5:0 sd1 8:176 active ready running
|  |- 0:0:6:0 sdm 8:192 active ready running
|  |- 1:0:7:0 sdq 65:0 active ready running
|  `-- 1:0:6:0 sdp 8:240 active ready running
`+- policy='round-robin 0' prio=10 status=enabled
   |- 0:0:3:0 sdj 8:144 active ready running
   |- 0:0:4:0 sdk 8:160 active ready running
   |- 1:0:4:0 sdn 8:208 active ready running
   `-- 1:0:5:0 sdo 8:224 active ready running
```

Additional information regarding volume migrations between I/O groups can be found on Knowledge Center:

<https://ibm.biz/BdjJQx>

6.4 I/O queues

Host operating system and host bus adapter software must have a way to fairly prioritize I/O to the storage. The host bus might run faster than the I/O bus or external storage. Therefore, you must have a way to queue I/O to the devices. Each operating system and host adapter have unique methods to control the I/O queue. The unique method to control I/O queue can be host adapter-based or memory and thread resources-based, or based on the number of commands that are outstanding for a device.

You have several configuration parameters available to control the I/O queue for your configuration. The storage adapters (volumes on IBM Spectrum Virtualize) have host adapter parameters and queue depth parameters. Algorithms are also available within multipathing software, such as the `qdepth_enable` attribute.

6.4.1 Queue depths

Queue depth is used to control the number of concurrent operations that occur on different storage resources. Queue depth is the number of I/O operations that can be run in parallel on a device.

Guidance about limiting queue depths in large SANs as described in previous IBM documentation was replaced with a calculation for homogeneous and nonhomogeneous FC hosts. This calculation is for an overall queue depth per I/O group. You can use this number to reduce queue depths that are lower than the recommendations or defaults for individual host adapters.

For more information, see Chapter 3, “Storage pools and managed disks” on page 75 and *Queue depth in Fibre Channel hosts* topic in the V8.1 Documentation sites:

- ▶ IBM Spectrum Virtualize
<https://ibm.biz/BdjL6h>
- ▶ Storwize V7000
<https://ibm.biz/BdjL6g>

6.5 Host clustering and reserves

To prevent hosts from sharing storage inadvertently, establish a storage reservation mechanism. The mechanisms for restricting access to IBM Spectrum Virtualize volumes use the SCSI-3 persistent reserve commands or the SCSI-2 reserve and release commands.

The host software uses several methods to implement host clusters. These methods require sharing the volumes on IBM Spectrum Virtualize between hosts. To share storage between hosts, maintain control over accessing the volumes. Some clustering software use software locking methods.

You can choose other methods of control by the clustering software or by the device drivers to use the SCSI architecture reserve or release mechanisms. The multipathing software can change the type of reserve that is used from an earlier reserve to persistent reserve, or remove the reserve.

Persistent reserve refers to a set of SCSI-3 standard commands and command options that provide SCSI initiators with the ability to establish, preempt, query, and reset a reservation policy with a specified target device. The functions that are provided by the persistent reserve commands are a superset of the original reserve or release commands.

The persistent reserve commands are incompatible with the earlier reserve or release mechanism. Also, target devices can support only reservations from the earlier mechanism or the new mechanism. Attempting to mix persistent reserve commands with earlier reserve or release commands results in the target device returning a reservation conflict error.

Earlier reserve and release mechanisms (SCSI-2) reserved the entire LUN (volume) for exclusive use down a single path. This approach prevents access from any other host or even access from the same host that uses a different host adapter. The persistent reserve design establishes a method and interface through a reserve policy attribute for SCSI disks. This design specifies the type of reservation (if any) that the operating system device driver establishes before it accesses data on the disk.

The following possible values are supported for the reserve policy:

- ▶ `No_reserve`: No reservations are used on the disk.
- ▶ `Single_path`: Earlier reserve or release commands are used on the disk.
- ▶ `PR_exclusive`: Persistent reservation is used to establish *exclusive host access* to the disk.
- ▶ `PR_shared`: Persistent reservation is used to establish *shared host access* to the disk.

When a device is opened (for example, when the AIX `varyonvg` command opens the underlying hdisks), the device driver checks the object data manager (ODM) for a `reserve_policy` and a `PR_key_value`. The driver then opens the device. For persistent reserve, each host that is attached to the shared disk must use a unique registration key value.

6.5.1 Clearing reserves

It is possible to accidentally leave a reserve on the IBM Spectrum Virtualize volume or on the IBM Spectrum Virtualize MDisk during migration into IBM Spectrum Virtualize or when disks are reused for another purpose. Several tools are available from the hosts to clear these reserves. The easiest tools to use are the `pcmquerypr` (AIX SDDPCM host) commands. Another tool is a menu-driven Windows SDDDSM tool.

The Windows Persistent Reserve Tool is called `PRTool.exe` and is installed automatically when SDDDSM is installed in the `C:\Program Files\IBM\Subsystem Device Driver\PRTool.exe` directory. You can clear the IBM Spectrum Virtualize volume reserves by removing all the host mappings.

Example 6-12 shows a failing `pcmquerypr` command to clear the reserve and the error.

Example 6-12 Output of the pcmquerypr command

```
# pcmquerypr -ph /dev/hdisk232 -V
connection type: fscsi0
open dev: /dev/hdisk232
couldn't open /dev/hdisk232, errno=16
```

Use the AIX `errno.h` include file to determine what error number 16 indicates. This error indicates a busy condition, which can indicate a legacy reserve or a persistent reserve from another host (or that this host is from a different adapter). However, some AIX technology levels have a diagnostic open issue that prevents the `pcmquerypr` command from opening the device to display the status or to clear a reserve.

For more information about older AIX technology levels that break the `pcmquerypr` command, see *IBM Multipath Subsystem Device Driver Path Control Module (PCM) Version 2.7.0.0 FOR AIX*, which is available at this website:

<http://www.ibm.com/support/docview.wss?uid=ssg1S4001363#SVC>

6.5.2 IBM Spectrum Virtualize MDisk reserves

There are instances in which a host image mode migration appears to succeed, but problems occur when the volume is opened for read or write I/O. The problems can result from not removing the reserve on the MDisk before image mode migration is used in IBM Spectrum Virtualize. You cannot clear a leftover reserve on an IBM Spectrum Virtualize MDisk from IBM Spectrum Virtualize. You must clear the reserve by mapping the MDisk back to the owning host and clearing it through host commands, or through back-end storage commands as advised by IBM technical support.

6.6 AIX hosts

This section describes various topics that are specific to AIX.

6.6.1 HBA parameters for performance tuning

You can use the example settings in this section to start your configuration in the specific workload environment. These settings are a guideline, and are not guaranteed to be the answer to all configurations. Always try to set up a test of your data with your configuration to see whether further tuning can help. For best results, it helps to have knowledge about your specific data I/O pattern.

The settings in the following sections can affect performance on an AIX host. These sections examine these settings in relation to how they affect the two workload types.

Transaction-based settings

The host attachment script sets the default values of attributes for IBM Spectrum Virtualize hdisks: `devices.fcp.disk.IBM.rte` or `devices.fcp.disk.IBM.mpio.rte`. You can modify these values as a starting point. In addition, you can use several HBA parameters to set higher performance or large numbers of hdisk configurations.

You can change all attribute values that are changeable by using the `chdev` command for AIX.

AIX settings that can directly affect transaction performance are the `queue_depth` hdisk attribute and `num_cmd_elem` attribute in the HBA attributes.

The queue_depth hdisk attribute

For the logical drive (which is known as the hdisk in AIX), the setting is the attribute `queue_depth`, as shown in the following example:

```
# chdev -l hdiskX -a queue_depth=Y -P
```

In this example, *X* is the hdisk number, and *Y* is the value to which you are setting *X* for `queue_depth`.

For a high-volume transaction workload of small random transfers, try a `queue_depth` value of 25 or more. For large sequential workloads, performance is better with shallow queue depths, such as a value of 4.

The num_cmd_elem attribute

For the HBA settings, the `num_cmd_elem` attribute for the fcs device represents the number of commands that can be queued to the adapter, as shown in the following example:

```
chdev -l fcsX -a num_cmd_elem=1024 -P
```

The default value is 200, but the following maximum values can be used:

- ▶ LP9000 adapters: 2048
- ▶ LP10000 adapters: 2048
- ▶ LP11000 adapters: 2048
- ▶ LP7000 adapters: 1024

Tip: For a high volume of transactions on AIX or many hdisks on the fcs adapter, increase `num_cmd_elem` to 1,024 for the fcs devices that are being used.

The AIX settings that can directly affect throughput performance with large I/O block size are the `lg_term_dma` and `max_xfer_size` parameters for the fcs device.

Throughput-based settings

In the throughput-based environment, you might want to decrease the queue-depth setting to a smaller value than the default from the host attach. In a mixed application environment, you do not want to lower the `num_cmd_elem` setting because other logical drives might need this higher value to perform. In a purely high throughput workload, this value has no effect.

Start values: For high throughput sequential I/O environments, use the start values `lg_term_dma = 0x400000` or `0x800000` (depending on the adapter type) and `max_xfer_size = 0x200000`.

First, test your host with the default settings. Then, make these possible tuning changes to the host parameters to verify whether these suggested changes enhance performance for your specific host configuration and workload.

The lg_term_dma attribute

The `lg_term_dma` AIX Fibre Channel adapter attribute controls the direct memory access (DMA) memory resource that an adapter driver can use. The default value of `lg_term_dma` is `0x200000`, and the maximum value is `0x8000000`.

One change is to increase the value of `lg_term_dma` to `0x400000`. If you still experience poor I/O performance after changing the value to `0x400000`, you can increase the value of this attribute again. If you have a dual-port Fibre Channel adapter, the maximum value of the `lg_term_dma` attribute is divided between the two adapter ports. Therefore, never increase the value of the `lg_term_dma` attribute to the maximum value for a dual-port Fibre Channel adapter because this value causes the configuration of the second adapter port to fail.

The max_xfer_size attribute

The `max_xfer_size` AIX Fibre Channel adapter attribute controls the maximum transfer size of the Fibre Channel adapter. Its default value is 100,000, and the maximum value is 1,000,000. You can increase this attribute to improve performance. You can change this attribute only with AIX V5.2 or later.

Setting the `max_xfer_size` attribute affects the size of the memory area that is used for data transfer by the adapter. With the default value of `max_xfer_size=0x100000`, the area is 16 MB, and for other allowable values of the `max_xfer_size` attribute, the memory area is 128 MB.

6.6.2 Configuring for fast fail and dynamic tracking

For host systems that run an AIX V5.2 or later operating system, you can achieve the best results by using the fast fail and dynamic tracking attributes. Before you configure your host system to use these attributes, ensure that the host is running the AIX operating system V5.2 or later.

To configure your host system to use the fast fail and dynamic tracking attributes, complete the following steps:

1. Set the Fibre Channel SCSI I/O Controller Protocol Device event error recovery policy to `fast_fail` for each Fibre Channel adapter, as shown in the following example:

```
chdev -l fscsi0 -a fc_err_recov=fast_fail
```

This command is for the `fscsi0` adapter.

2. Enable dynamic tracking for each Fibre Channel device, as shown in the following example:

```
chdev -l fscsi0 -a dyntrk=yes
```

This command is for the `fscsi0` adapter.

6.6.3 SDDPCM

As Fibre Channel technologies matured, AIX was enhanced by adding native multipathing support called *multipath I/O* (MPIO). By using the MPIO structure, a storage manufacturer can create software plug-ins for their specific storage. The IBM Spectrum Virtualize version of this plug-in is called SDDPCM, which requires a host attachment script called `devices.fcp.disk.ibm.mpio.rte`. For more information about SDDPCM, see *Host Attachment for SDDPCM on AIX*, S4001363, which is available at this website:

<http://www.ibm.com/support/docview.wss?uid=ssg1S4001363#SVC>

SDDPCM and AIX MPIO have been continually improved since their release. You must be at the latest release levels of this software. You will not see the preferred path indicator for SDDPCM until after the device is opened for the first time.

SDDPCM features the following types of reserve policies:

- ▶ No_reserve policy
- ▶ Exclusive host access single path policy
- ▶ Persistent reserve exclusive host policy
- ▶ Persistent reserve shared host access policy

Usage of the persistent reserve now depends on the `hdisk` attribute, `reserve_policy`. Change this policy to match your storage security requirements.

The following path selection algorithms are available:

- ▶ Failover
- ▶ Round-robin
- ▶ Load balancing
- ▶ Load balancing port

SDDPCM code 2.1.3.0 and later features improvements in failed path reclamation by a health checker, a failback error recovery algorithm, FC dynamic device tracking, and support for a SAN boot device on MPIO-supported storage devices.

More information can be found in the Multipath Subsystem Device Driver User's Guide at:

<https://ibm.biz/BdjL6V>

6.7 Virtual I/O Server

Virtual SCSI is based on a client/server relationship. The VIOS owns the physical resources and acts as the server or target device. Physical adapters with attached disks (in this case, volumes on IBM Spectrum Virtualize) on the VIOS partition can be shared by one or more partitions. These partitions contain a virtual SCSI client adapter that detects these virtual devices as standard SCSI-compliant devices and LUNs.

You can create the following types of volumes on a VIOS:

- ▶ Physical volume (PV) VSCSI hdisks
- ▶ Logical volume (LV) VSCSI hdisks

PV VSCSI hdisks are entire LUNs from the VIOS perspective. If you are concerned about failure of a VIOS and have configured redundant VIOSs for that reason, you must use PV VSCSI hdisks. Therefore, PV VSCSI hdisks are entire LUNs that are volumes from the virtual I/O client perspective. An LV VSCSI hdisk cannot be served up from multiple VIOSs. LV VSCSI hdisks are in LVM volume groups on the VIOS, and cannot span PVs in that volume group or be striped LVs. Because of these restrictions, use PV VSCSI hdisks.

Multipath support for IBM Spectrum Virtualize attachment to Virtual I/O Server is provided by MPIO with SDDPCM. Where Virtual I/O Server SAN Boot or dual Virtual I/O Server configurations are required, only MPIO with SDDPCM is supported. Because of this restriction in the latest IBM Spectrum Virtualize-supported levels, use MPIO with SDDPCM. For more information, see *IBM System Storage Interoperation Center (SSIC)*:

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

For more information about VIOS, see this website:

<http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/faq.html>

Common questions include how to migrate data into a virtual I/O environment, or how to reconfigure storage on a VIOS. This question is addressed at the previous web address.

Many clients want to know whether you can move SCSI LUNs between the physical and virtual environment “as is.” On a physical SCSI device (LUN) with user data on it that is in a SAN environment, can this device be allocated to a VIOS and then provisioned to a client partition and used by the client “as is”?

The answer is *no*. This function is not supported as of this writing. Virtual SCSI devices are new devices when they are created. The data must be put on them after creation, which often requires a type of backup of the data in the physical SAN environment with a restoration of the data onto the volume.

6.7.1 Methods to identify a disk for use as a virtual SCSI disk

The VIOS uses the following methods to uniquely identify a disk for use as a virtual SCSI disk:

- ▶ Unique device identifier (UDID)
- ▶ IEEE volume identifier
- ▶ Physical volume identifier (PVID)

Each of these methods can result in different data formats on the disk. The preferred disk identification method for volumes is the use of UDIDs.

6.7.2 UDID method for MPIO

Most multipathing software products for non-MPIO disk storage use the PVID method instead of the UDID method. Because of the different data formats that are associated with the PVID method, in non-MPIO environments, certain future actions that are performed in the VIOS logical partition (LPAR) can require data migration. That is, it might require a type of backup and restoration of the attached disks, including the following tasks:

- ▶ Conversion from a non-MPIO environment to an MPIO environment
- ▶ Conversion from the PVID to the UDID method of disk identification
- ▶ Removal and rediscovery of the disk storage ODM entries
- ▶ Updating non-MPIO multipathing software under certain circumstances
- ▶ Possible future enhancements to virtual I/O

Due in part to the differences in disk format, virtual I/O is supported only for new disk installations. AIX, virtual I/O, and SDDPCM development are working on changes to make this migration easier in the future. One enhancement is to use the UDID or IEEE method of disk identification. If you use the UDID method, you can contact IBM technical support to find a migration method that might not require restoration. A quick and simple method to determine whether a backup and restoration is necessary is to read the PVID off the disk by running the following command:

```
lquerypv -h /dev/hdisk## 80 10
```

If the output is different on the VIOS and virtual I/O client, you must use backup and restore.

6.8 Windows hosts

To release new enhancements more quickly, the newer hardware architectures are tested only on the SDDDSM code stream. Therefore, only SDDDSM packages are available.

For Microsoft Windows 2012 and Microsoft Windows 2008R2, download the latest version of SDDDSM from this website:

<http://www.ibm.com/support/docview.wss?uid=ssg1S4000350#SVC>

6.8.1 Clustering and reserves

Windows SDDDSM uses the persistent reserve functions to implement Windows clustering. A stand-alone Windows host does not use reserves.

When SDDDSM is installed, the reserve and release functions are converted into the appropriate persistent reserve and release equivalents to allow load balancing and multipathing from each host.

6.8.2 Tunable parameters

With Windows operating systems, the queue-depth settings are the responsibility of the host adapters. They are configured through the BIOS setting. Configuring the queue-depth settings varies from vendor to vendor. For more information about configuring your specific cards, see “Hosts running the Microsoft Windows Server operating system” in IBM Spectrum Virtualize IBM Knowledge Center:

<https://ibm.biz/BdsDBk>

Queue depth is also controlled by the Windows application program. The application program controls the number of I/O commands that it allows to be outstanding before waiting for completion. You might have to adjust the queue depth that is based on the overall I/O group queue depth calculation, as described in Chapter 4, “Volumes” on page 109.

6.8.3 Guidelines for disk alignment using Microsoft Windows with IBM Spectrum Virtualize volumes

You can set preferred settings for best performance with IBM Spectrum Virtualize when you use Microsoft Windows before 2008 operating systems and applications with a significant amount of I/O.

For more information, see “Performance Recommendations for Disk Alignment using Microsoft Windows” at this website:

http://www.ibm.com/support/docview.wss?rs=591&context=STPVGU&context=STPVFV&q1=microsoft&uid=ssg1S1003291&loc=en_US&cs=utf-8&lang=en

If you are using Microsoft Windows 2008 or later, there is no need for Disk Alignment.

6.9 Linux hosts

IBM Spectrum Virtualize multipathing supports Linux native DM-MPIO multipathing. Veritas DMP is also available for certain kernels.

For more information about which versions of each Linux kernel require DM-MPIO support, see *IBM System Storage Interoperation Center (SSIC)* available:

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

Certain types of clustering are now supported. However, the multipathing software choice is tied to the type of cluster and HBA driver. For example, Veritas Storage Foundation is supported for certain hardware and kernel combinations, but it also requires Veritas DMP multipathing. Contact IBM marketing for SCORE/RPQ support if you need Linux clustering in your specific environment and it is not listed.

New Linux operating systems support native DM-MPIO. An example configuration of `multipath.conf` is available:

<https://ibm.biz/BdjL6A>

For further reference on RHEL 6 and 7 operating systems, you can check the following sites:

- ▶ Red Hat Enterprise Linux 6 DM Multipath Configuration and Administration
<https://ibm.biz/BdjL6u>
- ▶ Red Hat Enterprise Linux 7 DM Multipath Configuration and Administration
<https://ibm.biz/BdjL6L>

6.9.1 Tunable parameters

Linux performance is influenced by HBA parameter settings and queue depth. The overall calculation for queue depth for the I/O group is described in Chapter 3, “Storage pools and managed disks” on page 75. In addition, the IBM Spectrum Virtualize IBM Knowledge Center provides maximums per HBA adapter or type. For more information, see these websites:

- ▶ IBM Spectrum Virtualize
<https://ibm.biz/BdjL6h>
- ▶ Storwize V7000
<https://ibm.biz/BdjL6g>

For more information about the settings for each specific HBA type and general Linux OS tunable parameters, see the *Attaching to a host running the Linux operating system* topic in the IBM Spectrum Virtualize IBM Knowledge Center:

- ▶ IBM Spectrum Virtualize
<https://ibm.biz/BdjJgN>
- ▶ Storwize V7000
<https://ibm.biz/BdjJhc>

In addition to the I/O and operating system parameters, Linux has tunable file system parameters.

You can use the `tune2fs` command to increase file system performance that is based on your specific configuration. You can change the journal mode and size, and index the directories. For more information, see “Learn Linux, 101: Maintain the integrity of filesystems” in IBM developerWorks® at this website:

<https://www.ibm.com/developerworks/library/l-lpic1-104-2/>

6.10 Solaris hosts

Two options are available for multipathing support on Solaris hosts: Symantec Veritas Volume Manager and Solaris MPxIO. The option that you choose depends on your file system requirements and the operating system levels in the latest interoperability matrix. For more information, see *IBM System Storage Interoperation Center (SSIC)*:

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

IBM SDD is no longer supported because its features are now available natively in the multipathing driver for Solaris MPxIO. If SDD support is still needed, contact your IBM marketing representative to request an RPQ for your specific configuration.

From Solaris 10 and later, Oracle has released a combined file system and logical volume manager called ZFS, designed by Sun Microsystems. It uses MPxIO and is inbound to the Solaris 11, being a native option to Veritas Volume Manager. For more information about Oracle ZFS, see the following links:

<http://www.oracle.com/technetwork/systems/hands-on-labs/s11-intro-zfs-1408637.html>
<https://docs.oracle.com/cd/E19253-01/819-5461/819-5461.pdf>
https://docs.oracle.com/cd/E23824_01/pdf/821-1448.pdf

6.10.1 Solaris MPxIO

SAN boot and clustering support is available for V5.9, V5.10, and 5.11, depending on the multipathing driver and HBA choices. Support for load balancing of the MPxIO software was included with IBM Spectrum Virtualize V4.3. If you want to run MPxIO on your Sun SPARC host, configure your IBM Spectrum Virtualize host object with the type attribute set to `tpgs`, as shown in the following example:

```
svctask mkhost -name new_name_arg -hbawpn wwpn_list -type tpgs
```

In this command, `-type` specifies the type of host. Valid entries are `hpux`, `tpgs`, `generic`, `openvms`, `adminlun`, and `hide_secondary`. The `tpgs` option enables an extra target port unit. The default is `generic`.

6.10.2 Symantec Veritas Volume Manager

When you are managing IBM Spectrum Virtualize storage in Symantec volume manager products, you must install an ASL on the host so that the volume manager is aware of the storage subsystem properties (active/active or active/passive). If the appropriate Array Support Library (ASL) is not installed, the volume manager did not claim the LUNs. Usage of the ASL is required to enable the special failover or failback multipathing that IBM Spectrum Virtualize requires for error recovery. Use the commands that are shown in Example 6-13 to determine the basic configuration of a Symantec Veritas server.

Example 6-13 Determining the Symantec Veritas server configuration

```
pkginfo -l (lists all installed packages)
showrev -p |grep vxvm (to obtain version of volume manager)
vxddladm listsupport (to see which ASLs are configured)
vxdisk list
vxdmpadm listctrl all (shows all attached subsystems, and provides a type where
possible)
vxdmpadm getsubpaths ctrl=cX (lists paths by controller)
vxdmpadm getsubpaths dmpnodename=cctxdxs2' (lists paths by LUN)
```

The commands that are shown in Example 6-14 and Example 6-15 determine whether the IBM Spectrum Virtualize is properly connected. They show at a glance which ASL is used (native DMP ASL or SDD ASL).

Example 6-14 on page 269 shows what you see when Symantec Volume Manager correctly accesses IBM Spectrum Virtualize by using the SDD pass-through mode ASL.

Example 6-14 Symantec Volume Manager using SDD pass-through mode ASL

```
# vxdmpadm list enclosure all
ENCLR_NAME ENCLR_TYPE ENCLR_SNO STATUS
=====
OTHER_DISKS OTHER_DISKS OTHER_DISKS CONNECTED
VPATH_SANVCO VPATH_SANVC 0200628002faXX00 CONNECTED
```

Example 6-15 shows what you see when IBM Spectrum Virtualize is configured by using native DMP ASL.

Example 6-15 IBM Spectrum Virtualize that is configured by using native ASL

```
# vxdmpadm listenclosure all
ENCLR_NAME ENCLR_TYPE ENCLR_SNO STATUS
=====
OTHER_DISKS OTHER_DSKSI OTHER_DISKS CONNECTED
SAN_VCO SAN_VC 0200628002faXX00 CONNECTED
```

6.10.3 DMP multipathing

For the latest ASL levels to use native DMP, see the array-specific module table at this website:

<https://sort.symantec.com/asl>

For the latest Veritas Patch levels, see the patch table at this website:

<https://sort.symantec.com/patch/matrix>

To check the installed Symantec Veritas version, enter the following command:

```
showrev -p |grep vxvm
```

To check which IBM ASLs are configured into the Volume Manager, enter the following command:

```
vxddladm listsupport |grep -i ibm
```

After you install a new ASL by using the **pkgadd** command, restart your system or run the **vxdtc1 enable** command. To list the ASLs that are active, enter the following command:

```
vxddladm listsupport
```

6.10.4 Troubleshooting configuration issues

Example 6-16 shows that the appropriate ASL is not installed or the system is enabling the ASL. The key is the enclosure type OTHER_DISKS.

Example 6-16 Troubleshooting ASL errors

```
vxdmpadm listctlr all
CTLR-NAME          ENCLR-TYPE          STATE          ENCLR-NAME
=====
c0                  OTHER_DISKS         ENABLED        OTHER_DISKS
c2                  OTHER_DISKS         ENABLED        OTHER_DISKS
c3                  OTHER_DISKS         ENABLED        OTHER_DISKS

vxdmpadm listenclosure all
```

ENCLR_NAME	ENCLR_TYPE	ENCLR_SNO	STATUS
OTHER_DISKS	OTHER_DISKS	OTHER_DISKS	CONNECTED
Disk	Disk	DISKS	DISCONNECTED

6.11 VMware server

To determine the various VMware ESX levels that are supported, see the IBM System Storage Interoperation Center (SSIC) available at:

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

On this website you can also find information about the available support in Spectrum Virtualize V8.1 for VMware vStorage APIs for Array Integration (VAAI).

IBM Spectrum Virtualize V7.2 and later supports VMware vStorage APIs. IBM Spectrum Virtualize implemented new storage-related tasks that were previously performed by VMware, which helps improve efficiency and frees server resources for more mission-critical tasks. The new functions include full copy, block zeroing, and hardware-assisted locking.

If you are not using the new API functions, the minimum supported VMware level is V3.5. If earlier versions are required, contact your IBM marketing representative and ask about the submission of an RPQ for support. The required patches and procedures are supplied after the specific configuration is reviewed and approved.

For more information about host attachment recommendations, see the *Attachment requirements for hosts running VMware operating systems* topic in the IBM Spectrum Virtualize Version V8.1 at:

- ▶ IBM Spectrum Virtualize
<https://ibm.biz/BdjJ9a>
- ▶ Storwize V7000
<https://ibm.biz/BdjJ9G>

6.11.1 Multipathing solutions supported

Multipathing is supported at VMware ESX level 2.5.x and later. Therefore, installing multipathing software is not required. The following multipathing algorithms are available on Native Multipathing (NMP):

- ▶ Fixed-path
- ▶ Round-robin
- ▶ Most recently used (MRU)

VMware multipathing was improved to use the IBM Spectrum Virtualize preferred node algorithms starting with V4.0. Preferred paths are ignored in VMware versions before V4.0. VMware multipathing software performs static load balancing for I/O, which defines the fixed path for a volume.

The round-robin algorithm rotates path selection for a volume through all paths. For any volume that uses the fixed-path policy, the first discovered preferred node path is chosen. The VMW_PSP_MRU policy selects the first working path, discovered at system boot time. If this

path becomes unavailable, the ESXi/ESX host switches to an alternative path and continues to use the new path while it is available.

All these algorithms were modified with V4.0 and later to honor the IBM Spectrum Virtualize preferred node that is discovered by using the **TPGS** command. Path failover is automatic in all cases. If the round-robin algorithm is used, path fallback might not return to a preferred node path. Therefore, manually check pathing after any maintenance or problems occur.

Update: From vSphere version 5.5 and later, VMware multipath driver fully supports IBM Spectrum Virtualize/Storwize V7000 ALUA preferred path algorithms. VMware administrators should select Round Robin and validate that `VMW_SATP_ALUA` is displayed. This configuration reduces operational burden and improves cache hit rate by sending the I/O to the preferred node.

6.11.2 Multipathing configuration maximums

The VMware multipathing software supports the following maximum configuration:

- ▶ A total of 256 SCSI devices
- ▶ Up to 32 paths to each volume
- ▶ Up to 1024 paths per server

Tip: Each path to a volume equates to a single SCSI device.

Refer to the following VMware document for a complete list of maximums:

<https://www.vmware.com/pdf/vsphere5/r55/vsphere-55-configuration-maximums.pdf>

<https://www.vmware.com/pdf/vsphere6/r60/vsphere-60-configuration-maximums.pdf>

<https://www.vmware.com/pdf/vsphere6/r65/vsphere-65-configuration-maximums.pdf>

6.12 Monitoring

A consistent set of monitoring tools is available when IBM SDDDSM and SDDPCM are used for the multipathing software on the various operating system environments. You can use the **datapath query device** and **datapath query adapter** commands for path monitoring. You can also monitor path performance by using either of the following **datapath** commands:

```
datapath query devstats
```

```
pcmpath query devstats
```

The **datapath query devstats** command shows performance information for a single device, all devices, or a range of devices. Example 6-17 shows the output of the **datapath query devstats** command for two devices.

Example 6-17 Output of the datapath query devstats command

```
C:\Program Files\IBM\Subsystem Device Driver>datapath query devstats
Total Devices : 2
```

```
Device #: 0
```

```
=====
```

	Total Read	Total Write	Active Read	Active Write	Maximum
I/O:	1755189	1749581	0	0	3
SECTOR:	14168026	153842715	0	0	256

Transfer Size:	<= 512	<= 4k	<= 16K	<= 64K	> 64K
	271	2337858	104	1166537	0
Device #: 1					
=====					
	Total Read	Total Write	Active Read	Active Write	Maximum
I/O:	20353800	9883944	0	1	4
SECTOR:	162956588	451987840	0	128	256
Transfer Size:	<= 512	<= 4k	<= 16K	<= 64K	> 64K
	296	27128331	215	3108902	0

Also, the **datapath query adaptstats** adapter-level statistics command is available (mapped to the **pcmpath query adaptstats** command). Example 6-18 shows the use of two adapters.

Example 6-18 Output of the datapath query adaptstats command

```
C:\Program Files\IBM\Subsystem Device Driver>datapath query adaptstats
```

Adapter #: 0					
=====					
	Total Read	Total Write	Active Read	Active Write	Maximum
I/O:	11060574	5936795	0	0	2
SECTOR:	88611927	317987806	0	0	256
Adapter #: 1					
=====					
	Total Read	Total Write	Active Read	Active Write	Maximum
I/O:	11048415	5930291	0	1	2
SECTOR:	88512687	317726325	0	128	256

You can clear these counters so that you can script the usage to cover a precise amount of time. By using these commands, you can choose devices to return as a range, single device, or all devices. To clear the counts, use the following command:

```
datapath clear device count
```

6.12.1 Load measurement and stress tools

Load measurement tools often are specific to each host operating system. For example, the AIX operating system has the **iostat** tool, and Windows has the **perfmon.msc /s** tool.

Industry standard performance benchmarking tools are available by joining the Storage Performance Council.

These tools are available to create stress and measure the stress that was created with a standardized tool. Use these tools to generate stress for your test environments to compare them with the industry measurements.

Iometer is another stress tool that you can use for Windows and Linux hosts. For more information about **Iometer**, see the **Iometer** page at this website:

<http://www.iometer.org>



IBM Easy Tier function

This chapter describes the functions that are provided by the IBM Easy Tier feature of the IBM Spectrum Virtualize and Storwize family products for disk performance optimization. It also describes some implementation guidelines. Finally, an overview of the monitoring capabilities is described.

This chapter includes the following sections:

- ▶ Easy Tier
- ▶ Easy Tier implementation considerations
- ▶ Monitoring tools

7.1 Easy Tier

In today's storage market, SSDs and flash arrays are emerging as an attractive alternative to hard disk drives (HDDs). Because of their low response times, high throughput, and IOPS-energy-efficient characteristics, SSDs and flash arrays have the potential to enable your storage infrastructure to achieve significant savings in operational costs.

However, the current acquisition cost per gibibyte (GiB) for SSDs or flash arrays is higher than for HDDs. SSD and flash array performance depends greatly on workload characteristics/ Therefore, they should be used with HDDs for optimal performance.

Choosing the correct mix of drives and the correct data placement is critical to achieve optimal performance at low cost. Maximum value can be derived by placing "hot" data with high I/O density and low response time requirements on SSDs or flash arrays, while targeting HDDs for "cooler" data which is accessed more sequentially and at lower rates.

Easy Tier automates the placement of data among different storage tiers, and it can be enabled for internal and external storage. This IBM Spectrum Virtualize and Storwize family system feature boosts your storage infrastructure performance to achieve optimal performance through a software, server, and storage solution.

Additionally, the Easy Tier feature called *storage pool balancing*, introduced in V7.3, automatically moves extents within the same storage tier from overloaded to less loaded managed disks (MDisks). Storage pool balancing ensures that your data is optimally placed among all disks within storage pools.

7.1.1 Easy Tier concepts

IBM Spectrum Virtualize and Storwize products implement Easy Tier enterprise storage functions, which were originally available on IBM DS8000 enterprise class storage systems. It enables automated subvolume data placement throughout different or within the same storage tiers. This feature intelligently aligns the system with current workload requirements and optimizes the usage of SSDs or flash arrays.

This functions includes the ability to automatically and non-disruptively relocate data (at the extent level) from one tier to another tier, or even within the same tier, in either direction. This process achieves the best available storage performance for your workload in your environment. Easy Tier reduces the I/O latency for hot spots, but it does not replace storage cache.

Both Easy Tier and storage cache solve a similar access latency workload problem. However, these two methods weigh differently in the algorithmic construction that is based on *locality of reference*, recency, and frequency. Because Easy Tier monitors I/O performance from the device end (after cache), it can pick up the performance issues that cache cannot solve, and complement the overall storage system performance.

Figure 7-1 shows placement of the Easy Tier engine within the IBM Spectrum Virtualize software stack.

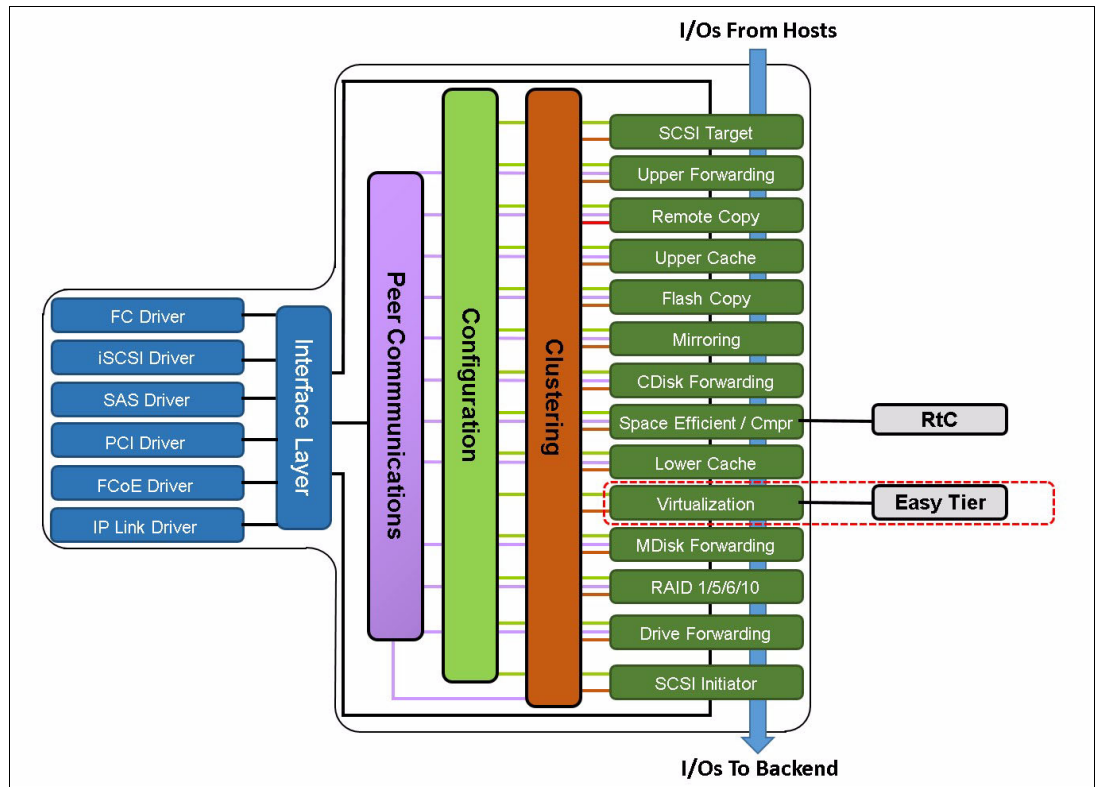


Figure 7-1 Easy Tier in the software stack

In general, the storage environment's I/O is monitored at a volume level, and the entire volume is always placed inside one appropriate storage tier. Determining the amount of I/O, moving part of the underlying volume to an appropriate storage tier, and reacting to workload changes is too complex for manual operation. This is where the Easy Tier feature can be used.

Easy Tier is a performance optimization function that automatically migrates extents that belong to a volume between different storage tiers (see Figure 7-2 on page 276) or the same storage tier (see Figure 7-7 on page 280). Because this migration works at the extent level, it is often referred to as *sub-logical unit number (LUN) migration*. Movement of the extents is done online and is not visible from the host point of view. As a result of extent movement, the volume no longer has all its data in one tier, but rather in two or three tiers.

Figure 7-2 shows the basic Easy Tier principle of operation.

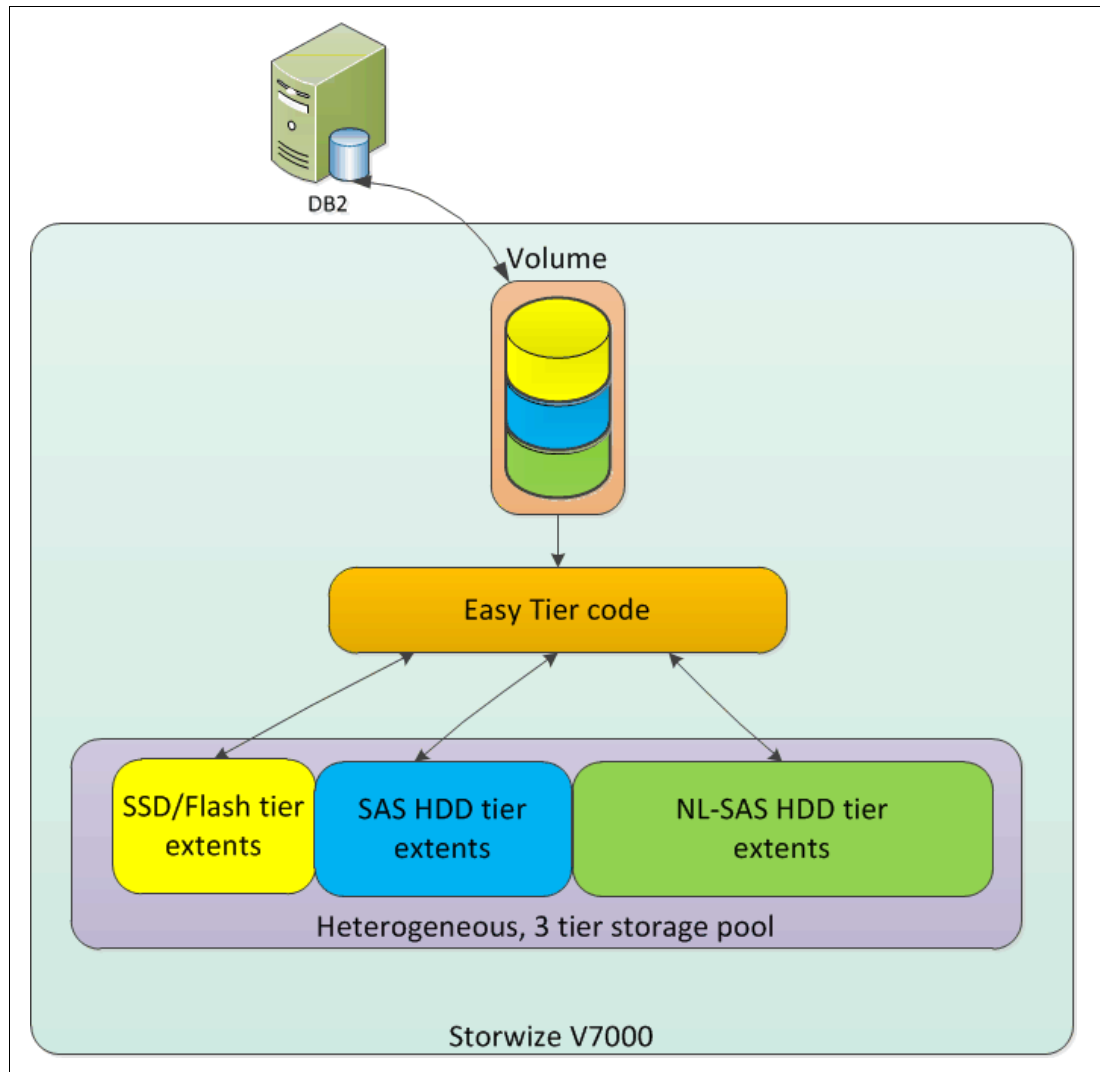


Figure 7-2 Easy Tier

You can enable Easy Tier on a volume basis. It monitors the I/O activity and latency of the extents on all Easy Tier enabled volumes over a 24-hour period. Based on the performance log, Easy Tier creates an extent migration plan and dynamically moves (promotes) high activity or hot extents to a higher disk tier within the same storage pool.

It also moves (demotes) extents whose activity dropped off, or cooled, from higher disk tier MDisk back to a lower tier MDisk. When Easy Tier runs in a storage pool rebalance mode, it moves extents from busy MDisk to less busy MDisk of the same type.

7.1.2 Four tiers Easy Tier and Read Intensive flash drive

The Easy Tier tiering model has been modified with V7.8 by adding a new tier to support Read-Intensive (RI) flash drives.

One of the reasons why flash technology is still expensive when compared to traditional HDD is that an over provisioning of the physical memory is provided to mitigate the Write Amplification issue. Read-Intensive flash drives are lower-cost flash drives with the cost reduction being achieved by having less redundant flash material. For more information, see the following website:

https://en.wikipedia.org/wiki/Write_amplification

Read Intensive support for IBM Spectrum Virtualize/Storwize systems was initially introduced with V7.7 and has been enhanced in V7.8 introducing, among other things, Easy Tier support for RI MDisks.

Even though Easy Tier still remains a three tier storage architecture, V7.8 added a new “user” tier specifically for RI MDisks (`tier1_flash`). From a user perspective, there are now four tiers (or Tech Types):

1. T0 or `tier0_flash` that represents enterprise flash technology
2. T1 or `tier1_flash` that represents RI flash technology
3. T2 or `tier2_hdd` that represents enterprise HDD technology
4. T3 or `tier3_nearline` that represents nearline HDD technology

These user tiers are mapped to Easy Tier tiers depending on the pool configuration. Figure 7-3 shows the possible combinations for the pool configuration of the four user tiers (the configurations that contain the RI user tier are highlighted in orange).

User Tiers	Easy Tier Tier (by pool configuration)													
	T0	T0+T1	T0+T1+T2	T0+T1+T2+T3	T0+T2	T0+T2+T3	T0+T3	T1	T1+T2	T1+T2+T3	T1+T3	T2	T2+T3	T3
T0 (Tier0 Flash)	1	1	1	1	1	1	1							
T1 (Tier1 Flash)		2	2	2				2	2	1	2			
T2 (Tier2 HDD)			3	2	2	2			3	2		2	2	
T3 (Tier3 Near Line)				3		3	2			3	3		3	3

Figure 7-3 Easy Tier mapping policy

The table columns represent all the possible pool configurations, while the rows show which Easy Tier tier each user tier is mapped in. For example, consider a pool with all the possible tiers configured that corresponds with the T0+T1+T2+T3 configuration in the table. With this configuration, T1 and T2 are mapped to the same Easy Tier tier (tier 2). Note that the 99 tier is only mapped to Easy Tier 1 or 2 tier.

7.1.3 SSD arrays and Flash MDisks

SSDs or flash arrays are treated no differently by the IBM Spectrum Virtualize or Storwize system than normal HDDs regarding RAID arrays or MDisks. For the Storwize systems, the individual SSDs in the storage enclosures are combined into an array, usually in RAID 10 or RAID 5 format. It is unlikely that RAID6 SSD arrays are used, because of the double parity resource requirements, with two logical SSDs used for parity only. As with usual HDDs, RAID is an MDisk of an array type and after creation is then managed the same way that the HDD MDisks are.

As is the case for HDDs, the SSD RAID array format helps to protect against individual SSD failures. Depending on your requirements, you can achieve more high availability (HA) protection above the RAID level by using volume mirroring.

The internal storage configuration of flash arrays can differ depending on an array vendor. Regardless of the methods used to configure flash-based storage, the flash system maps a volume to a host, in this case to the IBM Spectrum Virtualize or Storwize system. From the IBM Spectrum Virtualize or Storwize system perspective, a volume presented from a flash storage is also seen as a normal managed disk.

Starting with SVC 2145-DH8 nodes and software version 7.3, up to two expansion drawers can be connected to the one IBM Spectrum Virtualize I/O Group. Each drawer can have up to 24 SDDs, and only SDD drives are supported. The SDD drives are then gathered together to form RAID arrays in the same way that RAID arrays are formed in the IBM Storwize systems.

After creation of an SDD RAID array, it appears as a usual MDisk but with a tier of tier0_flash or tier1_flash, which differs from MDisk presented from external storage systems or RAID arrays made of HDDs. Because IBM Spectrum Virtualize/Storwize does not know what kind of physical disks that external MDisk are formed from, the default MDisk tier that the system adds to each external MDisk is tier2_hdd. It is up to the user or administrator to change the tier of MDisk to tier0_flash, tier1_flash, tier2_hdd, or tier3_nearline.

To change a tier of an MDisk in the CLI, use the **chmdisk** command as in Example 7-1.

Example 7-1 Changing MDisk tier

```
IBM_2145:SVC_ESC:superuser>lsmdisk -delim " "
id name status mode mdisk_grp_id mdisk_grp_name capacity ctrl_LUN_#
controller_name UID tier encrypt site_id site_name distributed dedupe
1 mdisk1 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000001 V7K_SITEB_C2
6005076802880102c000000000000200000000000000000000000000000000 tier2_hdd no 2
SITE_B no no
2 mdisk2 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000002 V7K_SITEB_C2
6005076802880102c00000000000021000000000000000000000000000000 tier2_hdd no 2
SITE_B no no
```

```
IBM_2145:SVC_ESC:superuser>chmdisk -tier tier3_nearline 1
```

```
IBM_2145:SVC_ESC:superuser>lsmdisk -delim " "
id name status mode mdisk_grp_id mdisk_grp_name capacity ctrl_LUN_#
controller_name UID tier encrypt site_id site_name distributed dedupe
1 mdisk1 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000001 V7K_SITEB_C2
6005076802880102c000000000000200000000000000000000000000000000 tier3_nearline no
2 SITE_B no no
2 mdisk2 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000002 V7K_SITEB_C2
6005076802880102c00000000000021000000000000000000000000000000 tier2_hdd no 2
SITE_B no no
```

It is also possible to change the MDisk tier from the graphical user interface (GUI), but this only applies to external MDisk. To change the tier, complete the following steps:

1. Click **Pools** → **External Storage** and click the **Plus** sign (+) next to the controller that owns the MDisk for which you want to change the tier.
2. Right-click the wanted MDisk and select **Modify Tier** (Figure 7-4 on page 279).

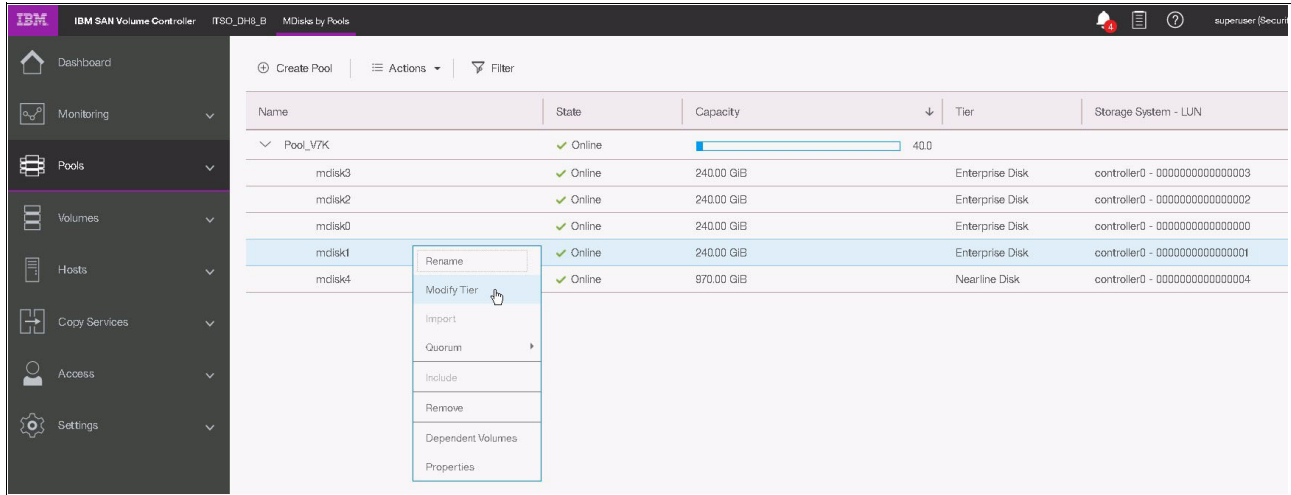


Figure 7-4 Change the MDisk tier

- The new window opens with options to change the tier (Figure 7-5).

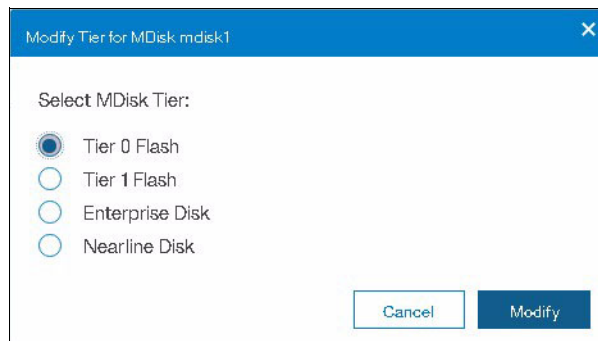


Figure 7-5 Select wanted MDisk tier

This change happens online and has no effect on hosts or availability of the volumes.

- If you do not see the *Tier* column, right-click the blue title row and select the **Tier** check box, as shown in Figure 7-6.

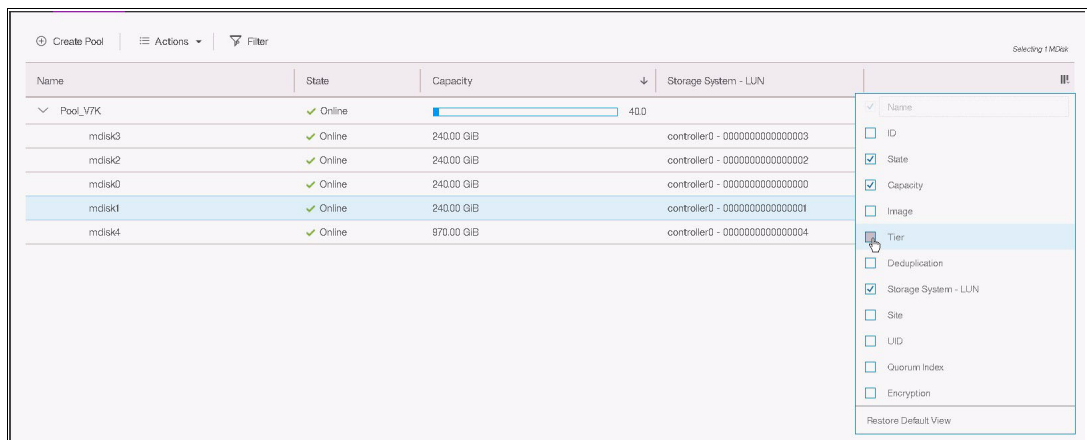


Figure 7-6 Customizing the title row to show the tier column

7.1.4 Disk tiers

The internal or external MDisks (LUNs) are likely to have different performance attributes because of the type of disk or RAID array on which they are. The MDisks can be created on 15,000 revolutions per minute (RPM) Fibre Channel (FC) or serial-attached SCSI (SAS) disks, nearline SAS (NL-SAS) or Serial Advanced Technology Attachment (SATA), or even SSDs or flash storage systems.

As mentioned in 7.1.3, “SSD arrays and Flash MDisks” on page 277, IBM Spectrum Virtualize and Storwize systems do not automatically detect the type of external MDisks. Instead, all external MDisks are initially put into the enterprise tier by default. The administrator must then manually change the MDisks tier and add them to storage pools. Depending on what type of disks are gathered to form a storage pool, two types of storage pools can be distinguished: *Single-tier* and *multitier*.

Single-tier storage pools

Figure 7-7 shows a scenario in which a single storage pool is populated with MDisks that are presented by an external storage controller. In this solution, the striped volumes can be measured by Easy Tier, and can benefit from *Storage Pool Balancing* mode, which moves extents between MDisks of the same type.

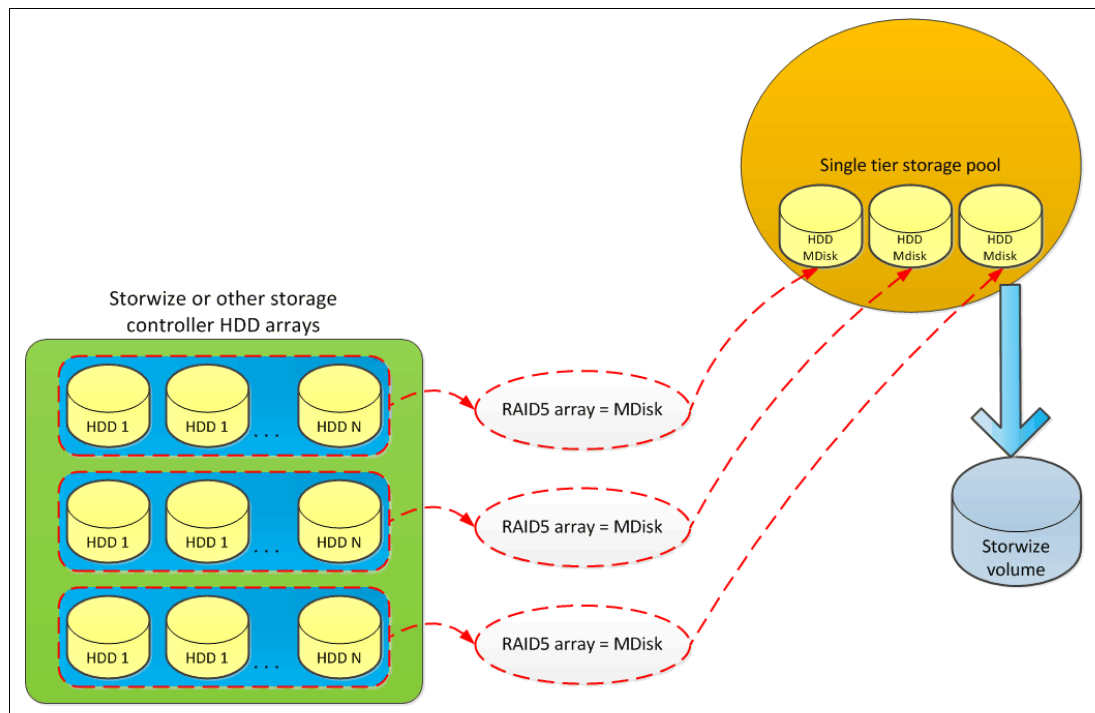


Figure 7-7 Single tier storage pool with striped volume

MDisks that are used in a single-tier storage pool should have the same hardware characteristics. These characteristics include the same RAID type, RAID array size, disk type, disk RPM, and controller performance characteristics.

Multitier storage pools

A multitier storage pool has a mix of MDisks with more than one type of disk tier attribute. This pool can be, for example, a storage pool that contains a mix of enterprise and SSD MDisks or enterprise and NL-SAS MDisks.

Figure 7-8 shows a scenario in which a storage pool is populated with three different MDisk types (one belonging to an SSD array, one belonging to an SAS HDD array, and one belonging to an NL-SAS HDD array). Although this example shows RAID 5 arrays, other RAID types can be used as well.

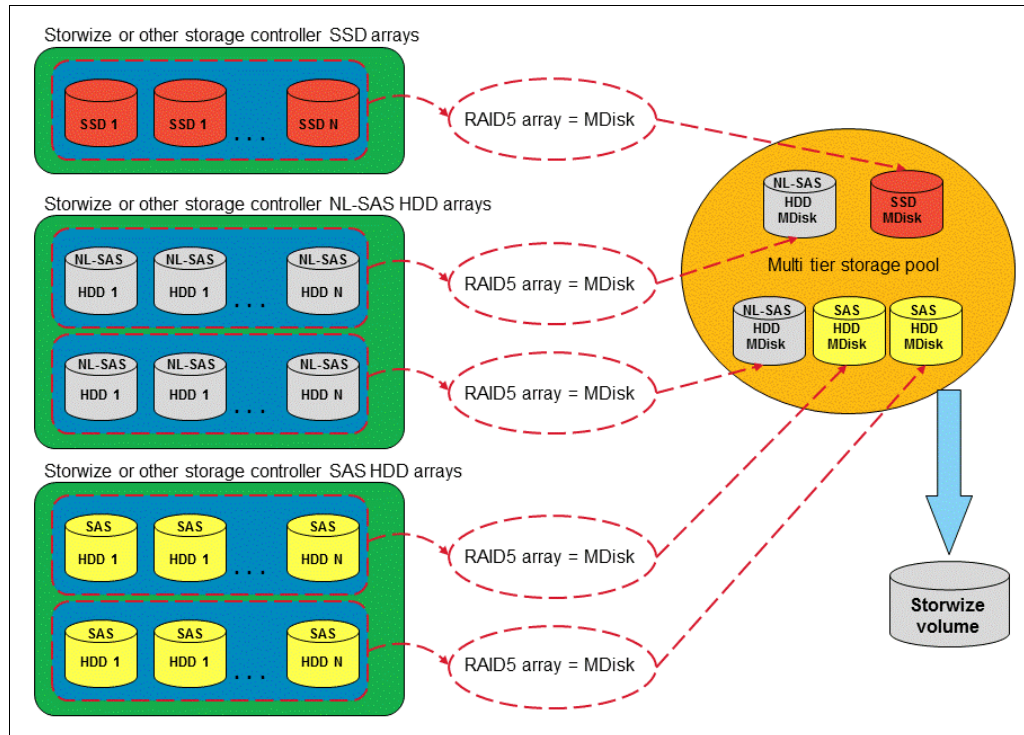


Figure 7-8 Multitier storage pool with striped volume

Adding SSDs to the pool also means that more space is now available for new volumes or volume expansion.

Note: Image mode and sequential volumes are not candidates for Easy Tier automatic data placement. They are not because all extents for those types of volumes must be on one specific MDisk, and cannot be moved.

The Easy Tier setting can be changed on a storage pool and volume level. Depending on the Easy Tier setting and the number of tiers in the storage pool, Easy Tier services might function in a different way. Table 7-1 shows possible combinations of Easy Tier setting.

Table 7-1 EasyTier settings

Storage pool Easy Tier setting	Number of tiers in the storage pool	Volume copy Easy Tier setting	Volume copy Easy Tier status
Off	One	off	inactive (see note 2)
Off	One	on	inactive (see note 2)
Off	Two - four	off	inactive (see note 2)
Off	Two - four	on	inactive (see note 2)
Measure	One	off	measured (see note 3)
Measure	One	on	measured (see note 3)

Storage pool Easy Tier setting	Number of tiers in the storage pool	Volume copy Easy Tier setting	Volume copy Easy Tier status
Measure	Two - four	off	measured (see note 3)
Measure	Two - four	on	measured (see note 3)
Auto	One	off	measured (see note 3)
Auto	One	on	balanced (see note 4)
Auto	Two - four	off	measured (see note 3)
Auto	Two - four	on	active (see note 5)
On	One	off	measured (see note 3)
On	One	on	balanced (see note 4)
On	Two - four	off	measured (see note 3)
On	Two - four	on	active (see note 5)

Table notes:

1. If the volume copy is in image or sequential mode, or is being migrated, the volume copy Easy Tier status is measured rather than active.
2. When the volume copy status is inactive, no Easy Tier functions are enabled for that volume copy.
3. When the volume copy status is measured, the Easy Tier function collects usage statistics for the volume, but automatic data placement is not active.
4. When the volume copy status is balanced, the Easy Tier function enables performance-based pool balancing for that volume copy.
5. When the volume copy status is active, the Easy Tier function operates in automatic data placement mode for that volume.

The default Easy Tier setting for a storage pool is Auto, and the default Easy Tier setting for a volume copy is On. Therefore, Easy Tier functions, except pool performance balancing, are disabled for storage pools with a single tier. Automatic data placement mode is enabled by default for all striped volume copies in a storage pool with two or more tiers.

7.1.5 Easy Tier process

The Easy Tier function includes the following four main processes:

► I/O Monitoring

This process operates continuously and monitors volumes for host I/O activity. It collects performance statistics for each extent, and derives averages for a rolling 24-hour period of I/O activity.

Easy Tier makes allowances for large block I/Os; therefore, it considers only I/Os of up to 64 kibibytes (KiB) as migration candidates.

This process is efficient and adds negligible processing resource use to the IBM Spectrum Virtualize/Storwize system nodes.

- ▶ Data Placement Advisor

The Data Placement Advisor uses workload statistics to make a cost-benefit decision as to which extents are to be candidates for migration to a higher performance tier.

This process also identifies extents that can be migrated back to a lower tier.

- ▶ Data Migration Planner (DMP)

By using the extents that were previously identified, the DMP builds the extent migration plans for the storage pool. The DMP builds two plans:

- The Automatic Data Relocation (ADR mode) plan to migrate extents across adjacent tiers
- The Rebalance (RB mode) plan to migrate extents within the same tier

- ▶ Data Migrator

This process involves the actual movement or migration of the volume's extents up to, or down from, the higher disk tier. The extent migration rate is capped so that a maximum of up to 30 megabytes per second (MBps) is migrated, which equates to approximately 3 terabytes (TB) per day that is migrated between disk tiers.

When enabled, Easy Tier performs the following actions across the tiers:

- ▶ Promote

Moves the hotter extents to a higher performance tier with available capacity. Promote occurs within adjacent tiers.

- ▶ Demote

Demotes colder extents from a higher tier to a lower tier. Demote occurs within adjacent tiers.

- ▶ Swap

Exchanges cold extent in an upper tier with hot extent in a lower tier.

- ▶ Warm Demote

Prevents performance overload of a tier by demoting a warm extent to a lower tier. This process is triggered when bandwidth or IOPS exceeds predefined threshold.

- ▶ Warm Promote

Introduced with version 7.8, this feature addresses the situation where a lower tier suddenly becomes very active. Instead of waiting for the next migration plan, Easy Tier can react immediately. Warm promote acts in a similar way to warm demote. If the 5-minute average performance shows that a layer is overloaded, Easy Tier immediately starts to promote extents until the condition is relieved.

- ▶ Cold Demote

Demotes inactive (or cold) extents that are on a higher performance tier to its adjacent lower-cost tier. In that way Easy Tier automatically frees extents on the higher storage tier before the extents on the lower tier become hot. Only supported between HDD tiers.

- ▶ Expanded Cold Demote

Demotes appropriate sequential workloads to the lowest tier to better use nearline disk bandwidth.

- ▶ Storage Pool Balancing

Redistributes extents within a tier to balance usage across MDisks for maximum performance. This process moves hot extents from high used MDisks to low used MDisks, and exchanges extents between high used MDisks and low used MDisks.

- ▶ Easy Tier attempts to migrate the most active volume extents up to SSD first.
- ▶ A previous migration plan and any queued extents that are not yet relocated are abandoned.

Note: Extent migration occurs only between adjacent tiers. For instance, in a three-tiered storage pool, Easy Tier will not move extents from the flash tier directly to the nearline tier and vice versa without moving them first to the enterprise tier.

Easy Tier extent migration types are presented in Figure 7-9.

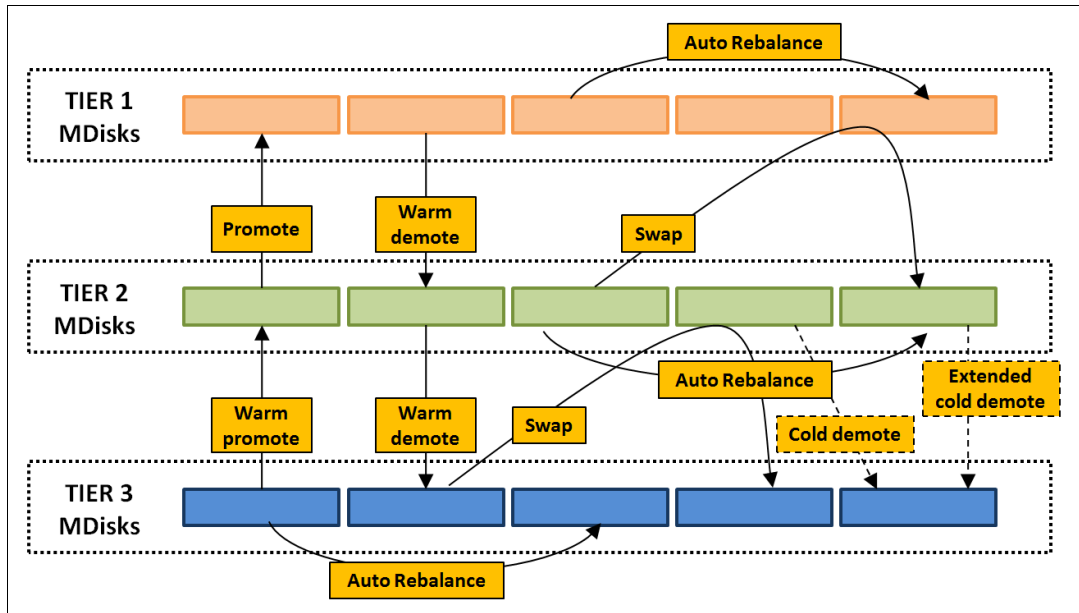


Figure 7-9 Easy Tier extent migration types

7.1.6 Easy Tier operating modes

Easy Tier includes the following main operating modes:

- ▶ Off
- ▶ Evaluation or measurement only
- ▶ Automatic data placement or extent migration
- ▶ Storage pool balancing

Easy Tier off mode

With Easy Tier turned off, no statistics are recorded, and no cross-tier extent migration occurs.

Evaluation or measurement only mode

Easy Tier Evaluation or measurement-only mode collects usage statistics for each extent in a single-tier storage pool where the Easy Tier value is set to On for both the volume and the pool. This collection is typically done for a single-tier pool that contains only HDDs so that the benefits of adding SSDs to the pool can be evaluated before any major hardware acquisition.

A `dpa_heat.nodeid.yymmdd.hhmmss.data` statistics summary file is created in the `/dumps` directory of the IBM Spectrum Virtualize node or Storwize node canisters. This file can be offloaded from the system with PuTTY Secure Copy Client (PSCP) `-load` command or by using the GUI, as described in 7.3.1, “Offloading statistics” on page 291. A web browser is used to view the report that is created by the tool.

Automatic Data Placement or extent migration mode

In Automatic data placement or extent migration operating mode, the storage pool parameter `-easytier on` or `auto` must be set, and the volumes in the pool must have `-easytier on`. The storage pool must also contain MDisks with different disk tiers, which makes it a multitier storage pool.

Dynamic data movement is transparent to the host server and application users of the data, other than providing improved performance. Extents are automatically migrated, as explained in “Implementation rules” on page 286.

The statistic summary file is also created in this mode. This file can be offloaded for input to the advisor tool. The tool produces a report on the extents that are moved to a higher tier, and a prediction of performance improvement that can be gained if more higher tier disks are available.

Options: The Easy Tier function can be turned on or off at the storage pool level *and* at the volume level.

Storage Pool Balancing

This feature assesses the extents that are written in a pool, and balances them automatically across all MDisks within the pool. This process works along with Easy Tier when multiple classes of disks exist in a single pool. In such a case, Easy Tier moves extents between the different tiers, and storage pool balancing moves extents within the same tier, to better use MDisks.

The process automatically balances existing data when new MDisks are added into an existing pool, even if the pool only contains a single type of drive. This fact does not mean that the process migrates extents from existing MDisks to achieve even extent distribution among all, old, and new MDisks in the storage pool. The Easy Tier rebalancing process within a tier migration plan is based on performance, not on the capacity of underlying MDisks.

Note: Storage pool balancing can be used to balance extents when mixing different size disks of the same performance tier. For example, when adding larger capacity drives to a pool with smaller capacity drives of the same class, Storage Pool Balancing redistributes the extents to take advantage of the additional performance of the new MDisks.

7.2 Easy Tier implementation considerations

Easy Tier comes as part of the IBM Spectrum Virtualize code. For Easy Tier to migrate extents between different tier disks, you must have disk storage available that offers different tiers (for example, a mix of SSD and HDD). With single tier (homogeneous) pools, Easy Tier uses Storage Pool Balancing only.

7.2.1 Implementation rules

Remember the following implementation and operational rules when you use the IBM System Storage Easy Tier function on the IBM Spectrum Virtualize/Storwize products:

- ▶ Easy Tier automatic data placement is not supported on image mode or sequential volumes. I/O monitoring for such volumes is supported, but you cannot migrate extents on these volumes unless you convert image or sequential volume copies to striped volumes.
- ▶ Automatic data placement and extent I/O activity monitors are supported on each copy of a mirrored volume. Easy Tier works with each copy independently of the other copy.

Volume mirroring consideration: Volume mirroring can have different workload characteristics on each copy of the data because reads are normally directed to the primary copy and writes occur to both copies. Therefore, the number of extents that Easy Tier migrates between the tiers might be different for each copy.

- ▶ If possible, the IBM Spectrum Virtualize or Storwize system creates volumes or expands volumes by using extents from MDisk from the HDD tier. However, if necessary, it uses extents from MDisk from the SSD tier.
- ▶ Do not provision the 100% of an Easy Tier enabled pool capacity. Reserve at least 16 extents for each tier for the Easy Tier movement operations.

When a volume is migrated out of a storage pool that is managed with Easy Tier, Easy Tier automatic data placement mode is no longer active on that volume. Automatic data placement is also turned off while a volume is being migrated, even when it is between pools that both have Easy Tier automatic data placement enabled. Automatic data placement for the volume is reenabled when the migration is complete.

7.2.2 Limitations

When you use Easy Tier on the IBM Spectrum Virtualize or Storwize system, remember the following limitations:

- ▶ Removing an MDisk by using the **-force** parameter
When an MDisk is deleted from a storage pool with the **-force** parameter, extents in use are migrated to MDisk in the same tier as the MDisk that is being removed, if possible. If insufficient extents exist in that tier, extents from the other tier are used.
- ▶ Migrating extents
When Easy Tier automatic data placement is enabled for a volume, you cannot use the **svctask migrateexts** CLI command on that volume.
- ▶ Migrating a volume to another storage pool
When IBM Spectrum Virtualize or Storwize system migrates a volume to a new storage pool, Easy Tier automatic data placement between the two tiers is temporarily suspended. After the volume is migrated to its new storage pool, Easy Tier automatic data placement between the generic SSD tier and the generic HDD tier resumes for the moved volume, if appropriate.
When the system migrates a volume from one storage pool to another, it attempts to migrate each extent to an extent in the new storage pool from the same tier as the original extent. In several cases, such as where a target tier is unavailable, the other tier is used. For example, the generic SSD tier might be unavailable in the new storage pool.

- ▶ Migrating a volume to an image mode
Easy Tier automatic data placement does not support image mode. When a volume with active Easy Tier automatic data placement mode is migrated to an image mode, Easy Tier automatic data placement mode is no longer active on that volume.
- ▶ Image mode and sequential volumes cannot be candidates for automatic data placement. However, Easy Tier supports evaluation mode for image mode volumes.

7.2.3 Easy Tier settings

The Easy Tier setting for storage pools and volumes can only be changed from the command-line interface. All of the changes are done online without any effect on hosts or data availability.

Turning Easy Tier on and off

Use the `chvdisk` command to turn off or turn on Easy Tier on selected volumes. Use the `chmdiskgrp` command to change status of Easy Tier on selected storage pools, as shown in Example 7-2.

Example 7-2 Changing Easy Tier setting

```
IBM_Storwize:V7000 Gen 2:superuser>chvdisk -easytier on test_vol_2
IBM_Storwize:V7000 Gen 2:superuser>chmdiskgrp -easytier auto test_pool_1
```

Tuning Easy Tier

It is also possible to change more advanced parameters of Easy Tier. These parameters should be used with caution because changing the default values can affect system performance.

Easy Tier acceleration

The first setting is called *Easy Tier acceleration*. This is a system-wide setting, and is disabled by default. Turning on this setting makes Easy Tier move extents up to four times faster than when in default setting. In accelerate mode, Easy Tier can move up to 48 GiB every 5 minutes, while in normal mode it moves up to 12 GiB. Enabling Easy Tier acceleration is advised only during periods of low system activity. The following two use cases for acceleration are the most likely:

- ▶ When adding capacity to the pool, accelerating Easy Tier can quickly spread existing volumes onto the new MDisks.
- ▶ When migrating the volumes between the storage pools in cases where the target storage pool has more tiers than the source storage pool, accelerating Easy Tier can quickly promote or demote extents in the target pool.

This setting can be changed online, without any effect on host or data availability. To turn Easy Tier acceleration mode on or off, use the `chsystem` command, as shown in Example 7-3.

Example 7-3 The chsystem command

```
IBM_Storwize:ITS0 Gen2:superuser>chsystem -easytieracceleration on
```

MDisk Easy Tier load

The second setting is called *MDisk Easy Tier load*. This setting is set on an MDisk basis, and indicates how much load Easy Tier can put on that particular MDisk. This setting has been introduced to handle situations where Easy Tier is either underutilizing or overutilizing an external MDisk. This setting doesn't apply to internal MDisks (array).

For an external MDisk, Easy Tier uses specific performance profiles based on the characteristics of the external controller and on the tier assigned to the MDisk. These performance profiles are generic, which means that they do not take into account the actual backend configuration. For instance, the same performance profile is used for a DS8000 with 300 GB 15K RPM and 1.8 TB 10K RPM. This feature is why the user is allowed to change the Easy Tier load setting to better align it with a specific external controller configuration.

There are several different values that can be set to each MDisk for the Easy Tier load:

- ▶ Default
- ▶ Low
- ▶ Medium
- ▶ High
- ▶ Very high

The system uses a default setting based on controller performance profile and the storage tier setting of the presented MDisks. If the disk drives are internal, the Easy Tier load setting is not allowed. However, an external MDisk tier should be changed by the user to align it with underlying storage.

Change the default setting to any other value only when you are certain that a particular MDisk is underutilized and can handle more load, or that the MDisk is overutilized and the load should be lowered. Change this setting to very high only for SDD and flash MDisks.

This setting can be changed online, without any effect on the hosts or data availability.

To change this setting, use command `chmdisk` as seen in Example 7-4.

Example 7-4 The chmdisk command

```
ITS0_SVC:superuser>chmdisk -easytierload high mdisk0
```

Extent size considerations

The extent size plays a major role in Easy Tier efficiency. In fact, the extent size determines the granularity level at which Easy Tier operates, which is the size of the chunk of data that Easy Tier moves across the tiers. By definition, a hot extent refers to an extent that has more I/O workload compared to other extents in the same pool and in the same tier.

It is unlikely that all the data that is contained in an extent has the same I/O workload, and therefore the same temperature. So, moving a hot extent will probably also move data that is not actually hot. The overall Easy Tier efficiency to put hot data in the proper tier is then inversely proportional to the extent size.

Consider the following practical aspects:

- ▶ Easy Tier efficiency is affecting the storage solution cost-benefit ratio. It is more effective for Easy Tier to place hot data in the top tier. In this case, less capacity can be provided for the relatively more expensive Easy Tier top tier.
- ▶ The extent size determines the bandwidth requirements for Easy Tier background process. The smaller the extent size, the lower that the bandwidth consumption is.

However, Easy Tier efficiency should not be the only factor considered when choosing the extent size. Manageability and capacity requirement considerations must also be taken into account.

As a general rule, set an extent size of either 256 MB or 512 MB for Easy Tier enabled configurations. With these extent sizes, the maximum configurable capacity for an IBM Spectrum Virtualize/Storwize system is 1 PB and 2 PB. For systems with larger capacity requirements, bigger extent sizes must be used.

External controller tiering considerations

IBM Easy Tier is an algorithm that has been developed by IBM Almaden Research and made available to many members of the IBM storage family, such as the DS8000, IBM Spectrum Virtualize, and Storwize products. The DS8000 is the most advanced in Easy Tier implementation and currently provides features that are not yet available for IBM Spectrum Virtualize and Storwize technology, such as Easy Tier Application, Easy Tier Heat Map Transfer, and Easy Tier Control.

Before V7.3, IBM Spectrum Virtualize/Storwize had basically only two tiers and no autorebalance feature was available. For this reason, when using external controllers with more advanced tiering capabilities like the DS8000, the preferred practice was to enable tiering at the backend level and leave IBM Spectrum Virtualize/Storwize Easy Tier disabled.

With V7.3 and the introduction of the autorebalance function, Easy Tier can be effectively enabled at the IBM Spectrum Virtualize/Storwize systems level. However, consider that IBM Spectrum Virtualize is not aware of the tiering functions of an external controller, and vice versa. So each tiering function makes its decisions independently. What you need to avoid is rebalance over rebalance because this situation can cancel any performance benefits.

Consider the following two options:

- ▶ Easy Tier is done at IBM Spectrum Virtualize/Storwize level:
 - a. In this case, complete these steps at the backend level:
 - i. Set up homogeneous pools according to the tier technology available.
 - ii. Create volumes to present to IBM Spectrum Virtualize/Storwize from the homogeneous pool.
 - iii. Disable tiering functions.
 - b. At an IBM Spectrum Virtualize/Storwize level, you need to complete the following actions:
 - i. Discover the MDisks provided by the backend storage and set the tier properly.
 - ii. Create hybrid pools that aggregate the MDisks.
 - iii. Enable the Easy Tier function.
- ▶ Easy Tier is done at backend level.
 - a. In this case, complete these actions at the back-end level:
 - i. Set up hybrid pools according to the tier technology available.
 - ii. Create volumes to present to IBM Spectrum Virtualize/Storwize from the hybrid pools.
 - iii. Enable the tiering functions.
 - b. At IBM Spectrum Virtualize/Storwize level, you need to complete the following actions:
 - i. Discover the MDisks provided by the backend storage and set the same tier for all.
 - ii. Create standard pools that aggregate the MDisks.
 - iii. Disable the Easy Tier function.

Even though both of these options provide benefits in term of performance, they have different characteristics.

Option 2 provides some advantages when compared to option 1. One advantage is that Easy Tier can be enabled or disabled at volume level. This feature allows users to decide which volumes will benefit from Easy Tier and which will not. With option 1, this goal cannot be achieved. Another advantage of option 1 is that the volume heat map matches directly to the host workload profile using the volumes.

With option 2, the volume heat map on the backend storage is based on the IBM Spectrum Virtualize/Storwize workload. It therefore does not exactly represent the hosts workload profile because of the effects of the IBM Spectrum Virtualize/Storwize caching. Finally, with option 1 you have the chance to change the extent size to improve the overall Easy Tier efficiency (as described in “Extent size considerations” on page 288).

However, option 1, especially with DS8000 as the backend, offers some advantages when compared to option 2. For example, when using external storage, IBM Spectrum Virtualize/Storwize Easy Tier uses generic performance profiles to evaluate the workload that can be added to a specific MDisk, as described in “MDisk Easy Tier load” on page 287.

These profiles might not exactly match the actual backend capabilities, which can lead to a resource utilization that is not optimized. With option 2, this problem rarely happens because the performance profiles are based on the real back-end configuration.

Easy Tier and remote copy considerations

When Easy Tier is enabled, the workloads that are monitored on the primary and the secondary system can differ. Easy Tier at the primary system sees a normal workload, and at the secondary system, it sees only the write workloads.

This situation means that the optimized extent distribution on the primary system can differ considerably from the one on the secondary system. The optimized extent reallocation that is based on the workload learning on the primary system is not sent to the secondary system at this time to allow the same extent optimization on both systems based on the primary workload pattern.

In a disaster recovery situation with a failover from the primary site to a secondary site, the extent distribution of the volumes on the secondary system is not optimized to match the primary workload. Easy Tier relearns the production I/O profile and builds a new extent migration plan on the secondary system to adapt to the new production workload.

It will eventually achieve the same optimization and level of performance as on the primary system. This task takes a little time, so the production workload on the secondary system might not run at its optimum performance during that period. The Easy Tier acceleration (see “Easy Tier acceleration” on page 287) feature can be used to mitigate this situation.

IBM Spectrum Virtualize or Storwize remote copy configurations that use NearLine tier at the secondary system must be carefully planned, especially when practicing disaster recovery using FlashCopy. In these scenarios, FlashCopy is usually started just before the beginning of the disaster recovery test. It is very likely that the FlashCopy target volumes are in the NearLine tier due to prolonged inactivity.

As soon as the FlashCopy is initiated, an intensive workload is usually added to the FlashCopy target volumes due to both the background and foreground I/Os. This situation can easily lead to overloading, and then possibly performance degradation of the NearLine storage tier if it is not properly sized in terms of resources.

Easy Tier and Real-time Compression considerations

When Easy Tier is enabled on compressed volumes, special considerations apply. See Chapter 11, “IBM Real-time Compression” on page 381 for further details.

Tier sizing considerations

Tier sizing is a complex task that always requires an environment workload analysis to match the performance and costs expectations. Consider the following sample configurations that address some or most common customer requirements:

- ▶ 10-20% Flash, 80-90% Enterprise
This configuration provides Flash like performance with reduced costs.
- ▶ 5% Flash, 15% Read Intensive Flash, 80% Nearline
This configuration again provides Flash like performance with reduced costs.
- ▶ 3-5% Flash, 95-97% Enterprise
This configuration provides improved performance compared to a single tier solution, and all data is guaranteed to have at least enterprise performance. It also removes the requirement for over provisioning for high access density environments.
- ▶ 3-5% Flash, 25-50% Enterprise, 40-70% Nearline
This configuration provides improved performance and density compared to a single tier solution. It also provides significant reduction in environmental costs.
- ▶ 20-50% Enterprise, 50-80% Nearline
This configuration provides reduced costs and comparable performance to a single tier Enterprise solution.

7.3 Monitoring tools

The IBM Storage Tier Advisor Tool (STAT) is a Microsoft Windows console application that analyzes heat data files produced by Easy Tier. STAT creates a graphical display of the amount of “hot” data per volume. It predicts, by storage pool, how more flash drives (or SSD capacity), enterprise drives, and nearline drives might improve system performance.

Heat data files are produced approximately once a day (that is, every 24 hours) when Easy Tier is active on one or more storage pools. These files summarize the activity per volume since the prior heat data file was produced. The heat data files can be found in the `/dumps` directory on the configuration node, and are named `dpa_heat.<node_name>.<time_stamp>.data`.

Any existing heat data file is erased after seven days. The file must be offloaded by the user and STAT must be started from a Windows console with the file specified as a parameter. The user can also specify the output directory. STAT creates a set of Hypertext Markup Language (HTML) files, and the user can then open the `index.html` file in a browser to view the results.

The IBM STAT tool can be downloaded from the IBM Support website:

<http://www.ibm.com/support/docview.wss?uid=ssg1S4000935>

7.3.1 Offloading statistics

To extract the summary performance data, use one of the following methods:

- ▶ CLI
- ▶ GUI

These methods are described next.

Using the CLI

To extract the performance data using the CLI, complete the following steps:

1. Find the most recent `dpa_heat.node_name.date.time.data` file in the cluster by entering the CLI `lsdumps` command, as shown in Example 7-5.

Example 7-5 Results for the `lsdumps` command

```
IBM_2145:SVC_ESC:superuser>lsdumps
id filename
0 reinst..trc
1 sel.000000.trc
2 ec_makevpd.000000.trc
3 rtc.race_mq_log.txt.000000.trc
...lines omitted.....
13 dpa_heat.75ACXP0.150527.123113.data
14 dpa_heat.75ACXP0.150528.123110.data
15 dpa_heat.75ACXP0.150529.021109.data
16 dpa_heat.75ACXP0.150529.181607.data
```

2. Next, perform the normal PSCP `-load` download process, as shown in Example 7-6.

Example 7-6 The `pscp` program to download the DPA heat maps

```
pscp -unsafe -load SVC_ESC
superuser@system_IP:/dumps/dpa_heat.75ACXP0.150527.123113.data your_local_directory
```

Using the GUI

If you prefer to use the GUI, complete the following steps:

1. Click **Settings** → **Support** to open the Support page, as shown in Figure 7-10. If the page does not display a list of individual log files, click **Show full log listing**.

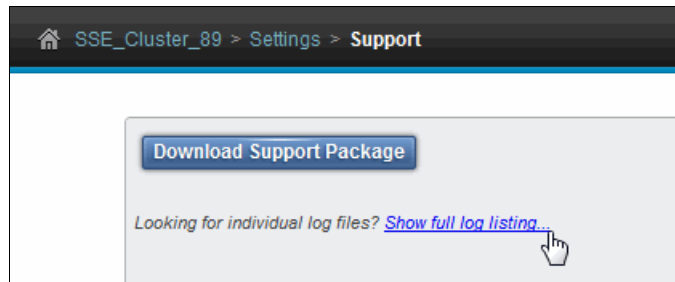


Figure 7-10 Accessing the “`dpa_heat`” file in the GUI

- Next, right-click the row for the `dpa_heat` file and choose **Download**, as shown in Figure 7-11.

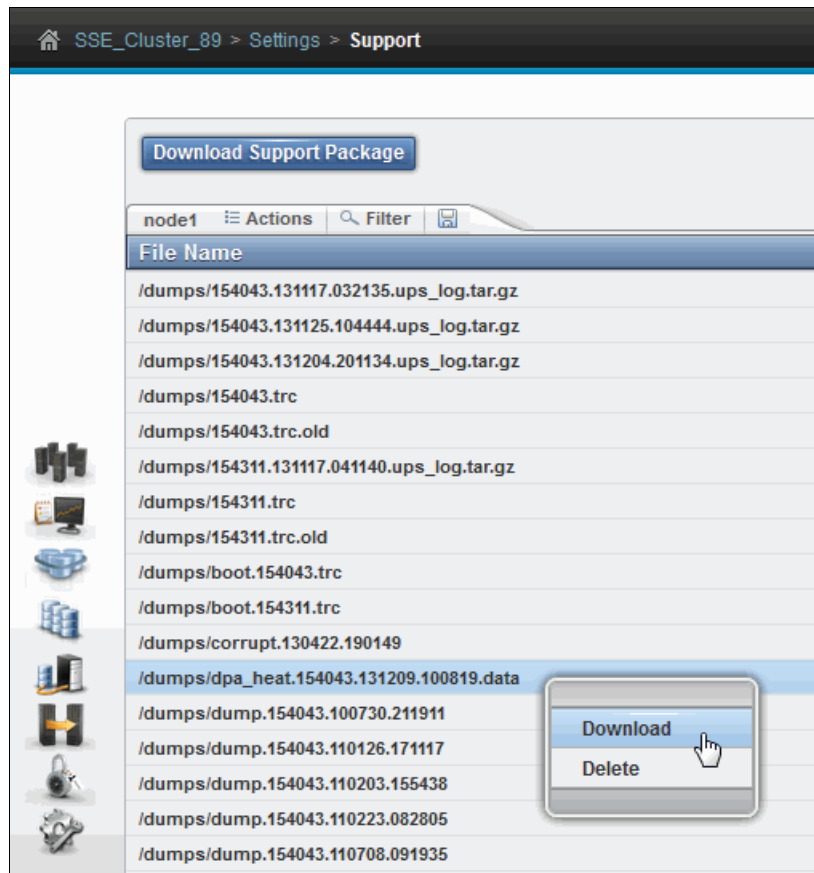


Figure 7-11 Downloading the `dpa_heat` file from the GUI

The file is downloaded to your local workstation.

Note: Starting with V8.1 the logs download is not available using the GUI.

7.3.2 Interpreting the STAT tool output

When you open the `index.html` file with your web browser, the System Summary window of the STAT output opens as shown in Figure 7-12.

On the left side of the window, two links are presented, as shown in Figure 7-12. These links allow the user to navigate between the System Report and the Systemwide Recommendation windows.

System Summary

This report is based on data from Mon Sep 05 14:16:29 2016. Easy Tier has been running continuously since Tue Aug 23 03:32:01 2016
Storage Tier Advisor Tool version: 9.2.1.1

Storage facility	IBM.2145-78STCZ0
Total storage pools monitored	2
Total volumes monitored	558
Total capacity monitored	124769.0 GiB
Hot data capacity (% of total)	4066.3 GiB (3%)
Data validity	Valid
System state	Latest Warmstart: No Warmstart Latest Failback: No Failback

Storage Pool ID *1	Capacity (GiB)	Configuration	Tier Status*2	Data Management Status*3	
P2	139909	SSD + Enterprise + NL		62134 GiB/100.00%	77775GiB
P3	139909	SSD + Enterprise + NL		62634 GiB/100.00%	77275GiB

20 Entries Per Page GO Easy Tier Management Unallocated Displaying Page 1 of 1

*1. Storage Pool ID displays the SVC extent pool ID, which is generated when the extent pool is created.
*2. Tier Status displays the status of each tier on the extent pool.

- Tier IOPS skewed indicates that the standard deviation of the current I/O per second (IOPS) utilization of disks in the tier exceeds a pre-defined threshold (10% of the average IOPS utilization across disks in the tier).
- Tier IOPS/BW overloaded indicates there is at least one disk in the tier with IOPS/BW exceeding IOPS/BW threshold.
- Tier assign stuck indicates that the process of assigning data to the tier could not make progress because the tier is probably performance overloaded.

*3. Data Management Status displays how data is managed in this extent pool.

- The dark purple portion of the Data Management Status bar displays **Assigned** for data managed by the Easy Tier Application.
- The light purple portion of the bar displays **Assign in progress** for data managed by the Easy Tier Application.
- The green portion of the bar represents data managed by Easy Tier.
- The black portion of the bar represents unallocated data.
- Each portion of the bar displays both capacity and IO percentage of the extent pool (except that the black portion of the bar only displays capacity of the unallocated data) by following the "Capacity/IO Percentage" format.

Figure 7-12 STAT main window

System Report

The System Summary window contains data that the Easy Tier monitor previously collected as shown in Figure 7-13.

System Summary

This report is based on data from Mon Sep 05 14:16:29 2016. Easy Tier has been running continuously since Tue Aug 23 03:32:01 2016
Storage Tier Advisor Tool version: 9.2.1.1

Storage facility	IBM.2145-78STCZ0
Total storage pools monitored	2
Total volumes monitored	558
Total capacity monitored	124769.0 GiB
Hot data capacity (% of total)	4066.3 GiB (3%)
Data validity	Valid
System state	Latest Warmstart: No Warmstart Latest Failback: No Failback

Storage Pool ID *1	Capacity (GiB)	Configuration	Tier Status*2	Data Management Status*3	
P2	139909	SSD + Enterprise + NL		62134 GiB/100.00%	77775GiB
P3	139909	SSD + Enterprise + NL		62634 GiB/100.00%	77275GiB

20 Entries Per Page GO Easy Tier Management Unallocated Displaying Page 1 of 1

*1. Storage Pool ID displays the SVC extent pool ID, which is generated when the extent pool is created.
*2. Tier Status displays the status of each tier on the extent pool.

- Tier IOPS skewed indicates that the standard deviation of the current I/O per second (IOPS) utilization of disks in the tier exceeds a pre-defined threshold (10% of the average IOPS utilization across disks in the tier).
- Tier IOPS/BW overloaded indicates there is at least one disk in the tier with IOPS/BW exceeding IOPS/BW threshold.
- Tier assign stuck indicates that the process of assigning data to the tier could not make progress because the tier is probably performance overloaded.

*3. Data Management Status displays how data is managed in this extent pool.

- The dark purple portion of the Data Management Status bar displays **Assigned** for data managed by the Easy Tier Application.
- The light purple portion of the bar displays **Assign in progress** for data managed by the Easy Tier Application.
- The green portion of the bar represents data managed by Easy Tier.

Figure 7-13 System Summary window

The System Summary window (Figure 7-13 on page 294) contains the following data:

- ▶ Total number of monitored pools.
- ▶ Total number of monitored volumes.
- ▶ Total capacity of monitored volumes.
- ▶ The hot data capacity, which is shown as the number of extents in GiB and percentage of the pool capacity.
- ▶ A storage pool table that shows the following information:
 - The storage pool ID.
 - The total capacity of the extent pool.
 - The configuration of the pool. Depending on whether the pool is a hybrid pool, one or two of the following options are shown:
 - Solid-state drives (SSD)
 - Enterprise
 - Nearline (NL)
 - The tier status, which potentially indicates whether a tier in the extent pool includes a skewed workload, and whether any MDisk is overloaded in terms of I/O per second (IOPS) or bandwidth.
 - The data management status, which displays how data is managed in this extent pool. The status bar includes the following indicators:
 - The dark purple portion of the bar represents data that is managed by the Easy Tier Application and the status displays **Assigned**.
 - The light purple portion of the bar represents data that is managed by the Easy Tier Application and the status displays **Assign in-progress**.
 - The green portion of the bar represents data that is managed by Easy Tier.
 - The black portion of the bar represents deallocated data.
 - Each portion of the bar displays the capacity and I/O percentage of the extent pool (except that the black portion of the bar displays only the capacity of the deallocated data) by following the “Capacity/IO Percentage” format.

Additionally, the following dates display under the report title:

- ▶ The first date refers to the time that the last data collection was considered for the sliding short-term monitoring window. That date corresponds to the last migration plan generation (if it exists) and is at most 24 hours from the Easy Tier data offload.
- ▶ The second date is the latest date when Easy Tier started to monitor the workload. It might be earlier in the past than the long-term monitoring window that is considered for the current migration plan.

Systemwide Recommendation

Figure 7-14 shows the Systemwide Recommendation window, which is opened by selecting the **Systemwide Recommendation** link in the left pane.

Systemwide Recommendation

This report is based on data from Mon Sep 05 14:16:29 2016 Easy Tier has been running continuously since Tue Aug 23 03:32:01 2016
Storage Tier Advisor Tool version: 9.2.1.1

Recommended SSD Configuration

Storage Pool ID	SSD Configuration	Predicted Performance Improvement	Total Improvement
0002	Performance Improved by Existing Spare SSD Capacity (1133.5 GB)	0% ~ 17%	0% ~ 17%
0003	Performance Improved by Adding 1 SSD_400G_R5_W8_50K Mdisk(s)	23% ~ 37%	23% ~ 37%

Recommended Enterprise Configuration

Storage Pool ID	Recommended Enterprise Configuration	Predicted IOPS Improvement	Total Improvement
0002	Existing Enterprise Mdisk(s) Free Capacity (42508.0 GB)	0% ~ 10%	0% ~ 10%
0003	Existing Enterprise Mdisk(s) Free Capacity (27722.5 GB)	0% ~ 10%	0% ~ 10%

Recommended NL Configuration

Storage Pool ID	Recommended NL Configuration	Cold Data Capacity(GiB)	Total Improvement
0003	Adding 1 7.2K_NL_4000G_R10_W8 Mdisk(s)	52815.0	52815.0
0002	Adding 1 7.2K_NL_4000G_R10_W8 Mdisk(s)	37499.0	90314.0

LEGAL DISCLAIMER:

The "Storage Tier Advisor Tool" uses limited storage performance measurement data from a user's operational environment to model potential unbalanced workload (a.k.a skew) on disk and array resources. It is intended to supplement and support, but not replace, detailed pre-installation sizing and planning analysis. It is most useful to obtain a "rule of thumb" system-wide performance projection of cumulative latency reduction on arrays and disks when a Solid State Disk configuration and the IBM Easy Tier™ function are used in combination to handle workload growth or skew management.

The "hot data" identification methodology in the tool is an engineering estimation based on expected cumulative latency reduction if the suggested Solid State Device configuration is used with the measured workload and storage configuration. Care has been taken in the development of this tool, but the accuracy of any prediction of performance improvement is subject to a variety of storage system configurations, conditions and other variables beyond the scope of this tool. Accordingly, actual results may vary.

THIS TOOL IS PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, INCLUDING ANY WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR USE. IBM MAKES NO WARRANTIES, REPRESENTATIONS OR GUARANTEES OF ACTUAL PERFORMANCE IMPROVEMENT OR DEGRADATION. IBM products and services are warranted in accordance with the agreements under which they are acquired. The user remains responsible for the results obtained from the use of any IBM product or service.

Figure 7-14 Systemwide Recommendation window

In this example, the following levels of system-wide recommendations can be displayed in this window by using the statistics data that is offloaded from an IBM Spectrum Virtualize/Storwize system:

- ▶ **Recommended SSD configuration**

This level shows a list of pools that can benefit from promoting extents on existing or added flash capacity, and the estimated system performance improvement that results from this relocation. For example, you can see that the system performance can gain up to 37% performance improvement by adding capacity from one array with 400 GB SSDs to the pool with ID 0003. This performance improvement is based on 24 hours of activity and that improvement can be higher and lower in different periods.
- ▶ **Recommended enterprise configuration**

This level shows a list of pools that can benefit from moving extents on existing or added enterprise ranks and the predicted IOPS improvement that results from this relocation.
- ▶ **Recommended nearline (NL) configuration**

This level shows a list of pools that can benefit from demoting extents to existing or other nearline ranks, and the cold data capacity that results from this cold demotion.

If the system-wide recommendation suggests adding capacity, array specifications, including drive sizes, RAID levels, and characteristics, are also shown. For the extent pool ID 0003 of the Recommended flash configuration table in Figure 7-14, the *Performance Improved by Adding 1 SSD_400G_R5_W8_50K* item indicates capacity from arrays with a 400 GB SSDs, RAID 5, and Width 8 (which means eight data drives+Parity) configuration.

Migration costs: All storage pool IDs are selectable in this report, including those pools on which the system did not offer any system-wide recommendations. You might even see some hot or warm extents for those last pools, which means that the Easy Tier algorithm decided that the migration cost for those extents was too high when compared to the benefit.

Pool Performance Statistics and Improvement Recommendation

From the System Summary window, by click a pool ID to open the Storage Pool Performance Statistics and Improvement Recommendation window as shown in Figure 7-15.

Storage Pool 0003 Performance Statistics and Improvement Recommendation						
This report is based on data from Mon Sep 05 14:16:29 2016. Easy Tier has been running continuously since Tue Aug 23 03:32:01 2016						
Storage Tier Advisor Tool version: 9.2.1.1						
SSD Tier (Average Utilization of Mdisk IOPS is 0%)						
Mdisk ID* ¹	Storage Pool ID	Mdisk type	Number of IOPS Threshold Exceeded* ²	Utilization of Mdisk IOPS* ³		Projected Utilization of Mdisk IOPS* ⁴
36	0003	SSD	0	0%	1%	0%
37	0003	SSD	0	0%	1%	0%
38	0003	SSD	0	0%	1%	0%
39	0003	SSD	0	0%	1%	0%
40	0003	SSD	0	0%	1%	0%
41	0003	SSD	0	0%	1%	0%
42	0003	SSD	0	0%	1%	0%
43	0003	SSD	0	0%	1%	0%
44	0003	SSD	0	0%	1%	0%
45	0003	SSD	0	0%	1%	0%
46	0003	SSD	0	0%	1%	0%
47	0003	SSD	0	0%	1%	0%
Enterprise Tier (Average Utilization of Mdisk IOPS is 1%)						
Mdisk ID* ¹	Storage Pool ID	Mdisk type	Number of IOPS Threshold Exceeded* ²	Utilization of Mdisk IOPS* ³		Projected Utilization of Mdisk IOPS* ⁴
25	0003	Enterprise	0	1%	1%	1%
26	0003	Enterprise	0	1%	1%	1%
27	0003	Enterprise	0	1%	1%	1%
28	0003	Enterprise	0	1%	1%	1%
29	0003	Enterprise	0	1%	1%	1%
30	0003	Enterprise	0	1%	1%	1%
31	0003	Enterprise	0	1%	1%	1%
32	0003	Enterprise	0	1%	1%	1%
33	0003	Enterprise	0	1%	1%	1%
NL Tier (Average Utilization of Mdisk IOPS is 0%)						
Mdisk ID* ¹	Storage Pool ID	Mdisk type	Number of IOPS Threshold Exceeded* ²	Utilization of Mdisk IOPS* ³		Projected Utilization of Mdisk IOPS* ⁴
34	0003	NL	0	0%	1%	0%
35	0003	NL	0	0%	1%	0%

Figure 7-15 Storage Pool Performance Statistics and Improvement Recommendation for pool ID 0003

The table shown in Figure 7-15 displays these characteristics for each MDisk in the pool:

- ▶ The MDisk ID and type.
- ▶ The number of IOPS thresholds exceeded. This number represents the number of cycles since the last decision window where the MDisk IOPS exceeded the threshold IOPS that was specified for the device type.
- ▶ The usage of the MDisk IOPS. This field shows, in three colored bars (blue, orange, and red), the current percentage of the maximum allowed IOPS threshold for the MDisks device type. The blue portion represents the percentage of IOPS below the tier average usage of the MDisks IOPS. The orange portion represents the percentage between the average and the maximum allowed IOPS for the MDisks IOPS. The red portion represents the percentage above the maximum allowed IOPS for the MDisks IOPS.
- ▶ The projected usage of the MDisk IOPS. This field shows the expected percentage of the maximum allowed IOPS threshold for the device type after the current migration plan is applied. The color code is the same as the usage of the MDisk IOPS. The percentage usage of the MDisk IOPS shows an improvement compared to the current usage, or at least the same percentage level.

Workload Distribution across tiers

Selecting Workload Distribution Across Tiers shows a figure that displays the skew of the workload, as shown in Figure 7-16. The Workload Distribution Across Tiers window includes the following components:

- ▶ The X-axis displays the top x-intensive data based on sorted data by a small I/O.
- ▶ The Y-axis denotes the accumulative small I/O percentage distributed on the top x-intensive data.

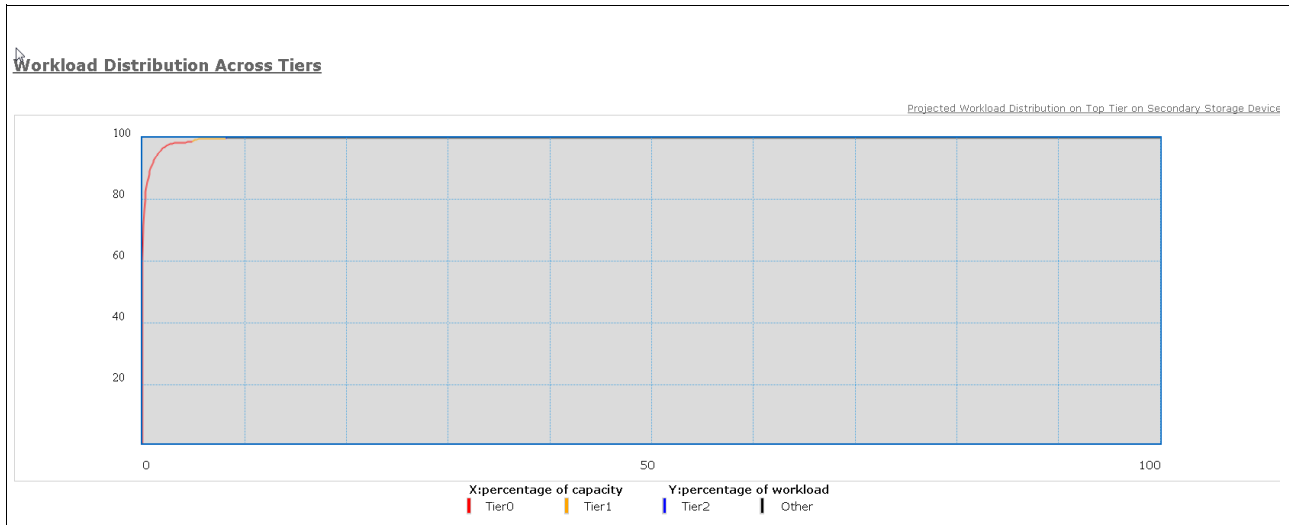


Figure 7-16 Workload Distribution Across Tier section

This report (pool workload distribution) uses the moving average of the small-block I/O only.

Recommended configurations

Click **Recommended SSD/Enterprise/NL Configuration** to open the table that contains the list of recommended SSD, Enterprise, NL, or a mix of these expands, as shown in Figure 7-17.

Recommended SSD Configuration					
SSD_600G_R10_W8_50K					
Storage Pool ID	SSD Configuration	Predicted Pool Performance Improvement	Predicted System Performance Improvement	Estimated Migration Time Range	Predicted Pool Capacity Increase (GiB)
0003	Performance Improved by Existing Spare SSD Capacity (430.5 GiB)	0% ~ 20%	0% ~ 13%	4 hour(s) ~ 38 hour(s)	-
Recommended SSD + Enterprise Configuration					
SSD_600G_R10_W8_50K					
Storage Pool ID	SSD Configuration	Predicted Pool Performance Improvement	Predicted System Performance Improvement	Estimated Migration Time Range	Predicted Pool Capacity Increase (GiB)
0003	Performance Improved by Existing Spare SSD Capacity (430.5 GiB)	0% ~ 20%	0% ~ 13%	4 hour(s) ~ 38 hour(s)	-
15K_ENT_600G_R10_W8					
Storage Pool ID	Enterprise Configuration	Predicted Pool IOPS Improvement	Predicted System IOPS Improvement	Estimated Migration Time Range	Predicted Pool Capacity Increase (GiB)
0003	Existing Enterprise Mdisk(s) IOPS Utilization Increase (1%)	-	-	0 hour(s) ~ 1 hour(s)	-
Recommended Enterprise Configuration					
15K_ENT_600G_R10_W8					
Storage Pool ID	Enterprise Configuration	Predicted Pool IOPS Improvement	Predicted System IOPS Improvement	Estimated Migration Time Range	Predicted Pool Capacity Increase (GiB)
0003	Existing Enterprise Mdisk(s) IOPS Utilization Increase (1%)	-	-	0 hour(s) ~ 1 hour(s)	-
Recommended NL Configuration					
7.2K_NL_4000G_R10_W8					
Storage Pool ID	NL Configuration	Cold Data Capacity (GiB)		Predicted Pool Capacity Increase (GiB)	
0003	Existing NL Mdisk(s) Free Capacity (49122.0 GiB)	49122.0		-	
0003	Adding 1 7.2K_NL_4000G_R10_W8 Mdisk(s)	52815.0		14901.0	

Figure 7-17 Recommendation section

The following fields are included in the tables shown in Figure 7-17 on page 298:

- ▶ The storage pool ID.
- ▶ The recommended configuration change for the specified type of MDisk and the expected result. As with the main summary report, the characteristics of the MDisk are shown (drive capacity, RAID level, and width).
- ▶ The predicted pool performance improvement percentage compared to the previous configuration.
- ▶ The predicted system performance improvement percentage compared to the previous configuration (as shown in Figure 7-14 on page 296).
- ▶ The estimated migration time range, in the use of the existing SSDs in the pool (within the current migration plan), or after the SSD capacity is added to the pool.
- ▶ For nearline: The cold data capacity that can be expected to be used on the proposed configuration.
- ▶ The predicted pool capacity increase after the potential add-on (no value if the system-wide recommendation was to use existing ranks in the tier).

By using the drop-down menu, you can change the display to another MDisk configuration, if another MDisk configuration is proposed for that selected tier. Figure 7-18 shows the drop-down menu for the Enterprise tier.

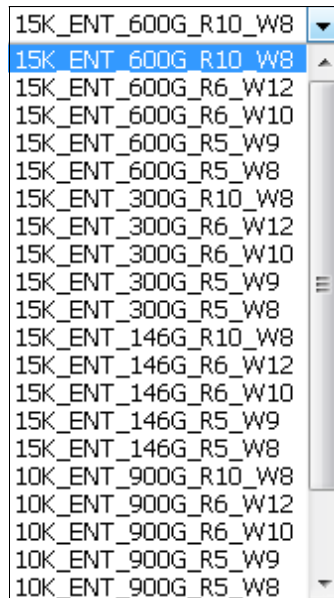


Figure 7-18 Drop-down enterprise menu

Volume Heat Distribution

By clicking Volume Heat Distribution, the heat distribution table opens, as shown on Figure 7-19. For each volume in the corresponding pool, the heat distribution table shows the following fields:

- ▶ The volume ID.
- ▶ The Copy ID.
- ▶ The volume's configured capacity.
- ▶ The three tiers (SSD, Enterprise, and NL), with the extent capacities already allocated on the respective tiers.
- ▶ The heat distribution of the volume, which is visible through the following color-coded table cells:
 - The blue portion of the bar represents the capacity of cold data on the volume.
 - The orange portion of the bar represents the capacity of warm data on the volume.
 - The red portion of the bar represents the capacity of hot data on the volume.
- ▶ Depending on the number of volumes in the extent pool, the display is divided into pages, between which you can browse by clicking the double-left and double-right angle brackets (<< and >>) on the line below the heat map. It is also possible to change the number of displayed volumes per page or to enter a page number and click **GO** to jump to that page.

Volume Heat Distribution						
Vdisk ID *7	Copy ID *8	Configured Size *9	IO Percentage of Extent Pool	Tier	Capacity on Tier *10	Heat Distribution *11
0	1	50.00 GiB	0.00%	SSD Tier	50.00 GiB	50.00 GiB
				Enterprise Tier	0.00 GiB	
				NL Tier	0.00 GiB	
256	1	100.00 GiB	0.02%	SSD Tier	60.50 GiB	60.50 GiB
				Enterprise Tier	38.50 GiB	38.50 GiB
				NL Tier	1.00 GiB	1.00 GiB
2	1	4097.00 GiB	0.09%	SSD Tier	1.50 GiB	0.50 GiB
				Enterprise Tier	4095.50 GiB	4095.50 GiB
				NL Tier	0.00 GiB	1.00 GiB
194	1	250.00 GiB	0.14%	SSD Tier	24.50 GiB	20.00 GiB
				Enterprise Tier	225.50 GiB	4.50 GiB
				NL Tier	0.00 GiB	1.00 GiB
4	1	400.00 GiB	1.84%	SSD Tier	44.50 GiB	5.50 GiB
				Enterprise Tier	353.50 GiB	38.50 GiB
				NL Tier	2.00 GiB	23.00 GiB
5	1	1024.00 GiB	4.56%	SSD Tier	280.00 GiB	115.00 GiB
				Enterprise Tier	741.00 GiB	662.50 GiB
				NL Tier	3.00 GiB	163.50 GiB
6	1	1024.00 GiB	0.01%	SSD Tier	1.00 GiB	3.00 GiB
				Enterprise Tier	1019.50 GiB	1.00 GiB
				NL Tier	3.50 GiB	0.50 GiB
7	1	100.00 GiB	0.00%	SSD Tier	0.00 GiB	94.50 GiB
				Enterprise Tier	94.50 GiB	5.50 GiB
				NL Tier	5.50 GiB	0.50 GiB
10	0	50.00 GiB	0.01%	SSD Tier	0.50 GiB	0.50 GiB
				Enterprise Tier	49.50 GiB	49.00 GiB
				NL Tier	0.00 GiB	0.50 GiB
11	1	50.00 GiB	0.07%	SSD Tier	4.50 GiB	3.50 GiB
				Enterprise Tier	44.00 GiB	43.50 GiB
				NL Tier	1.50 GiB	1.00 GiB

Figure 7-19 Volume Heat Distribution section

In the Heat Distribution column, the red areas indicate hot extents. When the entire cell is red, all extents of the corresponding volume are considered hot. Hot extents that are not already on the higher tier ranks are prioritized for promotion. The orange section indicates warm data, which data that is promoted after hot data is promoted and capacity becomes available. Cooler warm data can also be demoted as the result of a cold demotion.

The blue section indicates extents that are considered cold and currently not candidates to move to the higher tier's ranks. However, they might be moved onto nearline ranks (cold demotion), if applicable.

IBM Spectrum Virtualize works internally with a higher granularity for its Easy Tier heat buckets. Externally, only three heat categories (hot, warm, and cold) are shown.

7.3.3 IBM STAT Charting Utility

Updates to STAT have introduced more capability for reporting. As a result, when the STAT tool is run on a heat map file, an extra three comma-separated values (CSV) files are created and placed in the Data_files directory.

Figure 7-20 shows the CSV files highlighted in the Data_files directory after running the STAT tool against an IBM Storwize V7000 heat map.

```
Directory of C:\stats\Data_files
05/22/2014 10:15 AM <DIR>      .
05/22/2014 10:15 AM <DIR>      ..
05/11/2014 07:49 PM          271 banner_background.gif
05/11/2014 07:49 PM       2,819 banner_right.gif
05/11/2014 07:42 PM       9,811 banner_title.gif
05/11/2014 07:42 PM        942 head.html
05/22/2014 10:15 AM        886 innerBottom.html
05/22/2014 10:15 AM        355 innerTop.html
05/22/2014 10:15 AM        161 KD8P1BP_data_movement.csv
05/22/2014 10:15 AM     19,796 KD8P1BP_skew_curve.csv
05/22/2014 10:15 AM        1,712 KD8P1BP_workload_ctg.csv
05/11/2014 07:49 PM        8,515 product.jpg
05/22/2014 10:15 AM     51,468 pool_rec_p0000.html
05/22/2014 10:15 AM     59,218 pool_rec_p0001.html
05/11/2014 07:42 PM        8,236 product.jpg
05/22/2014 10:15 AM     14,565 System Summary.html
05/22/2014 10:15 AM        4,220 Systemwide Recommendation.html
          15 File(s)      177,289 bytes
          2 Dir(s)    1,290,362,880 bytes free

C:\stats\Data_files>
```

Figure 7-20 CSV files created by the STAT tool for Easy Tier

In addition to the STAT tool, IBM Spectrum Virtualize has another utility. This is a Microsoft Excel file for creating additional graphical reports of the workload that Easy Tier performs. The *IBM STAT Charting Utility* takes the output of the three CSV files and turns them into graphs for simple reporting. The STAT Charting Utility is a powerful tool for the Easy Tier planning activities. It offers a set of pre-configured Pivot Charts that provide detailed information about the workload profiles, the Easy Tier activity, and the workload skew.

With the STAT Charting Utility, it is possible to make a comprehensive and detailed analysis of the environment for a more effective tier sizing and workload analysis.

The new graphs display the following information:

► Workload Categorization

New workload visuals help you compare activity across tiers within and across pools to help determine the optimal drive mix for the current workloads. The output is illustrated in Figure 7-21.

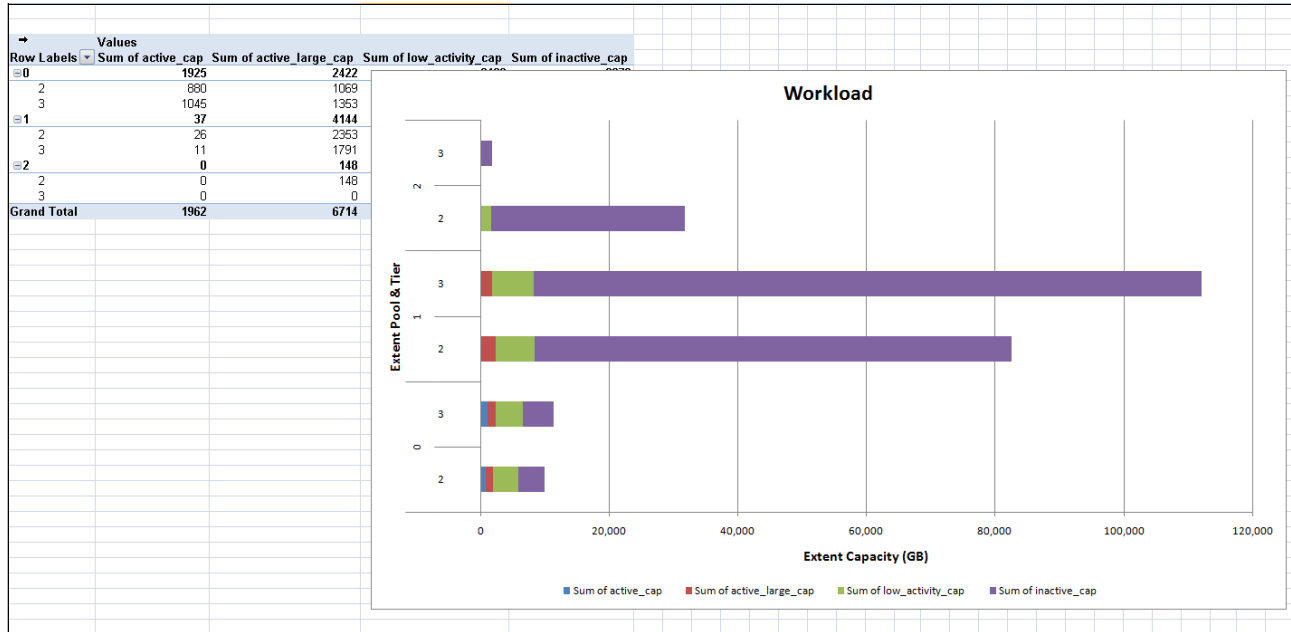


Figure 7-21 STAT Charting Utility Workload Categorization report

Using the pivot table features, you can obtain information at tier, pool, and even single volume level. The data is classified into four types depending on the workload profile:

- *Inactive*: Data with zero IOPS /Extent access density (no recent activity)
- *Low_activity*: Data with less than 0.1 IOPS /Extent access density
- *Active*: Data with more than 0.1 IOPS/Extent access density for small IOPS (transfer size < 57 KiB for CKD and < 64 KiB for FB)
- *Active_large*: All data that is not classified above (transfer size >= 57 KiB for CKD and >= 64 KiB for FB)

For each of these data types, many statistics are available that can be used for a detailed workload analysis.

► Data Movement report

The data movement reports provide information about the extents moving activity in 5-minute intervals. The output is illustrated in Figure 7-22.

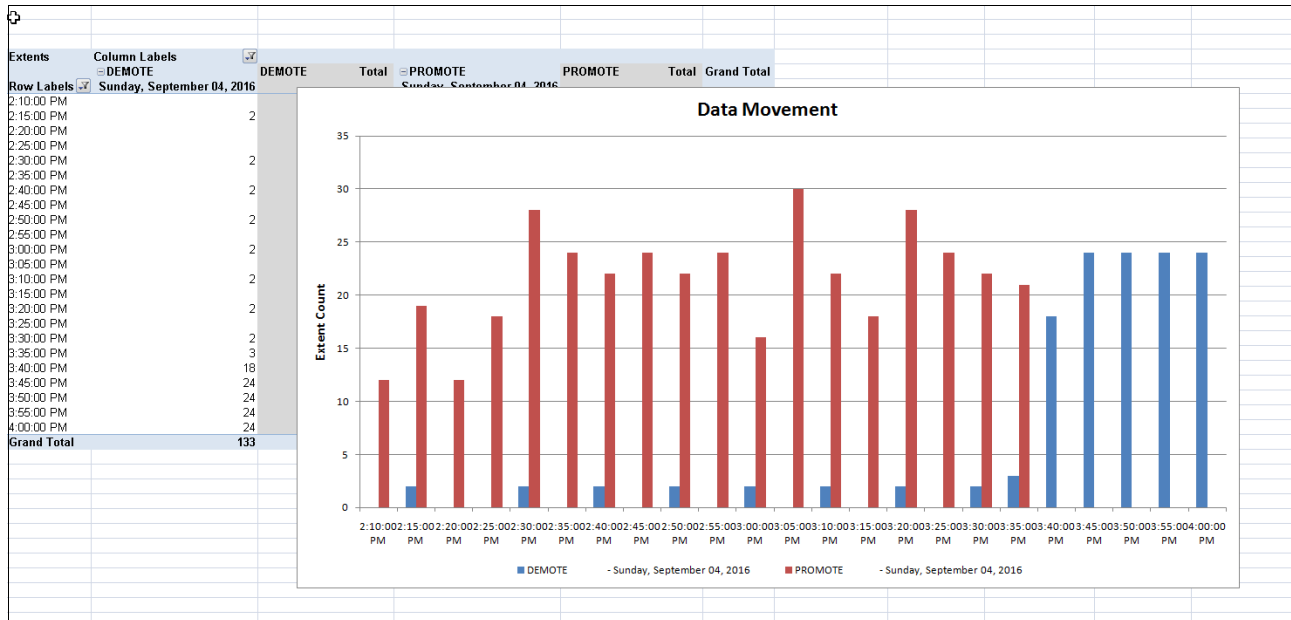


Figure 7-22 STAT Charting Utility Data Movement report

Using the pivot table features, you can obtain information at system and even single volume level. The data movement is classified into five types:

- Promote
- Demote
- Swap
- Auto Rebalance
- Warm Demote

Workload Skew report

This report shows the skew of all workloads across the system in a graph to help you visualize and accurately tier configurations when you add capacity or a new system. The output is illustrated in Figure 7-23.

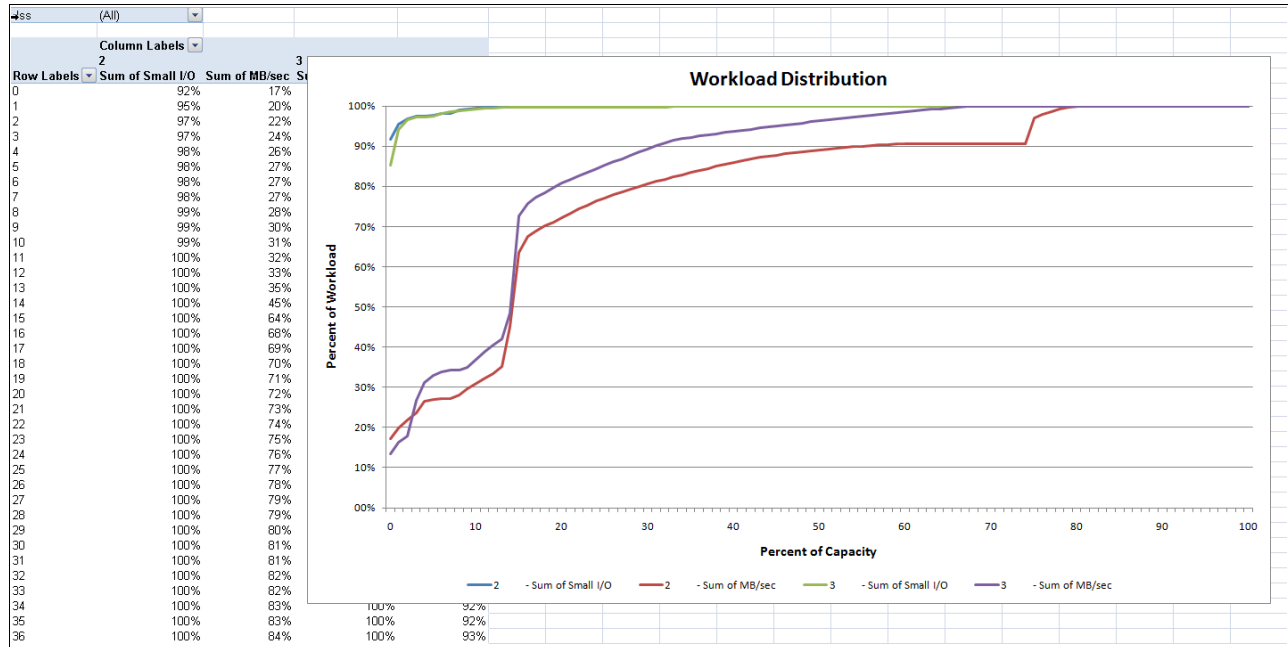


Figure 7-23 STAT Charting Utility Workload Skew report

This report provides detailed information about the workload distribution regarding the capacity. Both throughput and IOPS statistics are used.



Monitoring

Monitoring in a storage environment is crucial and it is part of what usually is called *storage governance*.

With a robust and reliable storage monitoring system, you can save significant money and minimize pain in your operation, by monitoring and predicting utilization bottlenecks in your virtualized storage environment.

This chapter provides suggestions and the basic concepts of how to implement a storage monitoring system for IBM Spectrum Virtualize/Storwize using their specific functions or external IBM Tools.

This chapter includes the following sections:

- ▶ Generic monitoring
- ▶ Performance Monitoring
- ▶ Metro and Global Mirror monitoring with IBM Copy Services Manager and scripts
- ▶ Monitoring Tier1 SSD

8.1 Generic monitoring

With IBM Spectrum Virtualize/Storwize, you can implement generic monitoring using IBM Spectrum Virtualize/Storwize specific functions that are integrated with the product itself without adding any external tools or cost.

Note: At the time of writing, a hot spare node which is in `online_spare` status shows as failed/offline in most scripts and also on IBM Spectrum Control 5.2.15, which was released before the Hot Spare Node feature. If you use manual scripts, add the value `online_spare` as a healthy condition as well.

8.1.1 Monitoring with the GUI

The management GUI is the primary tool that is used to service your system. Regularly monitor the status of the system by using the management GUI. If you suspect a problem, use the management GUI first to diagnose and resolve the problem.

Use the views that are available in the management GUI to verify the status of the system, the hardware devices, the physical storage, and the available volumes. The **Monitoring** → **Events** window provides access to all problems that exist on the system. Use the **Recommended Actions** filter to display the most important events that need to be resolved.

If there is a service error code for the alert, you can run a fix procedure that assists you in resolving the problem. These fix procedures analyze the system and provide more information about the problem. They suggest actions to take and step you through the actions that automatically manage the system where necessary. Finally, they check that the problem is resolved.

If an error is reported, always use the fix procedures within the management GUI to resolve the problem. Always use the fix procedures for both system configuration problems and hardware failures. The fix procedures analyze the system to ensure that the required changes do not cause volumes to be inaccessible to the hosts. The fix procedures automatically perform configuration changes that are required to return the system to its optimum state.

Email notification

The Call Home feature transmits operational and event-related data to you and IBM through a Simple Mail Transfer Protocol (SMTP) server connection in an event notification email. When configured, this function alerts IBM service personnel about hardware failures and potentially serious configuration or environmental issues.

SNMP notification

Simple Network Management Protocol (SNMP) is a standard protocol for managing networks and exchanging messages. The system can send SNMP messages that notify personnel about an event. You can use an SNMP manager to view the SNMP messages that are sent by the SVC.

The MIB file describes the format of the SNMP messages that are sent by IBM Spectrum Virtualize/Storwize. Use this MIB file to configure a network management program to receive SNMP event notifications that are sent from an IBM Spectrum Virtualize/Storwize system. This MIB file is suitable for use with SNMP messages from all versions of IBM Spectrum Virtualize/Storwize.

IBM Spectrum Virtualize/Storwize MIB file can be downloaded at:

ftp://ftp.software.ibm.com/storage/san/sanvc/SVC_MIB_8.1.0.MIB

Syslog notification

The syslog protocol is a standard protocol for forwarding log messages from a sender to a receiver on an IP network. The IP network can be IPv4 or IPv6. The system can send Syslog messages that notify personnel about an event. You can configure a syslog server to receive log messages from various systems and store them in a central repository.

IBM Storage Mobile Dashboard

IBM Storage Mobile Dashboard is a no charge application that provides basic monitoring capabilities for IBM storage systems. You can securely check the health and performance status of your IBM Spectrum Virtualize/Storwize system by viewing events and performance metrics.

To install IBM Storage Mobile Dashboard on an iOS device, open the App Store and search for *IBM Storage Mobile Dashboard*.

8.1.2 Monitoring using quotas and alert

In an IBM Spectrum Virtualize/Storwize system, the space usage of storage pools and Thin Provisioned or Compressed Volumes can be monitored by setting some specific quota alerts.

Storage pool

During storage pool configuration, you can set a warning such that when the pool capacity reaches this quota setting, an alert is issued. This setting generates a warning when the used capacity in the storage pool first exceeds the specified threshold. You can specify a `disk_size` integer, which defaults to megabytes (MB) unless the `-unit` parameter is specified. Or you can specify a `disk_size%`, which is a percentage of the storage pool size. To disable warnings, specify 0 or 0%. The default value is 0.

Volumes

Thin Provisioned and Compressed Volumes near their size limits are monitored at specified thresholds to preserve data integrity. If a volume can be shrunk to below the recommended new limit, you are advised to do so. If volume capacity cannot be reduced to meet the recommended limit, you are advised to create a non-compressed mirror of the data (if one does not exist) and delete the primary copy.

8.2 Performance Monitoring

Monitoring performance and the ability to collect historical performance metrics statistics is almost compulsory for any storage subsystem, and is for IBM Spectrum Virtualize/Storwize as well.

The next sections show what performance analysis tools are integrated with IBM Spectrum Virtualize/Storwize systems, and what IBM external tools are available to collect performance statistics to allow historical retention as well.

Remember that performance statistics are useful not only to debug or prevent some potential bottlenecks, but also to make capacity planning for future growth easier, as shown in Figure 8-1 on page 309.

8.2.1 Performance monitoring with the GUI

In IBM Spectrum Virtualize/Storwize, real-time performance statistics provide short-term status information for your systems. The statistics are shown as graphs in the management GUI.

You can use system statistics to monitor the bandwidth of all the volumes, interfaces, and MDisks that are being used on your system. You can also monitor the overall CPUs utilization for the system. These statistics summarize the overall performance health of the system and can be used to monitor trends in bandwidth and CPU utilization.

You can monitor changes to stable values or differences between related statistics, such as the latency between volumes and MDisks. These differences can then be further evaluated by performance diagnostic tools.

Additionally, with system-level statistics, you can quickly view bandwidth of volumes, interfaces, and MDisks. Each of these graphs displays the current bandwidth in megabytes per second and a view of bandwidth over time.

Each data point can be accessed to determine its individual bandwidth use and to evaluate whether a specific data point might represent performance impacts. For example, you can monitor the interfaces, such as for Fibre Channel or SAS interfaces, to determine whether the host data-transfer rate is different from the expected rate.

You can also select node-level statistics, which can help you determine the performance impact of a specific node. As with system statistics, node statistics help you to evaluate whether the node is operating within normal performance metrics.

The CPU utilization graph shows the current percentage of CPU usage and specific data points on the graph that show peaks in utilization. If compression is being used, you can monitor the amount of CPU resources that are being used for compression and the amount that is available to the rest of the system.

The Interfaces graph displays data points for Fibre Channel (FC), iSCSI, serial-attached SCSI (SAS), and IP Remote Copy interfaces. You can use this information to help determine connectivity issues that might affect performance.

The Volumes and MDisks graphs on the Performance window show four metrics: Read, Write, Read latency, and Write latency. You can use these metrics to help determine the overall performance health of the volumes and MDisks on your system. Consistent unexpected results can indicate errors in configuration, system faults, or connectivity issues.

Each graph represents 5 minutes of collected statistics, updated every 5 seconds, and provides a means of assessing the overall performance of your system, as shown in Figure 8-1.

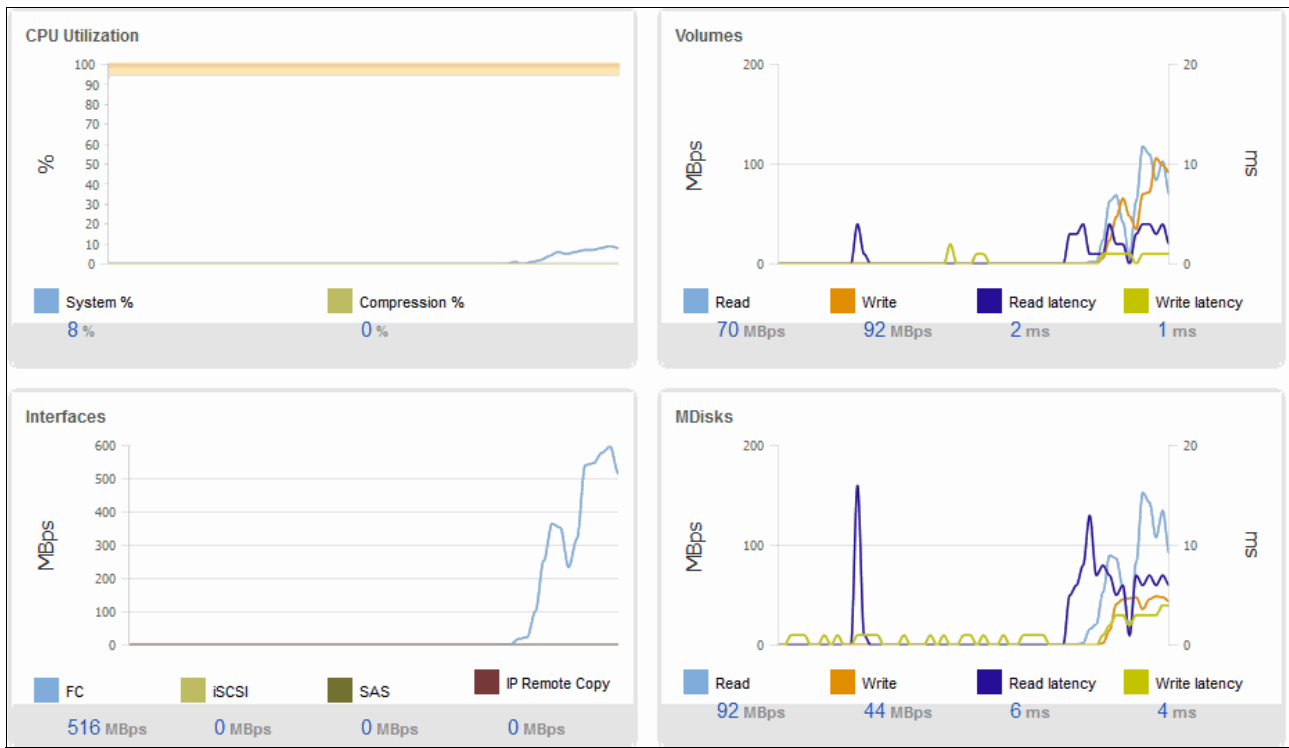


Figure 8-1 Monitoring GUI example

You can then choose the metrics that you want to be displayed, as shown in Figure 8-2.

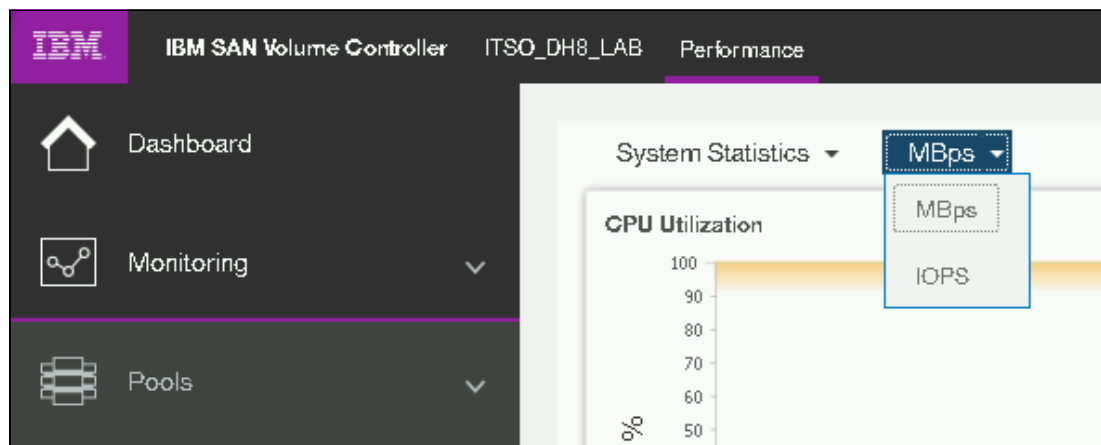


Figure 8-2 Selecting metrics

→

You can also obtain a quick overview by using the GUI option **System** → **Dashboard**, as shown in Figure 8-3.

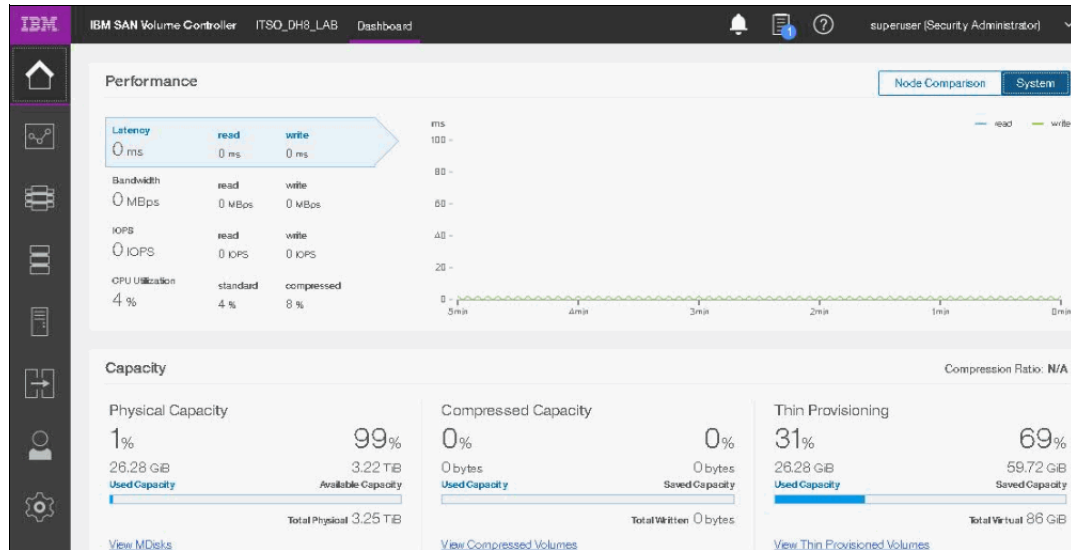


Figure 8-3 System > Dashboard

8.2.2 Performance monitoring with IBM Spectrum Control

IBM Spectrum Control offers several reports that you can use to monitor IBM Spectrum Virtualize/Storwize systems to identify performance problems. IBM Spectrum Control provides improvements to the web-based user interface that is designed to offer easy access to your storage environment.

IBM Spectrum Control provides a large amount of detailed information about IBM Spectrum Virtualize/Storwize systems. The next sections provides some basic suggestions about what metrics need to be monitored and analyzed to debug potential bottleneck problems. In addition, which alerts need to be set to be notified when some specific metrics exceed limits that are considered important for this specific environment.

For more information about the installation, configuration, and administration of IBM Tivoli® Storage Productivity Center (including how to add a storage system), see these websites:

- ▶ IBM Spectrum Control 5.2.15 Limitations and known issues:
<https://ibm.biz/BdJL79>
- ▶ Installing IBM Spectrum Control 5.2.15:
<https://ibm.biz/BdJL7C>

IBM Spectrum Control Dashboard

The performance dashboard provides Best Practice Performance Guidelines for the critical monitoring metrics. These guidelines do not represent the maximum operating limits of the related components, but are rather suggested limits that are selected with an emphasis on maintaining a stable and predictable performance profile.

The dashboard displays the *Last 24 hours* from the active viewing time and date. Selecting an individual element from the chart overlays the corresponding 24 hours for the previous day and seven days prior. This display allows for an immediate historical comparison of the respective metric. The day of reference can also be changed to allow historical comparison of previous days.

These dashboards provide two critical functions:

- ▶ Provides an “at-a-glance” view of all the critical SVC monitoring metrics.
- ▶ Provides a historical comparison of the current metric profile with previous days that enables rapid detection of anomalous workloads and behaviors.

Figure 8-4 shows how to change the day of reference.

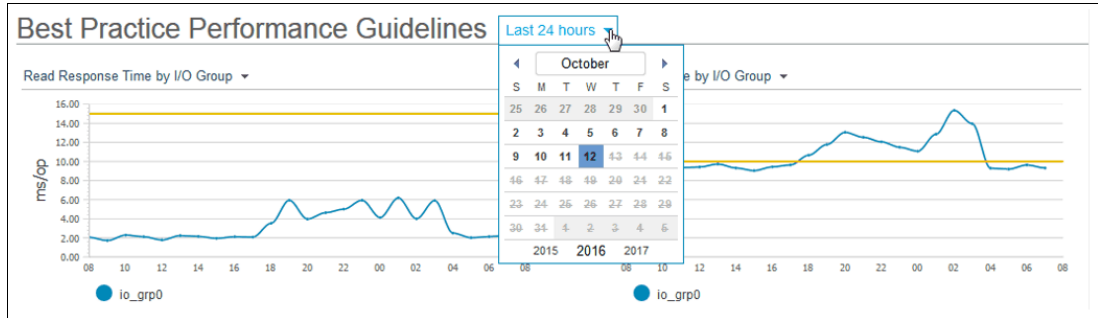


Figure 8-4 Change day of reference

Figure 8-5 shows a metric that is exceeding the best practice limit (orange line).

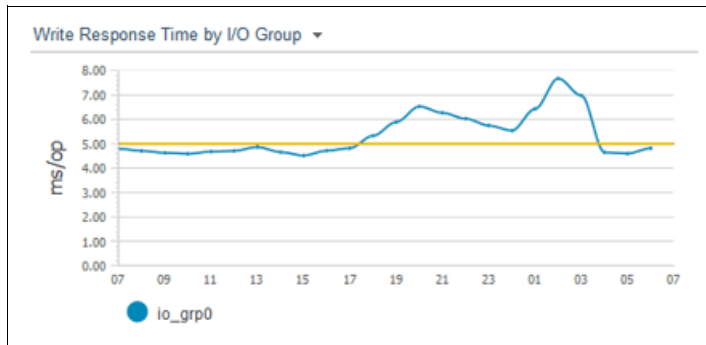


Figure 8-5 Metric exceeding best practice

Figure 8-6 shows the same chart as in Figure 8-5 with `io_grp0` selected, which overlays the previous day and 7 days prior.

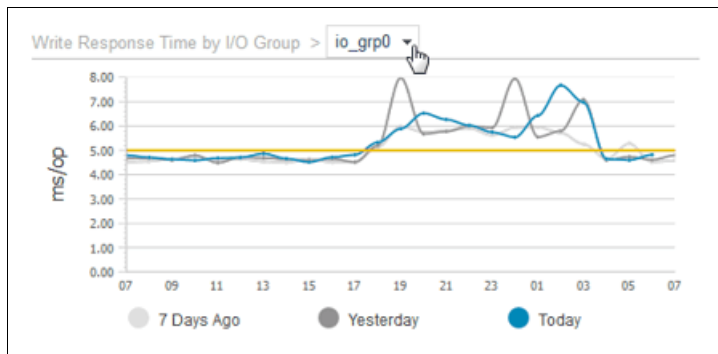


Figure 8-6 Changed chart due to iogrp selection

From this information, you can quickly conclude that this exception occurs every day at this same time, and is not a new phenomenon.

Best practice performance guidelines

You can view the key metrics that are outside of a standard range for storage systems that run IBM Spectrum Virtualize by using the performance guidelines. The guidelines were established by a historical analysis of storage environments.

Most of the performance charts show an orange line that indicates the best practice value for the metric. These guidelines are established as the levels that allow for a diverse set of workload characteristics while maintaining a stable performance profile. The other lines on each chart represent the measured values for the metric for the resources on your storage system: I/O groups, ports, or nodes.

You can use the lines to compare how close to potentially becoming overloaded your resources are. If your storage system is responding poorly and the charts indicate overloaded resources, you might have to better balance the workload. You can balance the workload between the nodes of the cluster, potentially adding more nodes to the cluster, or move some workload to other storage systems.

The charts show the hourly performance data measured for each resource on the selected day. Use the following charts to compare the workloads on your storage system with the best practice guidelines:

- ▶ **Node Utilization Percentage by Node:** Compare the guideline value for this metric, for example, 60% utilization, with the measured value from your system.
- ▶ **Overall Port Bandwidth Percentage by Port:** Compare the guideline value for this metric, for example, 50%, with the measured value from your system. Because a cluster can have many ports, the chart shows only the eight ports with the highest average bandwidth over the selected day.
- ▶ **Port-to-Local Node Send Response Time by Node:** Compare the guideline value for this metric, for example, 0.6 ms/op, with the measured value from your system.
- ▶ **Port-to-Remote Node Send Response Time by Node:** Because latencies for copy-services operations can vary widely, a guideline is not established for this metric. Use this chart to identify any discrepancies between the data rates of different nodes.
- ▶ **Read Response Time by I/O Group:** Compare the guideline value for this metric, for example, 15 ms/op, with the measured value from your system.
- ▶ **System CPU Utilization by Node:** Compare the guideline value for this metric, for example, 70% utilization, with the measured value from your system.
- ▶ **Total Data Rate by I/O Group:** Because data rates can vary widely, a guideline is not established for this metric. Use this chart to identify any significant discrepancies between the data rates of different I/O groups because these discrepancies indicate that the workload is not balanced.
- ▶ **Write Response Time by I/O Group:** Compare the guideline value for this metric, for example, 5 ms/op, with the measured value from your system.
- ▶ **Zero Buffer Credit Percentage by Node:** Compare the guideline value for this metric, for example, 20%, with the measured value from your system.

Note: The guidelines are not thresholds, and they are not related to the alerting feature in IBM Spectrum Control. To create performance alerts that use the guidelines as thresholds, go to a resource detail window in the web-based GUI, click **Alerts** in the General section, and then click **Definitions**.

8.2.3 Important metrics for debugging

The following are some of the most important metrics that need to be analyzed to debug performance problem in IBM Spectrum Virtualize/Storwize systems. Those metrics are valid to analyze the front end (by Node, by Host, by volume) or the back end (by MDisk, by Storage Pool):

Note: R/W stands for Read and Write operations.

- ▶ **I/O Rate R/W:** The term “I/O” is used to describe any program, operation, or device that transfers data to or from a computer, and to or from a peripheral device. Every transfer is an output from one device and an input into another. Typically measured in IOPS.
- ▶ **Data Rate R/W:** The data transfer rate (DTR) is the amount of digital data that is moved from one place to another in a specific time. In case of Disk or Storage Subsystem, this metric is the amount of data moved from a host to a specific storage device. Typically measured in MB per second.
- ▶ **Response time R/W:** This is the time taken for a circuit or measuring device, when subjected to a change in input signal, to change its state by a specified fraction of its total response to that change. In case of Disk or Storage Subsystem, this is the time used to complete an I/O operation. Typically measured in ms.
- ▶ **Cache Hit R/W:** This is the percentage of times where a read data or write data can be found already in cache or can find cache free space that it can be written to.
- ▶ **Average Data Block Size R/W:** The block size is the unit of work for the file system. Every read and write is done in full multiples of the block size. The block size is also the smallest size on disk that a file can have.
- ▶ **Port-to-Local Node Queue Time (Send):** The average time in milliseconds that a send operation spends in the queue before the operation is processed. This value represents the queue time for send operations that are issued to other nodes that are in the local cluster. A good scenario has less than 1 ms on average.
- ▶ **Port Protocol Errors (Zero Buffer Credit Percentage):** The amount of time, as a percentage, that the port was not able to send frames between ports because of insufficient buffer-to-buffer credit. The amount of time value is measured from the last time that the node was reset. In Fibre Channel technology, buffer-to-buffer credit is used to control the flow of frames between ports. In our experience less is better than more. However, in the real life this metric can be from 5% on average up to 20% peak without affecting performance.
- ▶ **Port data rate (send and receive):** The average number of data in MBps for operations in which the port receives or sends data.
- ▶ **Port Protocol Errors (Zero Buffer Credit Timer):** The number of microseconds that the port is not able to send frames between ports because there is insufficient buffer-to-buffer credit. In Fibre Channel technology, buffer-to-buffer credit is used to control the flow of frames between ports. Buffer-to-buffer credit is measured from the last time that the node was reset. This value is related to the data collection sample interval.
- ▶ **Port Congestion Index:** The estimated degree to which frame transmission was delayed due to a lack of buffer credits. This value is generally 0 - 100. The value 0 means there was no congestion. The value can exceed 100 if the buffer credit exhaustion persisted for an extended amount of time. When you troubleshoot a SAN, use this metric to help identify port conditions that might slow the performance of the resources to which those ports are connected.

- ▶ **Global Mirror (Overlapping Write Percentage):** The percentage of overlapping write operations that are issued by the Global Mirror primary site. Some overlapping writes are processed in parallel, and so they are excluded from this value.
- ▶ **Global Mirror (Write I/O Rate):** The average number of write operations per second that are issued to the Global Mirror secondary site. Keep in mind that IBM Spectrum Virtualize/Storwize systems have limited number of GM I/Os that can be delivered. This amount is around 90.000 for each I/O group.
- ▶ **Global Mirror (Secondary Write Lag):** The average number of extra milliseconds that it takes to service each secondary write operation for Global Mirror. This value does not include the time to service the primary write operations. Monitor the value of Global Mirror Secondary Write Lag to identify delays that occurred during the process of writing data to the secondary site.

Many others metrics are supplied to IBM Spectrum Control from IBM Spectrum Virtualize/Storwize systems. For more information about all metrics, see the following website:

<https://ibm.biz/BdjL73>

8.2.4 Performance support package

If you have performance issues on your system at any level (Host, Volume, Nodes, Pools, and so on), consult IBM Support, who require detailed performance data about the IBM Spectrum Virtualize/Storwize system to diagnose the problem. Generate a performance support package with detailed data by using IBM Spectrum Control.

In this scenario, you export performance data for a SAN Volume Controller to a compressed package. You then send the package to IBM Support, as shown in Figure 8-7.

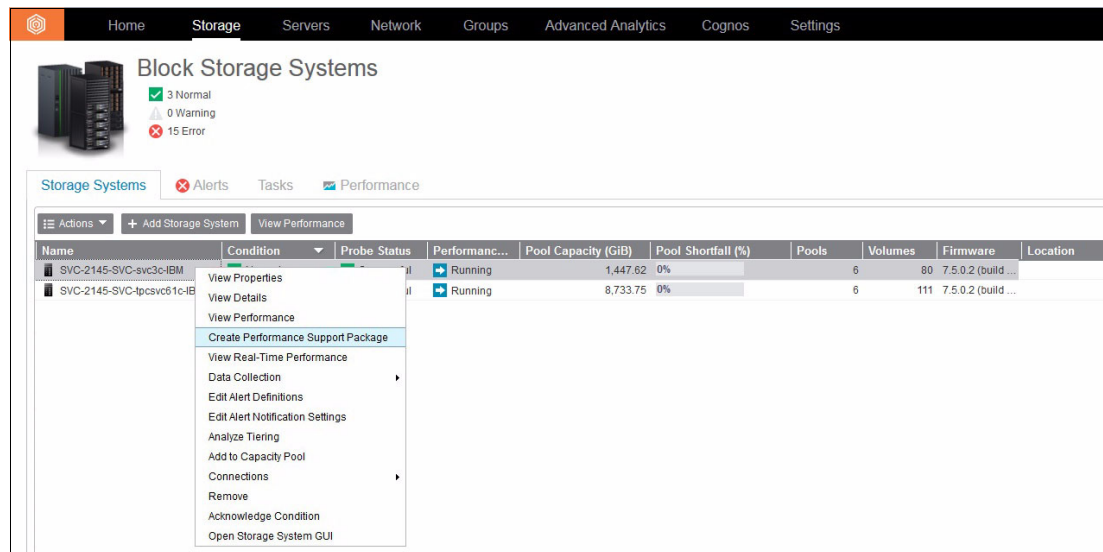


Figure 8-7 Performance support package creation

When the package has been created, you are requested to download it in .zip format. The package includes different reports in .csv format, as shown in Figure 8-8.

log.txt	Text Document	1 KB
<input type="checkbox"/> PerfReport_ITSO_SVC_ESC_Disks_20161017-233400_12hrs0mins.csv	CSV File	1 KB
<input type="checkbox"/> PerfReport_ITSO_SVC_ESC_HostConnections_20161017-233400_12hrs0mins.csv	CSV File	1 KB
<input type="checkbox"/> PerfReport_ITSO_SVC_ESC_IOGroups_20161017-233400_12hrs0mins.csv	CSV File	12 KB
<input type="checkbox"/> PerfReport_ITSO_SVC_ESC_ManagedDisks_20161017-233400_12hrs0mins.csv	CSV File	18 KB
<input type="checkbox"/> PerfReport_ITSO_SVC_ESC_Nodes_20161017-233400_12hrs0mins.csv	CSV File	18 KB
<input type="checkbox"/> PerfReport_ITSO_SVC_ESC_Pools_20161017-233400_12hrs0mins.csv	CSV File	16 KB
<input type="checkbox"/> PerfReport_ITSO_SVC_ESC_StoragePorts_20161017-233400_12hrs0mins.csv	CSV File	36 KB
<input type="checkbox"/> PerfReport_ITSO_SVC_ESC_StorageSystem_20161017-233400_12hrs0mins.csv	CSV File	12 KB
<input type="checkbox"/> PerfReport_ITSO_SVC_ESC_Volumes_20161017-233400_12hrs0mins.csv	CSV File	12 KB

Figure 8-8 Package files example

For more information about how to create a performance support package, see this website:

<https://ibm.biz/BdjLWr>

Note: The performance data might be large, especially if the data is for storage systems that have many volumes, or the performance monitors are running with a 1-minute sampling frequency. If the time range for the data is greater than 12 hours, volume data and 1-minute sample data is automatically excluded from the performance data. To include volume data and 1-minute sample data, select the **Advanced package** option on the Create Performance Support Package wizard.

8.3 Metro and Global Mirror monitoring with IBM Copy Services Manager and scripts

Copy Services Manager controls copy services in storage environments. Copy services are features that are used by storage systems such as IBM Spectrum Virtualize/Storwize systems to configure, manage, and monitor data-copy functions. Copy services include IBM FlashCopy, Metro Mirror, Global Mirror, and Global Mirror Change Volumes.

You can use Copy Services Manager to complete the following data replication tasks and help reduce the downtime of critical applications:

- ▶ Plan for replication when you are provisioning storage
- ▶ Keep data on multiple related volumes consistent across storage systems if there is a planned or unplanned outage
- ▶ Monitor and track replication operations
- ▶ Automate the mapping of source volumes to target volumes

One of the most important events that need to be monitored when IBM Spectrum Virtualize/Storwize systems are implemented in a DR solution with Metro Mirror (MM) or Global Mirror (GM) functions, is to check whether MM or GM has been suspended because of a 1920 error or 1720 error.

As explained in Chapter 5, “Copy services” on page 139, IBM Spectrum Virtualize/Storwize systems are able to suspend the MM or GM relationship to protect the performance on the primary site when MM or GM starts to affect write response time. That suspension can be caused by several factors.

IBM Spectrum Virtualize/Storwize systems do not restart the MM or GM automatically. They must be restarted manually.

Setting IBM Spectrum Virtualize/Storwize systems alert monitoring is explained in 8.1.1, “Monitoring with the GUI” on page 306. When MM or GM is managed by IBM CSM and if a 1920 error occurs, IBM CSM can automatically restart MM or GM sessions, and can set the delay time on the automatic restart option. This delay allows some time for the situation to correct itself.

Alternatively, if you have several sessions, you can stagger them so that they do not all restart at the same time, which can affect the system performance. Choose the set delay time feature to define a time, in seconds, for the delay between when Copy Services Manager processes the 1720/1920 event and when the automatic restart is issued.

CSM is also able to automatically restart unexpected suspends. When you select this option, the Copy Services Manager server automatically restarts the session when it unexpectedly suspends due to reason code 1720 or 1920. An automatic restart is attempted for every suspend with reason code 1720 or 1920 up to a predefined number of times within a 30-minute time period.

The number of times that a restart is attempted is determined by the storage server **gm1inktolerance** value. If the number of allowable automatic restarts is exceeded within the time period, the session does not restart automatically on the next unexpected suspend. Issue a **Start** command to restart the session, clear the automatic restart counters, and enable automatic restarts.

Warning: When you enable this option, the session is automatically restarted by the server. When this situation occurs, the secondary site is not consistent until the relationships are fully resynched.

You can specify the amount of time (in seconds) that the copy services management server waits after an unexpected suspend before automatically restarting the session. The range of possible values is 0 - 43200. The default is 0, which specifies that the session is restarted immediately following an unexpected suspend.

For more information about IBM Copy Service Manager, see this website:

<https://ibm.biz/BdjL7T>

8.3.1 Monitoring MM and GM with scripts

IBM Spectrum Virtualize/Storwize system provides a complete command-line interface (CLI), which allows you to interact with your systems by using scripts. Those scripts can run in the IBM Spectrum Virtualize/Storwize shell, but with a limited script command set available, or they can run out of the shell using any scripting language that you prefer.

An example of script usage is one to check at a specific interval time whether MM or GM are still active, if any 1920 errors have occurred, or to react to an SNMP or email alert received. The script can then start some specific recovery action based on your recovery plan and environment.

Customers who do not use IBM Copy Service Manager have created their own scripts. These scripts are sometimes supported by IBM as part of ITS professional services or IBM System Lab services. Tell your IBM representative what kind of monitoring you want to implement with scripts, and together try to find if one exists in the IBM Intellectual Capital Management repository that can be reused.

8.4 Monitoring Tier1 SSD

The Tier1 SSD that was released in 4Q/2016 requires that special attention is paid to the endurance events that can be triggered. For monitoring purposes, stay alert to the new fields listed in Table 8-1.

Table 8-1 Field changes to drive and array devices

Field	Description
write_endurance_used	Metric pulled from within drive (SAS spec) relating to the amount of data written across the life of the drive divided by the anticipated amount (2.42 PB for the 15.36 TB drive) Starts at 0, and can continue > 100
write_endurance_usage_rate	Measuring / Low / Marginal / High Takes 160 Days to get initial measurement; Low: Approximately 5.5 Years or more Marginal: Approximately 4.5 – 5.5 Years High: Approximately < 4.5 years High triggers event SS_EID_VL_ER_SSD_WRITE_ENDURANCE_USAGE_RATE_HIGH
replacement_date	The Current Date + Endurance Rate * Remaining Endurance Triggers event SS_EID_VL_ER_SSD_DRIVE_WRITE_ENDURANCE_LIMITED at 6 Months before limit

If you see either of these triggered events, contact your IBM service representative to put an action plan in place:

SS_EID_VL_ER_SSD_WRITE_ENDURANCE_USAGE_RATE_HI4GH
SS_EID_VL_ER_SSD_DRIVE_WRITE_ENDURANCE_LIMITED



Maintenance

Among the many benefits that the IBM Spectrum Virtualize software provides is to greatly simplify the storage management tasks that system administrators need to perform. However, as the IT environment grows and gets renewed, so does the storage infrastructure.

This chapter highlights guidance for the daily basis activities of storage administration by using the IBM Spectrum Virtualize software installed on IBM SAN Volume Controller, the IBM Storwize V7000. This guidance can help you to maintain your storage infrastructure with the levels of availability, reliability, and resiliency demanded by today's applications, and to keep up with storage growth needs.

This chapter focuses on the most important topics to consider in IBM Spectrum Virtualize administration so that you can use it as a checklist. It also provides tips and guidance.

Important: The practices that are described here were effective in many IBM Spectrum Virtualize installations worldwide for organizations in several areas. They all had one common need, which was to easily, effectively, and reliably manage their SAN storage environment. Nevertheless, whenever you have a choice between two possible implementations or configurations, if you look deep enough, you *always* have *both* advantages and disadvantages over one another.

Do not take these practices as absolute truth, but rather use them as a guide. The choice of which approach to use is ultimately yours.

This chapter includes the following sections:

- ▶ Documenting IBM Spectrum Virtualize and SAN environment
- ▶ Storage management users
- ▶ Standard operating procedures
- ▶ IBM Spectrum Virtualize code update
- ▶ SAN modifications
- ▶ Hardware upgrades for IBM Spectrum Virtualize
- ▶ Adding expansion enclosures
- ▶ I/O Throttling

9.1 Documenting IBM Spectrum Virtualize and SAN environment

This section focuses on the challenge of automating the documentation that is needed for an IBM Spectrum Virtualize solution. Consider the following points:

- ▶ Several methods and tools are available to automate the task of creating and updating the documentation. Therefore, the IT infrastructure might handle this task.
- ▶ Planning is key to maintaining sustained and organized growth. Accurate documentation of your storage environment is the blueprint with which you plan your approach to short-term and long-term storage growth.
- ▶ Your storage documentation must be conveniently available and easy to consult when needed. For example, you might need to determine how to replace your core SAN directors with newer ones, or how to fix the disk path problems of a single server. The relevant documentation might consist of a few spreadsheets and a diagram.

Storing documentation: Avoid storing IBM Spectrum Virtualize and SAN environment documentation only in the SAN. If your organization has a disaster recovery plan, include this storage documentation in it. Follow its guidelines about how to update and store this data. If no disaster recovery plan exists and you have the proper security authorization, it might be helpful to store an updated copy offsite.

In theory, this IBM Spectrum Virtualize and SAN environment documentation should be sufficient for any system administrator who has average skills in the products that are included. Make a copy that includes all of your configuration information. Use the copy to create a functionally equivalent copy of the environment by using similar hardware without any configuration, off-the-shelf media, and configuration backup files. You might need the copy if you ever face a disaster recovery scenario, which is also why it is so important to run periodic disaster recovery tests.

Create the first version of this documentation as you install your solution. If you completed forms to help plan the installation of your IBM Spectrum Virtualize solution, use these forms to help you document how your IBM Spectrum Virtualize solution was first configured. Minimum documentation is needed for an IBM Spectrum Virtualize solution. Because you might have more business requirements that require other data to be tracked, remember that the following sections do not address every situation.

9.1.1 Naming conventions

Whether you are creating your IBM Spectrum Virtualize and SAN environment documentation or you are updating what is already in place, first evaluate whether you have a good naming convention in place. With a good naming convention, you can quickly and uniquely identify the components of your IBM Spectrum Virtualize and SAN environment. System administrators can then determine whether a name belongs to a volume, storage pool, MDisk, host, or host bus adapter (HBA) by looking at it.

Because error messages often point to the device that generated an error, a good naming convention quickly highlights where to start investigating when an error occurs. Typical IBM Spectrum Virtualize and SAN component names limit the number and type of characters you can use. For example, IBM Spectrum Virtualize names are limited to 63 characters, which makes creating a naming convention a bit easier than in previous versions of IBM Spectrum Virtualize code.

Names: In previous versions of IBM Spectrum Virtualize code, names were limited to 15 characters. Starting with version 7.1, the limit is 63 characters.

Many names in IBM Spectrum Virtualize and SAN environment can be modified online. Therefore, you do not need to worry about planning outages to implement your new naming convention. The naming examples that are used in the following sections are effective in most cases, but might not be fully adequate for your particular environment or needs. The naming convention to use is your choice, but you must implement it in the whole environment.

Storage controllers

IBM Spectrum Virtualize names the storage controllers `controlerX`, with `X` being a sequential decimal number. If multiple controllers are attached to your IBM Spectrum Virtualize solution, change the name so that it includes, for example, the vendor name, the model, or its serial number. Therefore, if you receive an error message that points to `controlerX`, you do not need to log in to IBM Spectrum Virtualize to know which storage controller to check.

Note: IBM Spectrum Virtualize detects controllers based on their WWNN. If you have a storage controller that has one WWNN for each worldwide port name (WWPN), this configuration might lead to many `controlerX` names pointing to the same physical box. In this case, prepare a naming convention to cover this situation.

MDisks and storage pools

When IBM Spectrum Virtualize detects new MDisks, it names them by default as `mdiskXX`, where `XX` is a sequential number. Change the `XX` value to something more meaningful. For example, you can change it to include the following information:

- ▶ A reference to the storage controller it belongs to (such as its serial number or last digits)
- ▶ The extpool, array, or RAID group that it belongs to in the storage controller
- ▶ The LUN number or name it has in the storage controller

Consider the following examples of MDisk names with this convention:

- ▶ `23K45_A7V10`, where `23K45` is the serial number, `7` is the array, and `10` is the volume
- ▶ `75VXYZ1_02_0206`, where `75VXYZ1` is the serial number, `02` is the extpool, and `0206` is the LUN

Storage pools have several different possibilities. One possibility is to include the storage controller, the type of back-end disks, the RAID type, and sequential digits. If you have dedicated pools for specific applications or servers, another possibility is to use them instead. Consider the following examples:

- ▶ `P05XYZ1_3GR5`: Pool 05 from serial `75VXYZ1`, LUNs with 300 GB FC DDMs and RAID 5
- ▶ `P16XYZ1_EX01`: Pool 16 from serial `75VXYZ1`, pool 01 dedicated to Exchange Mail servers
- ▶ `XIV01_F9H02_ET`: Pool with disks from XIV named `XIV01` and Flash System 900 `F9H02`, both managed by Easy Tier.

Volumes (formerly VDisks)

Volume names should include the following information:

- ▶ The hosts or cluster to which the volume is mapped
- ▶ A single letter that indicates its usage by the host, as shown in the following examples:
 - `B`: For a boot disk, or `R` for a rootvg disk (if the server boots from SAN)
 - `D`: For a regular data disk

- Q: For a cluster quorum disk (do not confuse with IBM Spectrum Virtualize quorum disks)
- L: For a database logs disks
- T: For a database table disk
- ▶ A few sequential digits, for uniqueness

For example, ERPNY01_T03 indicates a volume that is mapped to server ERPNY01 and database table disk 03.

- ▶ Sessions standard for VMware datastores:
 - esx01-sessions-001: For a datastore composed of a single volume.
 - esx01-sessions-001a and esx01-sessions-001b: For a datastore composed by 2 volumes.

Hosts

In today's environment, administrators deal with large networks, the internet, and Cloud Computing. Use good server naming conventions so that they can quickly identify a server and determine the following information:

- ▶ Where it is (to know how to access it)
- ▶ What kind it is (to determine the vendor and support group in charge)
- ▶ What it does (to engage the proper application support and notify its owner)
- ▶ Its importance (to determine the severity if problems occur)

Changing a server's name in IBM Spectrum Virtualize is as simple as changing any other IBM Spectrum Virtualize object name. However, changing the name on the operating system of a server might have implications for application configuration and require a server reboot. Therefore, you might want to prepare a detailed plan if you decide to rename several servers in your network. The following example is for server name conventions for LLAATRFNN:

- ▶ LL is the location, which might designate a city, data center, building floor, or room.
- ▶ AA is a major application, for example, billing, ERP, and Data Warehouse.
- ▶ T is the type, for example, UNIX, Windows, and VMware.
- ▶ R is the role, for example, Production, Test, Q&A, and Development.
- ▶ FF is the function, for example, DB server, application server, web server, and file server.
- ▶ NN is numeric.

SAN aliases and zones

SAN aliases often need to reflect only the device and port that is associated to it. Including information about where one particular device port is physically attached on the SAN might lead to inconsistencies if you make a change or perform maintenance and then forget to update the alias. Create one alias for each device port WWPN in your SAN, and use these aliases in your zoning configuration. Consider the following examples:

- ▶ AIX_NYBIXTDB02_FC2: Interface fcs2 of AIX server NYBIXTDB02.
- ▶ LIN_POKBIXAP01_FC1: Interface fcs1 of Linux Server POKBIXAP01.
- ▶ WIN_EXCHSRV01_HBA1: Interface HBA1 of physical Windows server EXCHSRV01.
- ▶ ESX_NYVMCLUSTER01_VMHBA2: Interface vmhba2 of ESX server NYVMCLUSTER01.
- ▶ IBM_NYSVC01_N1_P1_HOST: Port 1 of Node 1 from SVC Cluster NYSVC01, dedicated for hosts/backend.
- ▶ IBM_NYSVC01_N1_P5_INTRACLUSTER: Port 5 of Node 1 from SVC Cluster NYSVC01 dedicated to intracluster traffic.

- ▶ IBM_NYSVC01_N1_P7_REPLICATION: Port 7 of Node 1 from SVC Cluster NYSVC01 dedicated to replication.

Be mindful of the IBM Spectrum Virtualize port aliases. There are mappings between the last digits of the port WWPN and the node FC port, but these mappings vary depending on the SAN Volume Controller model or the Storwize product.

- ▶ IBM_D88870_75XY131_I0301: DS8870 serial number 75XY131, port I0301.
- ▶ IBM_TL01_TD06: Tape library 01, tape drive 06.
- ▶ EMC_VNX7500_01_SPAP2: EMC VNX7500 hostname VNX7500_01, SP A, port 2.

If your SAN does not support aliases, for example, in heterogeneous fabrics with switches in some interoperations modes, use WWPNs in your zones. However, remember to update every zone that uses a WWPN if you ever change it.

Have your SAN zone name reflect the devices in the SAN it includes (normally in a one-to-one relationship) as shown in the following examples:

- ▶ SERVERALIAS_T1_SVCCLUSTERNAME (from a server to the SAN Volume Controller, where you use T1 as an identifier to zones that uses for example, node ports P1 on Fabric A, and P2 on Fabric B)
- ▶ SERVERALIAS_T2_SVCCLUSTERNAME (from a server to the SAN Volume Controller, where you use T2 as an identifier to zones that uses for example, node ports P3 on Fabric A, and P4 on Fabric B)
- ▶ IBM_DS8870_75XY131_SVCCLUSTERNAME (zone between a backend storage and the SVC Cluster)
- ▶ NYC_SVC01_POK_SVC01_REPLICATION (for remote copy services)

9.1.2 SAN fabric documentation

The most basic piece of SAN documentation is a SAN diagram. It is likely to be one of the first pieces of information you need if you ever seek support from your SAN switches vendor. Also, a good spreadsheet with ports and zoning information eases the task of searching for detailed information, which, if included in the diagram, makes the diagram difficult to use.

Brocade SAN Health

The *Brocade SAN Health Diagnostics Capture tool* is a no-cost, automated tool that can help you retain this documentation. SAN Health consists of a data collection tool that logs in to the SAN switches that you indicate and collects data by using standard SAN switch commands. The tool then creates a compressed file with the data collection. This file is sent to a Brocade automated machine for processing by secure web or e-mail.

After some time (typically a few hours), the user receives an e-mail with instructions about how to download the report. The report includes a Visio Diagram of your SAN and an organized Microsoft Excel spreadsheet that contains all of your SAN information. For more information and to download the tool, see this website:

<http://www.brocade.com/sanhealth>

The first time that you use the SAN Health Diagnostics Capture tool, explore the options provided to learn how to create a well-organized and useful diagram.

Figure 9-1 shows an example of a poorly formatted diagram.

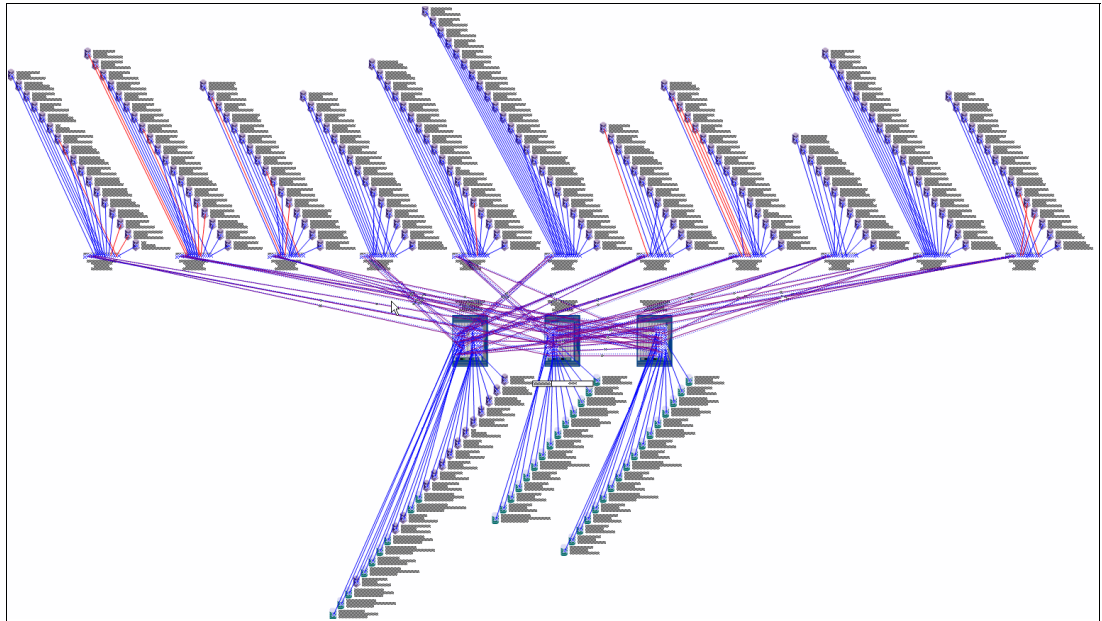


Figure 9-1 A poorly formatted SAN diagram

Figure 9-2 shows a tab of the SAN Health Options window in which you can choose the format of SAN diagram that best suits your needs. Depending on the topology and size of your SAN fabrics, you might want to manipulate the options in the Diagram Format or Report Format tabs.

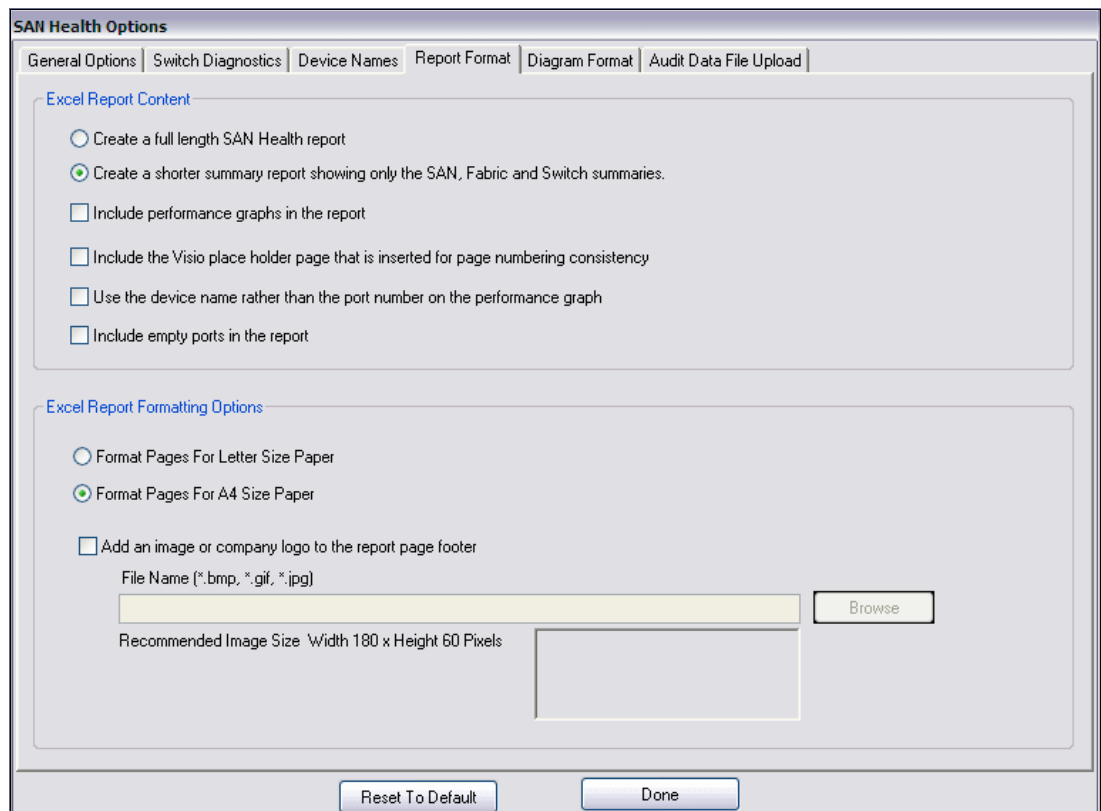


Figure 9-2 Brocade SAN Health Options window

SAN Health supports switches from manufacturers other than Brocade, such as McData and Cisco. Both the data collection tool download and the processing of files are available at no cost. You can download Microsoft Visio and Excel viewers at no cost from the Microsoft website.

Another tool, which is known as *SAN Health Professional*, is also available for download at no cost. With this tool, you can audit the reports in detail by using advanced search functions and inventory tracking. You can configure the SAN Health Diagnostics Capture tool as a Windows scheduled task.

Tip: Regardless of the method that is used, generate a fresh report at least once a month. Keep previous versions so that you can track the evolution of your SAN.

IBM Spectrum Control reporting

If you have IBM Spectrum Control running in your environment, you can use it to generate reports on your SAN. For more information about how to configure and schedule IBM Spectrum Control reports, see the IBM Spectrum Control documentation:

<https://ibm.biz/BdjL7E>

Ensure that the reports that you generate include all the information that you need. Schedule the reports with a period that you can use to backtrack any changes that you make.

9.1.3 IBM Spectrum Virtualize documentation

You can back up the configuration data for an IBM Spectrum Virtualize system after preliminary tasks are completed. Configuration data for the system provides information about your system and the objects that are defined in it.

Before you back up your configuration data, the following prerequisites must be met:

- ▶ No independent operations that change the configuration for the system can be running while the **backup** command is running.
- ▶ No object name can begin with an underscore character (_).

Note: The system automatically creates a backup of the configuration data each day at 1 AM. This backup is known as a *cron backup* and is written on the configuration node to `/dumps/svc.config.cron.xml_serial#>`.

Use these instructions to generate a manual backup at any time:

1. Issue the **svconfig backup** command to back up your configuration:

The command displays messages similar to the ones in Example 9-1.

Example 9-1 Sample svconfig backup command output

```
CMMVC6112W io_grp io_grp1 has a default name
CMMVC6112W io_grp io_grp2 has a default name
CMMVC6112W mdisk mdisk14 ...
CMMVC6112W node node1 ...
CMMVC6112W node node2 ...
.....
```

The **svconfig backup** command creates three files that provide information about the backup process and the configuration. These files are created in the `/dumps` directory of the configuration node. Table 9-1 describes the three files that are created by the backup process.

Table 9-1 Files created by the backup process

File name	Description
<code>svc.config.backup.xml_<serial#></code>	Contains your configuration data.
<code>svc.config.backup.sh_<serial#></code>	Contains the names of the commands that were issued to create the backup of the system.
<code>svc.config.backup.log_<serial#></code>	Contains details about the backup, including any reported errors or warnings.

2. Check that the **svconfig backup** command completes successfully, and examine the command output for any warnings or errors. The following output is an example of the message that is displayed when the backup process is successful:

```
CMMVC6155I SVCCONFIG processing completed successfully
```

3. If the process fails, resolve the errors and run the command again.
4. Copy the backup file from the configuration node. With MS Windows, use the PuTTY **pscp** utility. With UNIX or Linux, you can use the standard **scp** utility.

The configuration backup file is in Extensible Markup Language (XML) format and can be imported into your IBM Spectrum Virtualize documentation spreadsheet. The configuration backup file might contain too much data; for example, it contains information about each internal storage drive that is installed in the system. Importing the file into your IBM Spectrum Virtualize documentation spreadsheet might make it unreadable.

In this case, consider collecting the output of specific commands. At a minimum, you should collect the output of the following commands:

- ▶ **svcinfo lsfabric**
- ▶ **svcinfo lssystem**
- ▶ **svcinfo lsmdisk**
- ▶ **svcinfo lsmdiskgrp**
- ▶ **svcinfo lsvdisk**
- ▶ **svcinfo lshost**
- ▶ **svcinfo lshostvdiskmap**

Import the commands into a spreadsheet, preferably with each command output on a separate sheet.

One way to automate either task is to first create a batch file (Windows) or shell script (UNIX or Linux) that collects and stores this information. For more information, see 9.8, “I/O Throttling” on page 346. Then, use spreadsheet macros to import the collected data into your IBM Spectrum Virtualize documentation spreadsheet.

When you are gathering IBM Spectrum Virtualize information, consider the following preferred practices:

- ▶ If you are collecting the output of specific commands, use the **-delim** option of these commands to make their output delimited by a character other than tab, such as comma, colon, or exclamation mark. You can import the temporary files into your spreadsheet in comma-separated values (CSV) format, specifying the same delimiter.

Note: It is important to use a delimiter that is not already part of the output of the command. Commas can be used if the output is a particular type of list. Colons might be used for special fields, such as IPv6 addresses, WWPNs, or iSCSI names.

- ▶ If you are collecting the output of specific commands, save the output to temporary files. To make your spreadsheet macros simpler, you might want to preprocess the temporary files and remove any “garbage” or undesired lines or columns. With UNIX or Linux, you can use text edition commands such as **grep**, **sed**, and **awk**. Freeware software is available for Windows with the same commands, or you can use any batch text editor tool.

The objective is to fully automate this procedure so you can schedule it to run automatically on a regular basis. Make the resulting spreadsheet easy to consult and have it contain only the information that you use frequently. The automated collection and storage of configuration and support data (which is typically more extensive and difficult to use) are described in 9.1.7, “Automated support data collection” on page 329.

9.1.4 Storage documentation

Fully allocate all of the available space in the storage controllers that you use as back-end to the IBM Spectrum Virtualize solution. This way, you can perform all your Disk Storage Management tasks by using IBM Spectrum Virtualize.

You must generate only documentation of your back-end storage controllers manually one time after configuration. Then, you can update the documentation when these controllers receive hardware or code updates. As such, there is little point to automating this back-end storage controller documentation. The same applies to the IBM Spectrum Virtualize internal disk drives and enclosures.

However, if you use split controllers, this option might not be the best one. The portion of your storage controllers that is used outside the IBM Spectrum Virtualize solution might have its configuration changed frequently. In this case, see your back-end storage controller documentation for more information about how to gather and store the information that you need.

9.1.5 Technical support information

If you must open a technical support incident for your storage and SAN components, create and keep available a spreadsheet with all relevant information for all storage administrators. This spreadsheet should include the following information:

- ▶ Hardware information:
 - Vendor, machine and model number, serial number (example: IBM 2145-CF8 S/N 75ABCDE)
 - Configuration, if applicable
 - Current code level
- ▶ Physical location:
 - Datacenter, including the complete street address and phone number
 - Equipment physical location (room number, floor, tile location, and rack number)
 - Vendor’s security access information or procedure, if applicable
 - Onsite person’s contact name and phone or page number

- ▶ Support contract information:
 - Vendor contact phone numbers and website
 - Customer's contact name and phone or page number
 - User ID to the support website, if applicable
 - Do not store the password in the spreadsheet under any circumstances
 - Support contract number and expiration date

By keeping this data on a spreadsheet, storage administrators have all the information that they need to complete a web support request form or to provide to a vendor's call support representative. Typically, you are asked first for a brief description of the problem and then asked later for a detailed description and support data collection.

9.1.6 Tracking incident and change tickets

If your organization uses an incident and change management and tracking tool (such as IBM Tivoli Service Request Manager®), you or the storage administration team might need to develop proficiency in its use for several reasons:

- ▶ If your storage and SAN equipment are not configured to send SNMP traps to this incident management tool, manually open incidents whenever an error is detected.
- ▶ Disk storage allocation and deallocation and SAN zoning configuration modifications should be handled under properly submitted and approved change tickets.
- ▶ If you are handling a problem yourself, or calling your vendor's technical support desk, you might need to produce a list of the changes that you recently implemented in your SAN or that occurred since the documentation reports were last produced or updated.

When you use incident and change management tracking tools, adhere to the following guidelines for IBM Spectrum Virtualize and SAN Storage Administration:

- ▶ Whenever possible, configure your storage and SAN equipment to send SNMP traps to the incident monitoring tool so that an incident ticket is automatically opened and the proper alert notifications are sent. If you do not use a monitoring tool in your environment, you might want to configure e-mail alerts that are automatically sent to the mobile phones or pagers of the storage administrators on duty or on call.
- ▶ Discuss within your organization the risk classification that a storage allocation or deallocation change ticket is to have. These activities are typically safe and nondisruptive to other services and applications when properly handled.

However, they have the potential to cause collateral damage if a human error or an unexpected failure occurs during implementation. Your organization might decide to assume more costs with overtime and limit such activities to off-business hours, weekends, or maintenance windows if they assess that the risks to other critical applications are too high.

- ▶ Use templates for your most common change tickets, such as storage allocation or SAN zoning modification, to facilitate and speed up their submission.
- ▶ Do not open change tickets in advance to replace failed, redundant, hot-pluggable parts, such as disk drive modules (DDMs) in storage controllers with hot spares, or SFPs in SAN switches or servers with path redundancy. Typically, these fixes do not change anything in your SAN storage topology or configuration, and do not cause any more service disruption or degradation than you already had when the part failed.

Handle these fixes within the associated incident ticket because it might take longer to replace the part if you need to submit, schedule, and approve a non-emergency change ticket.

An exception is if you must interrupt more servers or applications to replace the part. In this case, you must schedule the activity and coordinate support groups. Use good judgment and avoid unnecessary exposure and delays.

- ▶ Keep handy the procedures to generate reports of the latest incidents and implemented changes in your SAN Storage environment. Typically, you do not need to periodically generate these reports because your organization probably already has a Problem and Change Management group that runs such reports for trend analysis purposes.

9.1.7 Automated support data collection

In addition to the easier-to-use documentation of your IBM Spectrum Virtualize and SAN Storage environment, collect and store for some time the configuration files and technical support data collection for all your SAN equipment.

For IBM Spectrum Virtualize, this information includes **snap** data. For other equipment, see the related documentation for more information about how to gather and store the support data that you might need.

You can create procedures that automatically create and store this data on scheduled dates, delete old data, or transfer the data to tape.

9.1.8 Subscribing to IBM Spectrum Virtualize support

Subscribing to IBM Spectrum Virtualize support is probably the most overlooked practice in IT administration, and yet it is the most efficient way to stay ahead of problems. With this subscription, you can receive notifications about potential threats before they can reach you and cause severe service outages.

To subscribe to this support and receive support alerts and notifications for your products, see the following IBM Support website:

<http://www.ibm.com/support>

If you do not have an IBM ID, create an ID.

You can subscribe to receive information from each vendor of storage and SAN equipment from the IBM website. You can often quickly determine whether an alert or notification is applicable to your SAN storage. Therefore, open them when you receive them and keep them in a folder of your mailbox.

9.2 Storage management users

Almost all organizations have IT security policies that enforce the use of password-protected user IDs when their IT assets and tools are used. However, some storage administrators still use generic, shared IDs, such as `superuser`, `admin`, or `root`, in their management consoles to perform their tasks. They might even use a factory-set default password. Their justification might be a lack of time, forgetfulness, or the fact that their SAN equipment does not support the organization's authentication tool.

SAN storage equipment management consoles often do not provide access to stored data, but one can easily shut down a shared storage controller and any number of critical applications along with it. Moreover, having individual user IDs set for your storage administrators allows much better backtracking of your modifications if you must analyze your logs.

IBM Spectrum Virtualize supports the following authentication methods:

- ▶ Local authentication by using password
- ▶ Local authentication by using SSH keys
- ▶ Remote authentication using LDAP
- ▶ Remote authentication using Tivoli

Regardless of the authentication method you choose, complete the following tasks:

- ▶ Create individual user IDs for your Storage Administration staff. Choose user IDs that easily identify the user. Use your organization's security standards.
- ▶ Include each individual user ID into the UserGroup with only enough privileges to perform the required tasks.
- ▶ If required, create generic user IDs for your batch tasks, such as Copy Services or Monitoring. Include them in a CopyOperator or Monitor UserGroup. Do not use generic user IDs with the SecurityAdmin privilege in batch tasks.
- ▶ Create unique SSH public and private keys for each of your administrators.
- ▶ Store your superuser password in a safe location in accordance to your organization's security guidelines and use it only in emergencies.

9.3 Standard operating procedures

To simplify the SAN storage administration tasks that you use most often (such as SAN storage allocation or removal, or adding or removing a host from the SAN), create step-by-step, predefined standard procedures for them. The following sections provide guidance for keeping your IBM Spectrum Virtualize environment working correctly and reliably.

9.3.1 Allocating and deallocating volumes to hosts

When you allocate and deallocate volumes to hosts, consider the following guidelines:

- ▶ Before you allocate new volumes to a server with redundant disk paths, verify that these paths are working well and that the multipath software is free of errors. Fix any disk path errors that you find in your server before you proceed.
- ▶ When you plan for future growth of space efficient VDisks, determine whether your server's operating system supports the particular volume to be extended online. Previous AIX releases, for example, do not support online expansion of rootvg LUNs. Test the procedure in a nonproduction server first.
- ▶ Always cross-check the host LUN ID information with the vdisk_UID of the SAN Volume Controller. Do not assume that the operating system recognizes, creates, and numbers the disk devices in the same sequence or with the same numbers as you created them in the SAN Volume Controller/Storwize.
- ▶ Ensure that you delete any volume or LUN definition in the server *before* you unmap it in IBM Spectrum Virtualize. For example, in AIX, remove the hdisk from the volume group (reducevg) and delete the associated hdisk device (rmdev).

- ▶ From version 7.4 onwards, consider enabling volume protection by using `chsystem vdiskprotectionenabled yes -vdiskprotectiontime <value_in_minutes>`. Volume protection ensures that some CLI actions (most of those that either explicitly or implicitly remove host-volume mappings or delete volumes) are policed to prevent the removal of mappings to volumes or deletion of volumes that are considered *active* (the system has detected I/O activity within the specified time in minutes to the volume from any host).

Note: Volume protection cannot be overridden by the use of the `-force` flag in the affected CLI commands. Volume protection must be disabled to carry on an activity that is currently blocked.

- ▶ Ensure that you explicitly remove a volume from any volume-to-host mappings and any copy services relationship to which it belongs *before* you delete it.

Attention: You must avoid the use of the `-force` parameter in `rmvdisk`.

If you issue the `svctask rmvdisk` command and it still has pending mappings, IBM Spectrum Virtualize prompts you to confirm and is a hint that you might have done something incorrectly.

- ▶ When you are deallocating volumes, plan for an interval between unmapping them to hosts (`rmvdiskhostmap`) and deleting them (`rmvdisk`). The IBM internal Storage Technical Quality Review Process (STQRP) asks for a minimum of a 48-hour period, and having at least a one business day interval so that you can perform a quick backout if you later realize you still need some data on that volume.

9.3.2 Adding and removing hosts

When you add and remove host (or hosts) in IBM Spectrum Virtualize, consider the following guidelines:

- ▶ Before you map new servers to IBM Spectrum Virtualize, verify that they are all error free. Fix any errors that you find in your server and IBM Spectrum Virtualize before you proceed. In IBM Spectrum Virtualize, pay special attention to anything inactive in the `svcinfo lsfabric` command.
- ▶ Plan for an interval between updating the zoning in each of your redundant SAN fabrics, such as at least 30 minutes. This interval allows for failover to occur and stabilize, and for you to be notified if unexpected errors occur.
- ▶ After you perform the SAN zoning from one server's HBA to IBM Spectrum Virtualize, you should list its WWPN by using the `svcinfo lshbaportcandidate` command. Use the `svcinfo lsfabric` command to certify that it was detected by the IBM Spectrum Virtualize nodes and ports that you expected. When you create the host definition in IBM Spectrum Virtualize (`svctask mkhost`), try to avoid the `-force` parameter. If you do not see the host's WWPNs, it might be necessary to scan fabric from the host. For example, use the `cfgmgr` command in AIX.

9.4 IBM Spectrum Virtualize code update

Because IBM Spectrum Virtualize might be at the core of your disk and SAN storage environment, its update requires planning, preparation, and verification. However, with the appropriate precautions, an update can be conducted easily and transparently to your servers and applications. This section highlights applicable guidelines for IBM Spectrum Virtualize update. Before upgrading your IBM Spectrum Virtualize to V8.1, you must enable the Host Offloading Throttle, for more details check the following website:

<http://www.ibm.com/support/docview.wss?uid=ssg1S1010697>

Most of the following sections explain how to prepare for the IBM Spectrum Virtualize update. The last two sections present version-independent guidelines to update the IBM Spectrum Virtualize system and disk drive.

9.4.1 Current and target IBM Spectrum Virtualize code level

First, determine your current and target IBM Spectrum Virtualize code level. Log in to your IBM Spectrum Virtualize web-based GUI and find the current version. The specific tab to use varies depending on the version itself. Alternatively, if you are using the CLI, run the `svcinfo 1ssystem` command.

IBM Spectrum Virtualize code levels are specified by four digits in the format:

- ▶ V is the major version number
- ▶ R is the release level
- ▶ M is the modification level
- ▶ F is the fix level

As target, use the latest general availability (GA) IBM Spectrum Virtualize release unless you have a specific reason not to update:

- ▶ The specific version of an application or other component of your SAN Storage environment has a known problem or limitation.
- ▶ The latest IBM Spectrum Virtualize GA release is not yet cross-certified as compatible with another key component of your SAN storage environment.
- ▶ Your organization has mitigating internal policies, such as the use of the “latest minus 1” release, or prompting for “seasoning” in the field before implementation.

For more information, see the following websites:

- ▶ Storwize V7000 Concurrent Compatibility and Code Cross-Reference:
<http://www.ibm.com/support/docview.wss?uid=ssg1S1003705>
- ▶ SAN Volume Controller Concurrent Compatibility and Code Cross-Reference:
<http://www.ibm.com/support/docview.wss?uid=ssg1S1001707>

9.4.2 IBM Spectrum Virtualize Upgrade Test Utility

Install and run the latest IBM Spectrum Virtualize Upgrade Test Utility before you update the IBM Spectrum Virtualize code. To download the Upgrade Test Utility, see this website:

<https://www.ibm.com/support/docview.wss?uid=ssg1S4000585>

This tool verifies the health of your IBM Spectrum Virtualize solution for the update process. It also checks for unfixed errors, degraded MDisks, inactive fabric connections, configuration conflicts, hardware compatibility, disk drives firmware, and many other issues that might otherwise require cross-checking a series of command outputs.

Note: The Upgrade Test Utility does not log in storage controllers or SAN switches. Instead, it reports the status of the connections of IBM Spectrum Virtualize to these devices. It is the users' responsibility to check these components for internal errors.

You can use the GUI or the CLI to install and run the Upgrade Test Utility.

Figure 9-3 shows the Storwize version 8.1.0.1 GUI window that is used to install and run the Upgrade Test Utility. It is uploaded and installed like any other software update. The Test Only option is only available from version 7.6 onwards.

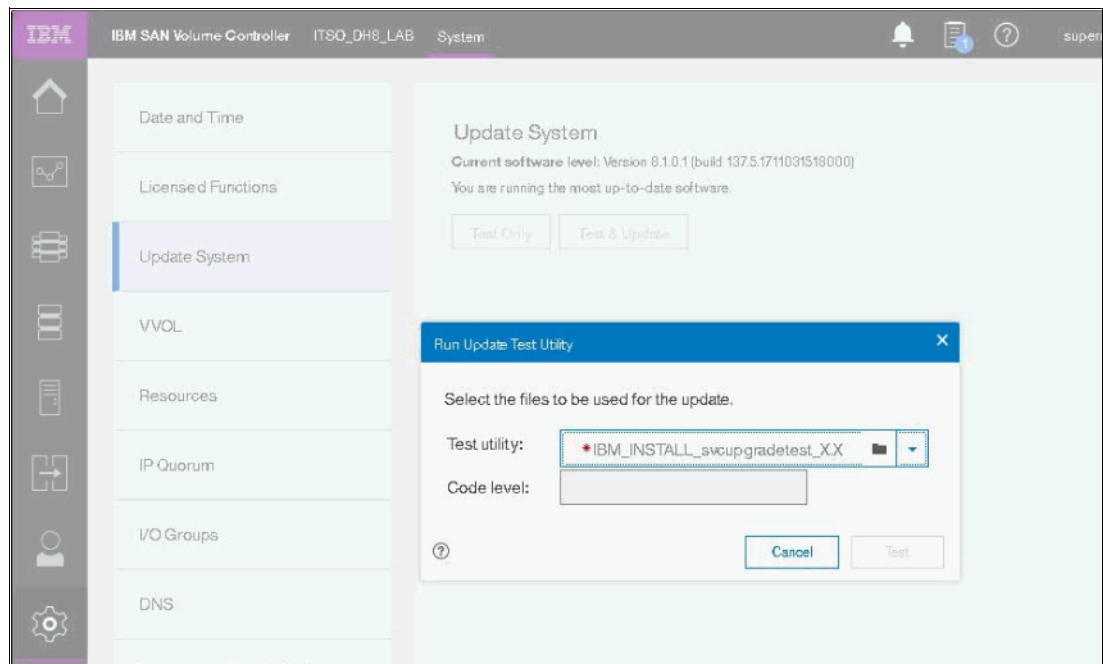


Figure 9-3 IBM Spectrum Virtualize Upgrade Test Utility installation using the GUI

Example 9-2 shows how to install and run Upgrade Test Utility in the CLI. In this case, the Upgrade Test Utility found warnings and errors and indicates recommended actions on a Storwize V7000.

Example 9-2 Upgrade test by using the CLI

```
IBM_Storwize:Spectrum_Virtualize_Cluster:superuser>svctask applysoftware -file
IBM_INSTALL_svcupgradetest_20.9
CMMVC9001I The package installed successfully.
IBM_Storwize:Spectrum_Virtualize_Cluster:superuser>svcupgradetest -v 7.7.1.2 -d
svcupgradetest version 20.9
```

Please wait, the test may take several minutes to complete.

***** Warning found *****

The upgrade utility has detected that email notifications for error

reporting have either not been configured or that the Call Home function has not been configured to automatically open a problem record. This may be caused by an invalid or missing email address. Please review the following technote to understand the benefits of enabling call home and inventory emails. <http://www.ibm.com/support/docview.wss?rs=591&uid=ssg1S1004537>

***** Warning found *****

This tool has found the internal disks of this system are not running the recommended firmware versions. Details follow:

Model	Latest FW	Current FW	Drive	Info
ST9300603SS	B53E	B53B	Drive 2	in slot 19 in enclosure 1
			Drive 3	in slot 18 in enclosure 1
HK230041S	2936	291E	Drive 0	in slot 24 in enclosure 1
			Drive 1	in slot 23 in enclosure 1

We recommend that you upgrade the drive microcode at an appropriate time. If you believe you are running the latest version of microcode, then check for a later version of this tool. You do not need to upgrade the drive firmware before starting the software upgrade.

***** Error found *****

The system identified that one or more drives in the system are running microcode with a known issue.

The following flashes are appropriate for your drives:

* <http://www.ibm.com/support/docview.wss?rs=591&uid=ssg1S1004327>

The following drives are affected by this issue:

0, 1

Results of running svcupgradetest:

=====

The tool has found 1 errors and 2 warnings.

9.4.3 IBM Spectrum Virtualize hardware considerations

Before you start the update process, always check whether your IBM Spectrum Virtualize hardware and target code level are compatible.

If part or all your current hardware is not supported at the target code level that you want to update to, replace the unsupported hardware with newer models before you update to the target code level.

Conversely, if you plan to add or replace hardware with new models to an existing cluster, you might have to update your IBM Spectrum Virtualize code first.

9.4.4 Attached hosts preparation

If the appropriate precautions are taken, the IBM Spectrum Virtualize update is not apparent to the attached servers and their applications. The automated update procedure updates one IBM Spectrum Virtualize node at a time, while the other node in the I/O group covers for its designated volumes.

However, to ensure that this feature works, the *failover capability* of your multipath software must be working properly. This capability can be mitigated by enabling NPIV if your current code level supports this function. For more information about NPIV, see Chapter 6, “Hosts” on page 241.

Before you start IBM Spectrum Virtualize update preparation, check the following items for every server that is attached to IBM Spectrum Virtualize that you update:

- ▶ The operating system type, version, and maintenance or fix level
- ▶ The make, model, and microcode version of the HBAs
- ▶ The multipath software type, version, and error log

For information about troubleshooting, see these websites (require an IBMid):

- ▶ The IBM Support page on SAN Volume Controller Troubleshooting:
<https://ibm.biz/BdjL7z>
- ▶ The IBM Support page on Storwize V7000 Troubleshooting:
<https://ibm.biz/BdjL7P>

Fix every problem or “suspect” that you find with the disk path failover capability. Because a typical IBM Spectrum virtualize environment has several dozens of servers to a few hundred servers attached to it, a spreadsheet might help you with the Attached Hosts Preparation tracking process. If you have some host virtualization, such as VMware ESX, AIX LPARs, IBM VIOS, or Solaris containers in your environment, verify the redundancy and failover capability in these virtualization layers.

9.4.5 Storage controllers preparation

As critical as with the attached hosts, the attached storage controllers must correctly handle the failover of MDisk paths. Therefore, they must be running supported microcode versions and their own SAN paths to IBM Spectrum Virtualize must be free of errors.

9.4.6 SAN fabrics preparation

If you are using symmetrical, redundant, independent SAN fabrics, preparing these fabrics for an IBM Spectrum Virtualize update can be safer than hosts or storage controllers. This statement is true assuming that you follow the guideline of a 30-minute minimum interval between the modifications that you perform in one fabric to the next. Even if an unexpected error brings down your entire SAN fabric, the IBM Spectrum Virtualize environment must continue working through the other fabric and your applications must remain unaffected.

Because you are updating your IBM Spectrum Virtualize, also update your SAN switches code to the latest supported level. Start with your principal core switch or director, continue by updating the other core switches, and update the edge switches last. Update one entire fabric (all switches) before you move to the next one so that any problem you might encounter affects only the first fabric. Begin your other fabric update only after you verify that the first fabric update has no problems.

If you are not running symmetrical, redundant independent SAN fabrics, fix this problem as a high priority because it represents a single point of failure (SPOF).

9.4.7 SAN components update sequence

Check the compatibility of your target IBM Spectrum Virtualize code level with all components of your SAN storage environment (SAN switches, storage controllers, server HBAs) and its attached servers (operating systems and eventually, applications).

Applications often certify only the operating system that they run under and leave to the operating system provider the task of certifying its compatibility with attached components (such as SAN storage). However, various applications might use special hardware features or raw devices and certify the attached SAN storage. If you have this situation, consult the compatibility matrix for your application to certify that your IBM Spectrum Virtualize target code level is compatible.

The IBM Spectrum Virtualize Supported Hardware List provides the complete information for using your IBM Spectrum Virtualize SAN storage environment components with the current and target code level. For links to the Supported Hardware List, Device Driver, Firmware, and Recommended Software Levels for different products and different code levels, see the following resources:

- ▶ Support Information for SAN Volume Controller:
<http://www.ibm.com/support/docview.wss?uid=ssg1S1003658>
- ▶ Support Information for IBM Storwize V7000:
<http://www.ibm.com/support/docview.wss?uid=ssg1S1003741>

By cross-checking the version of IBM Spectrum Virtualize is compatible with the versions of your SAN environment components, you can determine which one to update first. By checking a component's update path, you can determine whether that component requires a multistep update.

If you are not making major version or multistep updates in any components, the following update order is less prone to eventual problems:

1. SAN switches or directors
2. Storage controllers
3. Servers HBAs microcodes and multipath software
4. IBM Spectrum Virtualize system
5. IBM Spectrum Virtualize internal disk drives

Attention: Do *not* update two components of your IBM Spectrum Virtualize SAN storage environment simultaneously, such as the IBM Spectrum Virtualize system and one storage controller. This caution is true even if you intend to do it with your system offline. An update of this type can lead to unpredictable results, and an unexpected problem is much more difficult to debug.

9.4.8 IBM Spectrum Virtualize participating in Metro Mirror or Global Mirror

When you update an IBM Spectrum Virtualize system that participates in an intercluster Copy Services relationship, do *not* update both clusters in the relationship simultaneously. This situation is not verified or monitored by the automatic update process, and might lead to a loss of synchronization and unavailability.

You must successfully finish the update in one cluster before you start the next one. Try to update the next cluster as soon as possible to the same code level as the first one. Avoid running them with different code levels for extended periods.

Note: When you are updating from version 7.1 or earlier to version 7.2 or later, you *must* stop all Global Mirror (GM) relationships that have their secondary volume on the system that is being updated before starting the update process. This requirement is because of performance improvements in GM code in version 7.2. You can restart these relationships after the update process completes. Other remote copy relationships, such as Metro Mirror (MM) or Global Mirror with Change Volumes (GMCV), do not have to be stopped.

9.4.9 IBM Spectrum Virtualize update

Adhere to the following version-independent guidelines for your IBM Spectrum Virtualize code update:

- ▶ Schedule the IBM Spectrum Virtualize code update for a low I/O activity time. The update process puts one node at a time offline. It also disables the write cache in the I/O group that node belongs to until both nodes are updated. Therefore, with lower I/O, you are less likely to notice performance degradation during the update.
- ▶ Never power off, reboot, or reset an IBM Spectrum Virtualize node during code update unless you are instructed to do so by IBM Support. Typically, if the update process encounters a problem and fails, it backs out.
- ▶ Check whether you are running a web browser type and version that are supported by the IBM Spectrum Virtualize target code level on every computer that you intend to use to manage your IBM Spectrum Virtualize.
- ▶ If you are planning for a major IBM Spectrum Virtualize version update, update your current version to its latest fix level before you run the major update.

9.4.10 IBM Spectrum Virtualize disk drive update

Update of disk drive firmware is concurrent whether it is HDD or SSD. However, with SSD, the FPGA level can also be updated. Update of FPGA is *not* concurrent, so all I/Os to the SSDs must be stopped before the update. It is not a problem if SSDs are not yet configured. However, if you have any SSD arrays in storage pools, you must remove SSD MDisks from the pools before the update.

This task can be challenging because removing MDisks from storage pool means migrating all extents from these MDisks to the remaining MDisk in the pool. You cannot remove SSD MDisks from the pool if there is no space left on the remaining MDisks. In such a situation, one option is to migrate some volumes to other storage pools to free enough extents so the SSD MDisk can be removed.

Important: More precaution must be taken if you are updating the FPGA on SSD in a 2-tiers hybrid storage pool with Easy Tier running. If the Easy Tier setting on the storage pool has value of auto, Easy Tier switches off after SSD MDisks are removed from that pool, which means it loses all its historical data.

After SSD MDisks are added back to this pool, Easy Tier must start its analysis from the beginning. If you want to avoid such a situation, switch the Easy Tier setting on the storage pool to on. This setting ensures that Easy Tier retains its data after SSD removal.

9.5 SAN modifications

When you administer shared storage environments, human error can occur when a failure is fixed or a change is made that affects one or more servers or applications. That error can then affect other servers or applications because appropriate precautions were not taken.

Human error can include the following examples:

- ▶ Disrupting or disabling the working disk paths of a server while trying to fix failed ones.
- ▶ Disrupting a neighbor SAN switch port while inserting or pulling out an FC cable or SFP.
- ▶ Disabling or removing the working part in a redundant set instead of the failed one.
- ▶ Making modifications that affect both parts of a redundant set without an interval that allows for automatic failover during unexpected problems.

Adhere to the following guidelines to perform these actions with assurance:

- ▶ Uniquely and correctly identify the components of your SAN.
- ▶ Use the proper failover commands to disable only the failed parts.
- ▶ Understand which modifications are necessarily disruptive, and which can be performed online with little or no performance degradation.

9.5.1 Cross-referencing HBA WWPNs

With the WWPn of an HBA, you can uniquely identify one server in the SAN. If a server's name is changed at the operating system level and not at the IBM Spectrum Virtualize host definitions, it continues to access its previously mapped volumes exactly because the WWPn of the HBA did not change.

Alternatively, if the HBA of a server is removed and installed in a second server and the first server's SAN zones and IBM Spectrum Virtualize host definitions are not updated, the second server can access volumes that it probably should not access.

Complete the following steps to cross-reference HBA WWPns:

1. In your server, verify the WWPns of the HBAs that are used for disk access. Typically, you can complete this task by using the SAN disk multipath software of your server. If you are using SDDPCM, run the `pcmpath query WWPn` command to see output similar to what is shown in Example 9-3.

Example 9-3 Output of the pcmpath query WWPn command

```
[root@nybixtdb02]> pcmpath query wwpn
Adapter Name PortWWN
fscsi0       10000000C925F5B0
fscsi1       10000000C9266FD1
```

If you are using server virtualization, verify the WWPns in the server that is attached to the SAN, such as AIX VIO or VMware ESX.

2. Cross-reference with the output of the IBM Spectrum Virtualize `lshost <hostname>` command, as shown in Example 9-4.

Example 9-4 Output of the `lshost <hostname>` command

```
IBM_2145:svccf8:admin>svcinfolshost NYBIXTDB02
id 0
name NYBIXTDB02
port_count 2
type generic
mask 1111
iogrp_count 1
WWPN 1000000C925F5B0
node_logged_in_count 2
state active
WWPN 1000000C9266FD1
node_logged_in_count 2
state active
IBM_2145:svccf8:admin>
```

3. If necessary, cross-reference information with your SAN switches, as shown in Example 9-5. (In Brocade, switches use `nodefind <WWPN>`.)

Example 9-5 Cross-referencing information with SAN switches

```
b1g32sw1_B64:admin> nodefind 10:00:00:00:C9:25:F5:B0
Local:
Type Pid   COS      PortName                               NodeName                               SCR
N   401000;  2,3;10:00:00:00:C9:25:F5:B0; 20:00:00:00:C9:25:F5:B0; 3
Fabric Port Name: 20:10:00:05:1e:04:16:a9
Permanent Port Name: 10:00:00:00:C9:25:F5:B0
Device type: Physical Unknown(initiator/target)
Port Index: 16
Share Area: No
Device Shared in Other AD: No
Redirect: No
Partial: No
Aliases: nybixtdb02_fcs0
b32sw1_B64:admin>
```

For storage allocation requests that are submitted by the server support team or application support team to the storage administration team, always include the server's HBA WWPNs to which the new LUNs or volumes are supposed to be mapped. For example, a server might use separate HBAs for disk and tape access, or distribute its mapped LUNs across different HBAs for performance. You cannot assume that any new volume is supposed to be mapped to every WWPN that server logged in the SAN.

If your organization uses a change management tracking tool, perform all your SAN storage allocations under approved change tickets with the servers' WWPNs listed in the Description and Implementation sessions.

9.5.2 Cross-referencing LUN IDs

Always cross-reference the IBM Spectrum Virtualize `vdisk_UID` with the server LUN ID before you perform any modifications that involve IBM Spectrum Virtualize volumes. Example 9-6 shows an AIX server that is running SDDPCM. The SAN Volume Controller `vdisk_name` has no relation to the AIX device name. Also, the first SAN LUN mapped to the server (`SCSI_id=0`) shows up as `hdisk4` in the server because it had four internal disks (`hdisk0 - hdisk3`).

Example 9-6 Results of running the `lshostvdiskmap` command

```
IBM_2145:svccf8:admin>lshostvdiskmap NYBIXTDB03
id name          SCSI_id vdisk_id vdisk_name      vdisk_UID
0 NYBIXTDB03 0      0      NYBIXTDB03_T01 60050768018205E12000000000000000
IBM_2145:svccf8:admin>
```

```
root@nybixtdb03:./> pcmpath query device
Total Dual Active and Active/Asymmetric Devices : 1
DEV#: 4  DEVICE NAME: hdisk4  TYPE: 2145  ALGORITHM: Load Balance
SERIAL: 60050768018205E12000000000000000
=====
Path#      Adapter/Path Name      State      Mode      Select      Errors
0*         fscsi0/path0           OPEN      NORMAL      7           0
1          fscsi0/path1           OPEN      NORMAL     5597        0
2*         fscsi2/path2           OPEN      NORMAL      8           0
3          fscsi2/path3           OPEN      NORMAL     5890        0
```

If your organization uses a change management tracking tool, include the `vdisk_UID` and LUN ID information in every change ticket that performs SAN storage allocation or reclaim.

Note: Because a host can have many volumes with the same `scsi_id`, always cross-reference the IBM Spectrum Virtualize volume UID with the host volume UID, and record the `scsi_id` and LUN ID of that volume.

9.5.3 HBA replacement

Replacing a failed HBA is a fairly trivial and safe operation if it is performed correctly. However, more precautions are required if your server has redundant HBAs and its hardware permits you to “hot” replace it (with the server still running).

Complete the following steps to replace a failed HBA and retain the good HBA:

1. In your server, using the multipath software, identify the failed HBA and record its WWPNs. For more information, see 9.5.1, “Cross-referencing HBA WWPNs” on page 338. Then, place this HBA and its associated paths offline, gracefully if possible. This approach is important so that the multipath software stops trying to recover it. Your server might even show a degraded performance while you perform this task.
2. Some HBAs have a label that shows the WWPNs. If you have this type of label, record the WWPNs before you install the new HBA in the server.
3. If your server does not support HBA hot-swap, power off your system, replace the HBA, connect the used FC cable into the new HBA, and power on the system.

If your server does support hot-swap, follow the appropriate procedures to perform a “hot” replace of the HBA. Do *not* disable or disrupt the good HBA in the process.

4. Verify that the new HBA successfully logged in to the SAN switch. If it logged in successfully, you can see its WWPNs logged in to the SAN switch port.
Otherwise, fix this issue before you continue to the next step.
Cross-check the WWPNs that you see in the SAN switch with the one you noted in step 1, and make sure that you did not get the WWNN mistakenly.
5. In your SAN zoning configuration tool, replace the old HBA WWPNs for the new ones in every alias and zone to which they belong. Do *not* touch the other SAN fabric (the one with the good HBA) while you perform this task.
Only one alias should use each WWPN, and zones must reference this alias.
If you are using SAN port zoning (though you should not be) and you did not move the new HBA FC cable to another SAN switch port, you do not need to reconfigure zoning.
6. Verify that the new HBA's WWPNs appear in the IBM Spectrum Virtualize system by using the `lsfcportcandidate` command.
If the WWPNs of the new HBA do not appear, troubleshoot your SAN connections and zoning.
7. Add the WWPNs of this new HBA in the IBM Spectrum Virtualize host definition by using the `addhostport` command. Do not remove the old one yet. Run the `lshost <servername>` command. Then, verify that the good HBA shows as `active`, while the failed and old HBA should show as `inactive` or `offline`.
8. In software to recognize the new HBA and its associated SAN disk paths. Certify that all SAN LUNs have redundant disk paths through the good and the new HBAs.
9. Return to the IBM Spectrum Virtualize system and verify again (by using the `lshost <servername>` command) that both the good and the new HBA's WWPNs are active. In this case, you can remove the old HBA WWPNs from the host definition by using the `rmhostport` command.
Do not remove any HBA WWPNs from the host definition until you ensure that you have at least two active ones that are working correctly.

By following these steps, you avoid removing your only good HBA by mistake.

9.6 Hardware upgrades for IBM Spectrum Virtualize

The IBM Spectrum Virtualize scalability features allow significant flexibility in its configuration. As a consequence, several scenarios are possible for its growth. The following sections describe these processes:

- ▶ Adding IBM Spectrum Virtualize nodes to an existing cluster
- ▶ Upgrading IBM Spectrum Virtualize nodes in an existing cluster
- ▶ Moving to a new IBM Spectrum Virtualize cluster
- ▶ Splitting an IBM Spectrum Virtualize cluster

9.6.1 Adding IBM Spectrum Virtualize nodes to an existing cluster

If your existing IBM Spectrum Virtualize cluster is below the maximum I/O groups limit for your specific product and you intend to upgrade it, you might find yourself installing newer SAN Volume Controller nodes or Storwize control enclosures that are more powerful than your existing ones. Therefore, your cluster will have different node models in different I/O groups.

To install these newer nodes, determine whether you need to upgrade your IBM Spectrum Virtualize code level first. For more information, see 9.4.3, “IBM Spectrum Virtualize hardware considerations” on page 334.

After you install the newer nodes, you might need to redistribute your servers across the I/O groups. Consider the following points:

- ▶ Moving a server’s volume to different I/O groups can be done online because of a feature called Non-Disruptive Volume Movement (NDVM), which was introduced in version 6.4 of IBM Spectrum Virtualize. Although this process can be done without stopping the host, careful planning and preparation are advised.

Note: You cannot move a volume that is in any type of remote copy relationship.

- ▶ If each of your servers is zoned to only one I/O group, modify your SAN zoning configuration as you move its volumes to another I/O group. As best you can, balance the distribution of your servers across I/O groups according to I/O workload.
- ▶ Use the **-iogrp** parameter in the **mkhost** command to define which I/O groups of IBM Spectrum Virtualize that the new servers will use. Otherwise, IBM Spectrum Virtualize maps by default the host to all I/O groups, even if they do not exist and regardless of your zoning configuration. Example 9-7 shows this scenario and how to resolve it by using the **rmhostiogrp** and **addhostiogrp** commands.

Example 9-7 Mapping the host to I/O groups

```
IBM_2145:svccf8:admin>lshost NYBIXTDB02
id 0
name NYBIXTDB02
port_count 2
type generic
mask 1111
iogrp_count 4
WWPN 10000000C9648274
node_logged_in_count 2
state active
WWPN 10000000C96470CE
node_logged_in_count 2
state active
IBM_2145:svccf8:admin>lsiogrp
id name          node_count vdisk_count host_count
0 io_grp0        2          32          1
1 io_grp1        0          0           1
2 io_grp2        0          0           1
3 io_grp3        0          0           1
4 recovery_io_grp 0          0           0
IBM_2145:svccf8:admin>lshostiogrp NYBIXTDB02
id name
0 io_grp0
1 io_grp1
2 io_grp2
3 io_grp3
IBM_2145:svccf8:admin>rmhostiogrp -iogrp 1:2:3 NYBIXTDB02
IBM_2145:svccf8:admin>lshostiogrp NYBIXTDB02
id name
0 io_grp0
IBM_2145:svccf8:admin>lsiogrp
id name          node_count vdisk_count host_count
0 io_grp0        2          32          1
```

```

1 io_grp1      0      0      0
2 io_grp2      0      0      0
3 io_grp3      0      0      0
4 recovery_io_grp 0      0      0
IBM_2145:svccf8:admin>addhostiogrp -iogrp 3 NYBIXTDB02
IBM_2145:svccf8:admin>lshostiogrp NYBIXTDB02
id name
0 io_grp0
3 io_grp3
IBM_2145:svccf8:admin>lsiogrp
id name          node_count vdisk_count host_count
0 io_grp0         2          32          1
1 io_grp1         0          0          0
2 io_grp2         0          0          0
3 io_grp3         0          0          1
4 recovery_io_grp 0          0          0

```

- ▶ If possible, avoid setting a server to use volumes from different I/O groups that have different node types for extended periods of time. Otherwise, as this server's storage capacity grows, you might experience a performance difference between volumes from different I/O groups. This mismatch makes it difficult to identify and resolve eventual performance problems.

9.6.2 Upgrading IBM Spectrum Virtualize nodes in an existing cluster

If you are replacing the nodes of your existing SAN Volume Controller cluster with newer ones, the replacement procedure can be performed nondisruptively. The new node can assume the WWNN of the node you are replacing, which requires no changes in host configuration, SAN zoning, or multipath software. For more information about this procedure, see SAN Volume Controller at IBM Knowledge Center for your current code level:

<http://www.ibm.com/support/knowledgecenter/STPVGU>

From V7.8, IBM also offers the following Storwize node canisters upgrade option, from Storwize V7000 Gen2 to Storwize V7000 Gen2+.

The new node canister assumes the WWNN of the node you are replacing automatically, which requires no changes in host configuration, SAN zoning, or multipath software. For more information about this procedure, see IBM Knowledge Center for your product and current code level at the following Storwize V7000 website:

<https://ibm.biz/BdjGTt>

Nondisruptive node replacement uses failover capabilities to replace one node in an I/O group at a time. If a new node has a different version of IBM Spectrum Virtualize code, it installs the cluster version automatically during the node replacement procedure.

9.6.3 Moving to a new IBM Spectrum Virtualize cluster

You might have a highly populated, intensively used IBM Spectrum Virtualize cluster that you want to upgrade. You might also want to use the opportunity to overhaul your IBM Spectrum Virtualize and SAN storage environment.

Complete the following steps to replace your cluster entirely with a newer, bigger, and more powerful one:

1. Install your new IBM Spectrum Virtualize cluster.
2. Create a replica of your data in your new cluster.
3. Migrate your servers to the new IBM Spectrum Virtualize cluster when convenient.

If your servers can tolerate a brief, scheduled outage to switch from one IBM Spectrum Virtualize cluster to another, you can use the IBM Spectrum Virtualize remote copy services (Metro Mirror or Global Mirror) to create your data replicas, following these steps:

1. Select a host that you want to move to the new IBM Spectrum Virtualize cluster and find all the old volumes you must move.
2. Zone your host to the new IBM Spectrum Virtualize cluster.
3. Create remote copy relationships from the old volumes in the old IBM Spectrum Virtualize cluster to new volumes in the new IBM Spectrum Virtualize cluster.
4. Map the new volumes from the new IBM Spectrum Virtualize cluster to the host.
5. Discover new volumes on the host.
6. Stop all I/O from the host to the old volumes from the old IBM Spectrum Virtualize cluster.
7. Disconnect and remove the old volumes on the host from the old IBM Spectrum Virtualize cluster.
8. Unmap the old volumes from the old IBM Spectrum Virtualize cluster to the host.
9. Make sure remote copy relationships between old and new volumes in the old and new IBM Spectrum Virtualize cluster are synced.
10. Stop and remove remote copy relations between old and new volumes so that the target volumes in the new IBM Spectrum Virtualize cluster receive read/write access.
11. Import data from the new volumes and start your applications on the host.

If you must migrate a server online, instead, you must use host-based mirroring by completing these steps:

1. Select a host that you want to move to the new IBM Spectrum Virtualize cluster and find all the old volumes that you must move.
2. Zone your host to the new IBM Spectrum Virtualize cluster.
3. Create volumes in the new IBM Spectrum Virtualize cluster of the same size as the old volumes in the old IBM Spectrum Virtualize cluster.
4. Map the new volumes from the new IBM Spectrum Virtualize cluster to the host.
5. Discover new volumes on the host.
6. For each old volume, use host-based mirroring (such as AIX **mirrorvg**) to move your data to the corresponding new volume.
7. For each old volume, after the mirroring is complete, remove the old volume from the mirroring group.
8. Disconnect and remove the old volumes on the host from the old IBM Spectrum Virtualize cluster.
9. Unmap the old volumes from the old IBM Spectrum Virtualize cluster to the host.

This approach uses the server's computing resources (CPU, memory, and I/O) to replicate the data. It can be done online if properly planned. Before you begin, make sure it has enough spare resources.

The biggest benefit to using either approach is that they easily accommodate (if necessary) the replacement of your SAN switches or your back-end storage controllers. You can upgrade the capacity of your back-end storage controllers or replace them entirely, as you can replace your SAN switches with bigger or faster ones. However, you do need to have spare resources, such as floor space, power, cables, and storage capacity, available during the migration.

9.6.4 Splitting an IBM Spectrum Virtualize cluster

Splitting an IBM Spectrum Virtualize cluster might become a necessity if you have one or more of the following requirements:

- ▶ To grow the environment beyond the maximum number of I/O groups that a clustered system can support
- ▶ To grow the environment beyond the maximum number of attachable subsystem storage controllers
- ▶ To grow the environment beyond any other maximum system limit
- ▶ To achieve new levels of data redundancy and availability

By splitting the clustered system, you no longer have one IBM Spectrum Virtualize system that handles all I/O operations, hosts, and subsystem storage attachments. The goal is to create a second IBM Spectrum Virtualize system so that you can equally distribute the workload over the two systems.

After safely removing nodes from the existing cluster and creating a second IBM Spectrum Virtualize system, choose from the following approaches to balance the two systems:

- ▶ Attach new storage subsystems and hosts to the new system, and start putting only new workload on the new system.
- ▶ Migrate the workload onto the new system by using the approach described in 9.6.3, “Moving to a new IBM Spectrum Virtualize cluster” on page 343.

It can happen when you replace old nodes with new, more powerful ones. It can also occur in a remote partnership when more bandwidth is required on one site and spare bandwidth is on the other site.

9.7 Adding expansion enclosures

If you plan well, you can buy an IBM Spectrum Virtualize product with enough internal storage to run your business for some time. But as time passes and your environment grows, you will need to add more storage to your system.

Depending on the IBM Spectrum Virtualize product and the code level that you have installed, you can add different numbers of expansion enclosures to your system. Because all IBM Spectrum Virtualize systems were designed to make managing and maintaining them as simple as possible, adding an expansion enclosure is an easy task. However, here are some guidance and preferred practices that you should follow.

At the time of writing, the following IBM Spectrum Virtualize products only support one chain of expansion enclosures:

- ▶ Storwize V3500
- ▶ Storwize V3700
- ▶ Storwize V5010
- ▶ Storwize V5020

New expansion enclosures should be added at the bottom of the chain as long as the limit of enclosures for the product has not been reached.

These other IBM Spectrum Virtualize products support two chains of expansion enclosures:

- ▶ Storwize V5000
- ▶ Storwize V5030
- ▶ Storwize V7000 (Gen1, Gen2, Gen2+)
- ▶ Flash System V9000 (with SAS expansion option)
- ▶ SAN Volume Controller (with SAS expansion option)

As a preferred practice, the number of expansion enclosures should be balanced between both chains. This guideline means that the number of expansion enclosures in every chain cannot differ by more than one. For example, having five expansion enclosures in the first chain and only one in the second chain is incorrect.

Note: When counting the number of enclosures in a chain, remember that for Storwize V7000 Gen1 and Storwize V5000 Gen1, the control enclosure is part of the second chain of expansions.

Adding expansion enclosures is simplified because Storwize can automatically discover new expansion enclosures after the SAS cables are connected. It is possible to manage and use the new disk drives without managing the new expansion enclosures. However, unmanaged expansion enclosures are not monitored properly. This issue can lead to more difficult troubleshooting and can make problem resolution take longer. To avoid this situation, always manage newly added expansion enclosures.

Because of internal architecture and classical disk latency, it does not matter in which enclosure SAS or NL-SAS drives are placed. However, if you have some SSD drives and you want to use them in the most efficient way, place them in the control enclosure or in the first expansion enclosures in chains. This configuration ensures every I/O to SSD disk drives travel the shortest possible way through the internal SAS fabric.

Note: This configuration is even more important on Storwize V7000 Gen2 and Storwize Gen2+ because the drives in the control enclosure have double the bandwidth available compared to expansion enclosures and should be used for SSD drives if there are any in the system.

9.8 I/O Throttling

I/O Throttling is a mechanism that allows you to limit the volume of I/O processed by the storage controller at various levels to achieve QoS. The I/O rate is limited by queuing I/Os if it exceeds the preset limits. I/O Throttling is a way to achieve a better distribution of storage controller resources. V8.1 brings the possibility to set the throttling at a volume level, host, host cluster, storage pool, and then offload throttling by using the GUI. This section intends to describe some details of I/O throttling and show how to configure the feature in your system.

9.8.1 General information on I/O Throttling

This is a list of items to keep in mind when thinking about I/O Throttling:

- ▶ IOPS and BW throttles limits can be set
- ▶ Upper Bound QoS mechanism
- ▶ No minimum performance guaranteed
- ▶ Volumes, hosts, host clusters and managed disk groups can be throttled
- ▶ Queuing at microsecond granularity
- ▶ Internal I/Os are not throttled. (such as FlashCopy, cluster traffic, and so on)
- ▶ Reduces I/O bursts and smoothens I/O flow with variable delay in throttled I/Os
- ▶ Throttle limit is a per node value

9.8.2 I/O Throttling on front end I/O control

You can use throttling for a better front end I/O control, at volume, host, host cluster, and offload levels:

- ▶ In a multi tenant environment, hosts can have their own defined limits.
You can use this to allow restricted I/Os from a data mining server and a higher limit for an application server.
- ▶ An aggressive host consuming bandwidth of the controller can be limited by a throttle.
For example, a video streaming application can have a limit set to avoid consuming too much of the bandwidth.
- ▶ Restrict a group of hosts by their throttles.
For example, Department A gets more bandwidth than Department B.
- ▶ Each volume can have a throttle defined.
For example, a backup volume can have less bandwidth than a production volume.
- ▶ Offloaded I/Os.
[XCOPY/Writesame(VMware), ODX-WUT (HyperV)] can be confined by defined controller resources.

9.8.3 I/O Throttling on backend I/O control

You can also use throttling to control the backend I/O by throttling the storage pool, which can be useful in the following scenarios:

- ▶ Each storage pool can have a throttle defined.
- ▶ Both parent and child pool throttles are supported.
- ▶ Allows control of back-end I/Os from SVC.
- ▶ Useful to avoid overwhelming the backend storage.
- ▶ Useful in case of VVOLS since a VVOL gets created in a child pool. A child pool (mdiskgrp) throttle can control I/Os coming from that vvol.
- ▶ Parent and child pool throttles are independent of each other. A child pool can have higher throttle limits than its parent pool.

9.8.4 Overall benefits of using I/O Throttling

The overall benefits of using I/O Throttling is a better distribution all system resources:

- ▶ Avoids overwhelming the controller objects
- ▶ Avoids starving the external entities, *like hosts*, from their share of controller
- ▶ A scheme of distribution of controller resources that, in turn, results in better utilization of external resources such as host capacities.

With no throttling enabled, we have a scenario where Host 1 dominates the bandwidth, and after enabling the throttle, we see a much better distribution of the bandwidth among the hosts, as shown in Figure 9-4.

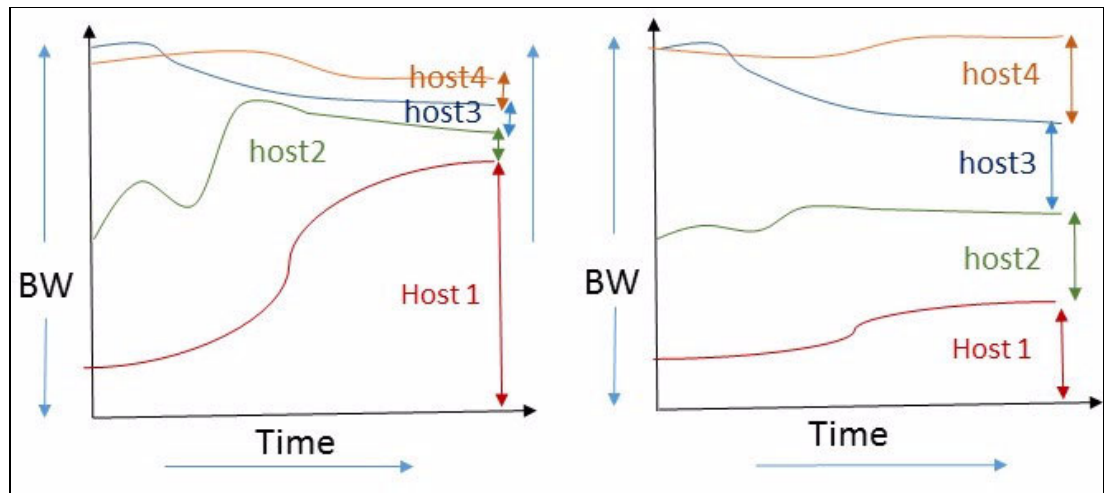


Figure 9-4 Distribution of controller resources after I/O Throttling

9.8.5 Considerations for I/O Throttling

When you are planning to use I/O Throttling there are a few points to be considered:

- ▶ The throttle cannot be defined for the host if it is part of a hostcluster which already has a hostcluster throttle.
- ▶ If the hostcluster does not have a throttle defined, its member hosts can have their individual host throttles defined.
- ▶ The storage pool throttles for child pool and parent pool work independently.
- ▶ If a volume has multiple copies then throttling would be done for the storage pool serving the primary copy. The throttling will not be applicable on the secondary pool for mirrored volumes and stretched cluster implementations.
- ▶ A host cannot be added to a hostcluster if both of them have their individual throttles defined. If just one of the host/hostcluster throttles is present, the command will succeed.
- ▶ A seeding host used for creating a hostcluster cannot have a host throttle defined for it.

Note: Throttling is only applicable at the I/Os that an SVC receives from hosts and hostclusters. The I/Os generated by SVC internally, like mirrored volume I/Os, cannot be throttled.

9.8.6 Configuring I/O Throttling using the CLI

In order to create a throttle using the CLI, you use the `mkthrottle` command, as shown in Example 9-8.

Example 9-8 Creating throttle using mkthrottle command on CLI

Syntax: (Highlighted are the new options on version 8.1)

```
mkthrottle -type [offload | vdisk | host | hostcluster | mdiskgrp]  
            [-bandwidth bandwidth_limit_in_mb]  
            [-iops iops_limit]  
            [-name throttle_name]  
            [-vdisk vdisk_id_or_name]  
            [-host host_id_or_name]  
            [-hostcluster hostcluster_id_or_name]  
            [-mdiskgrp mdiskgrp_id_or_name]
```

Usage examples:

```
IBM_2145:ITSO_DH8_LAB:superuser>mkthrottle -type host -bandwidth 100 -host  
ITSO_HOST3
```

```
IBM_2145:ITSO_DH8_LAB:superuser>mkthrottle -type hostcluster -iops 30000  
-hostcluster ITSO_HOSTCLUSTER1
```

```
IBM_2145:ITSO_DH8_LAB:superuser>mkthrottle -type mdiskgrp -iops 40000 -mdiskgrp 0
```

```
IBM_2145:ITSO_DH8_LAB:superuser>mkthrottle -type offload -bandwidth 50
```

```
IBM_2145:ITSO_DH8_LAB:superuser>mkthrottle -type vdisk -bandwidth 25 -vdisk  
volume1
```

```
IBM_2145:ITSO_DH8_LAB:superuser>lsthrottle
```

throttle_id	throttle_name	object_id	object_name	throttle_type	IOPs_limit	bandwidth_limit_MB
0	throttle0	2	ITSO_HOST3	host		100
1	throttle1	0	ITSO_HOSTCLUSTER1	hostcluster	30000	
2	throttle2	0	Pool0	mdiskgrp		40000
3	throttle3			offload		50
4	throttle4	10	volume1	vdisk		25

Note: You can change a throttle parameter by using the `chthrottle` command.

9.8.7 Configuring I/O Throttling using the GUI

The following pages shows how to configure the throttle by using the GUI.

Creating a volume throttle

To create a volume throttle, go to **Volumes** → **Volumes**, then select the desired volume, right click on it and chose *Edit Throttle*, as shown in Figure 9-5.

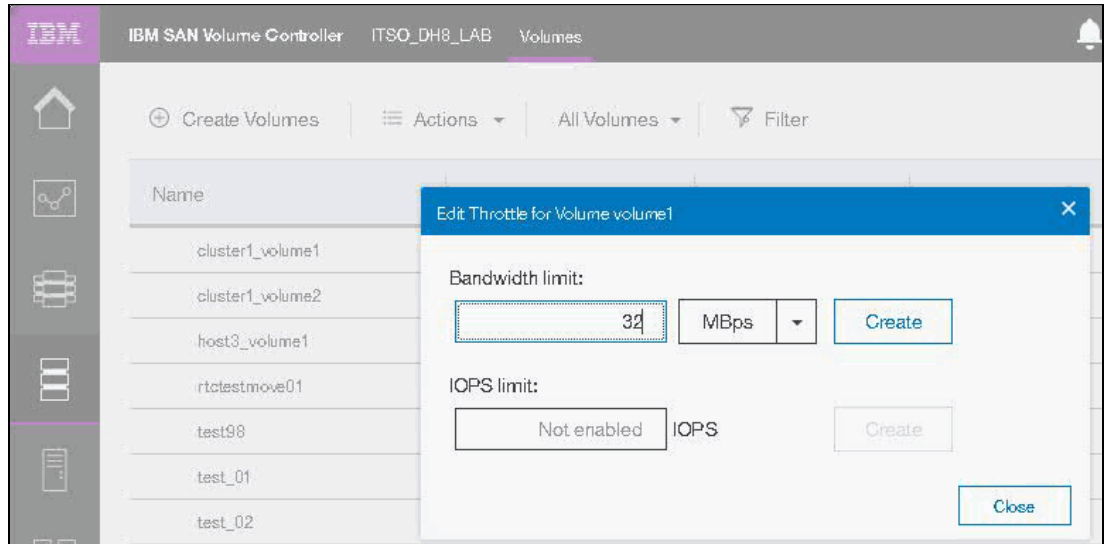


Figure 9-5 Creating a volume throttle on the GUI

Creating a host throttle

To create a host throttle, go to **Hosts** → **Hosts**, select the desired host, then right-click it and chose **Edit Throttle**, as shown in Figure 9-6.

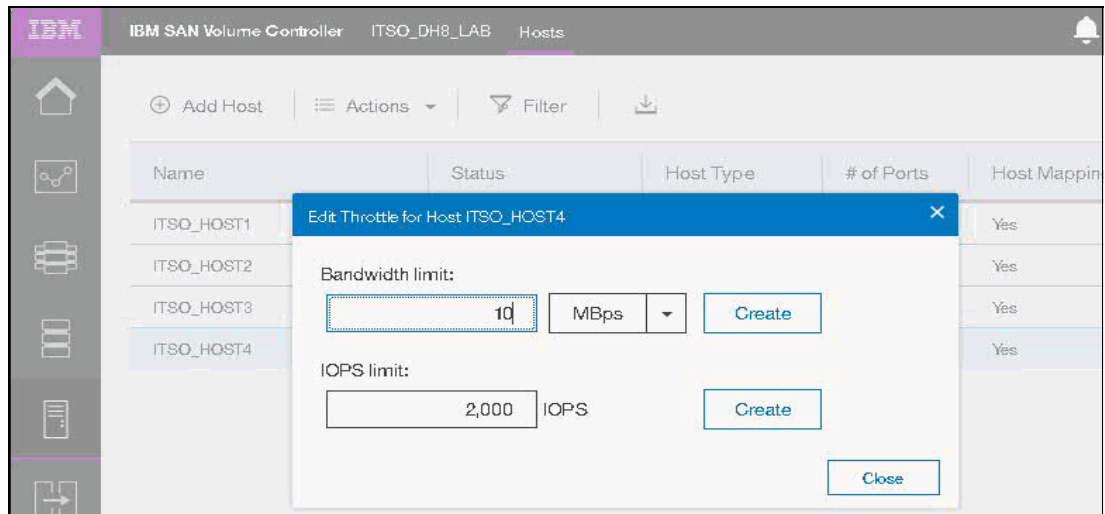


Figure 9-6 Creating a host throttle on GUI

Creating a host cluster throttle

To create a host cluster throttle, go to **Hosts** → **Host Clusters**, select the desired host cluster, then right-click it and chose **Edit Throttle**, as shown in Figure 9-7.

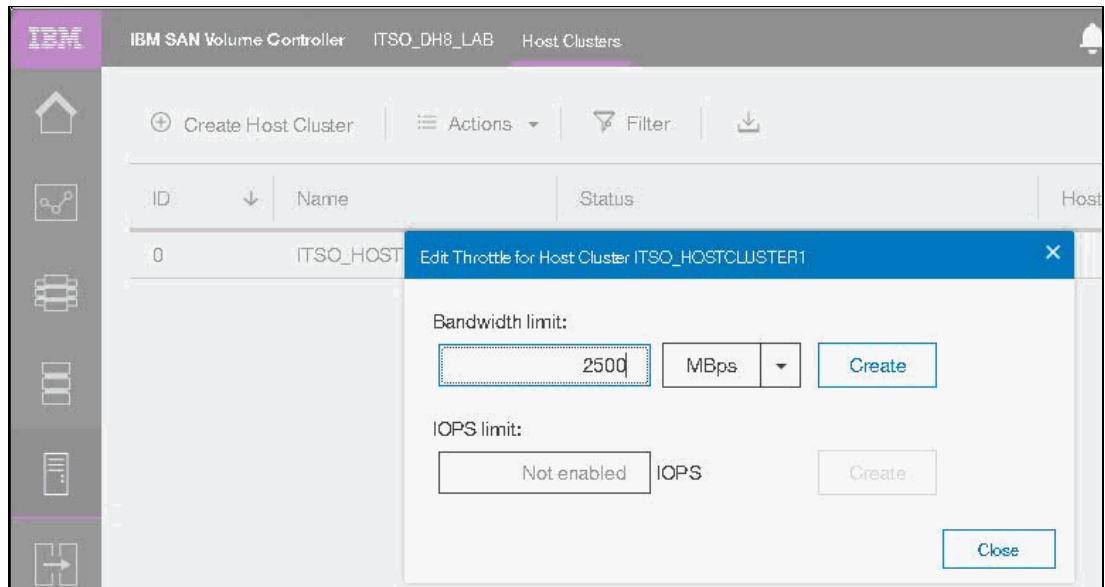


Figure 9-7 Creating a host cluster throttle on GUI

Creating a storage pool throttle

To create a storage pool throttle, go to **Pools** → **Pools**, select the desired storage pool, then right click on it and chose *Edit Throttle*, as shown in Figure 9-8.

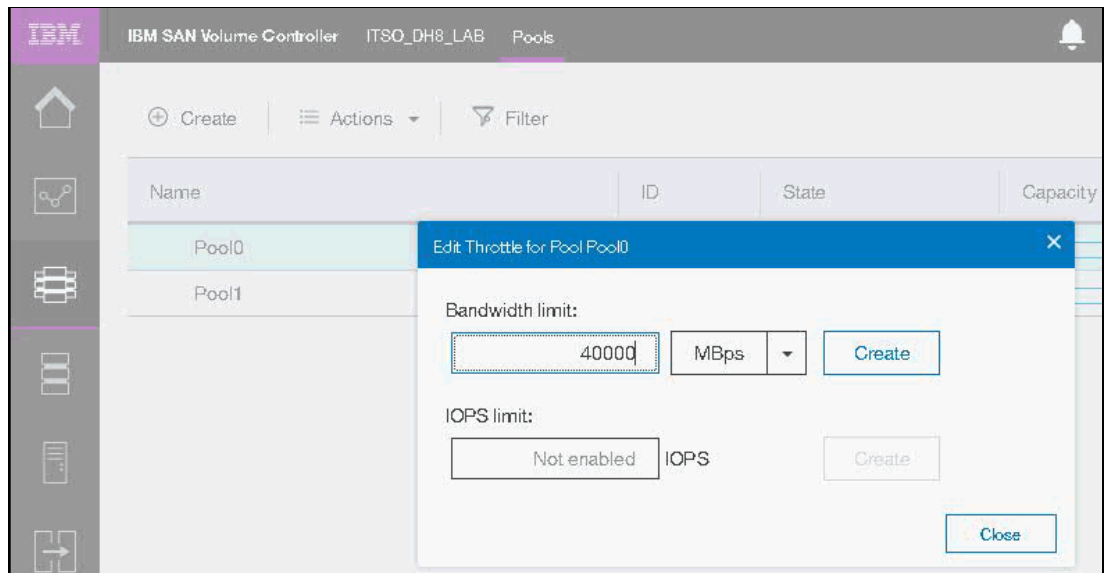


Figure 9-8 Creating a storage pool throttle on GUI

Creating an offload throttle:

To create an offload throttle, go to **Monitoring** → **System** → **Actions**, then select **Edit System Offload Throttle**, as shown in Figure 9-9.

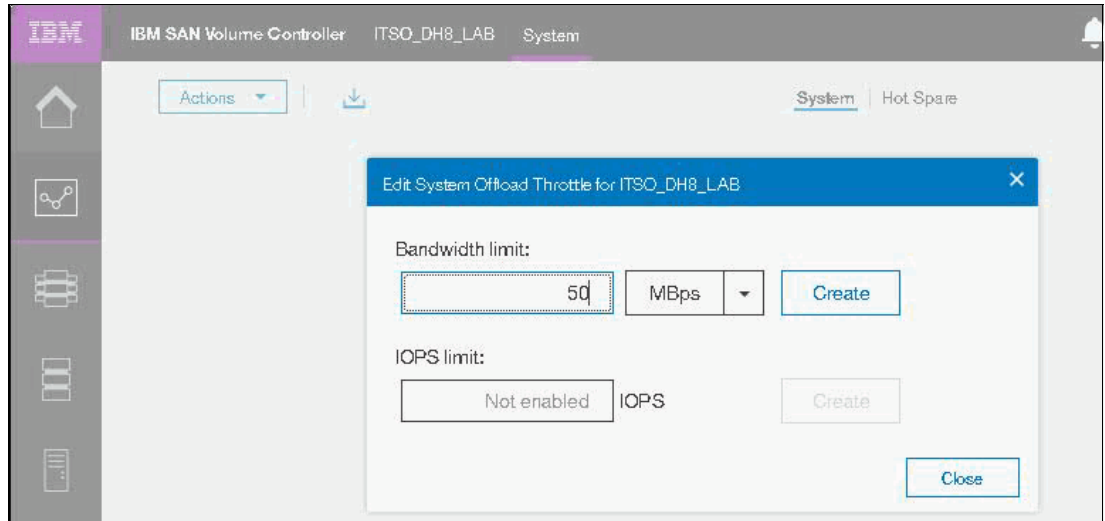


Figure 9-9 Creating system offload throttle on GUI



Troubleshooting and diagnostics

IBM Spectrum Virtualize is a robust and reliable virtualization engine that demonstrated excellent availability in the field. However, today's storage area networks (SANs), storage subsystems, and host systems are external components that might cause some events.

This chapter provides useful information to start your troubleshooting and an overview of common events that can occur in your environment. It describes situations that are related to IBM Spectrum Virtualize, Storwize, the SAN environment, storage subsystems, hosts, and multipathing drivers. It also explains how to collect the necessary problem determination data.

This chapter includes the following sections:

- ▶ Starting troubleshooting
- ▶ Remote Support Assistance
- ▶ Common issues
- ▶ Collecting data and isolating the problem
- ▶ Recovering from problems
- ▶ Health status during upgrade and known error
- ▶ Call Home Web and Health Checker feature

10.1 Starting troubleshooting

The Graphical User Interface (GUI) is a good start point for your troubleshooting. It has two icons at the top, which can be accessed from any panel of the GUI. As shown in Figure 10-1, the first icon shows IBM Spectrum Virtualize events, like an error or warning, and the second icon shows suggested tasks and background tasks that are running, or that were recently completed.

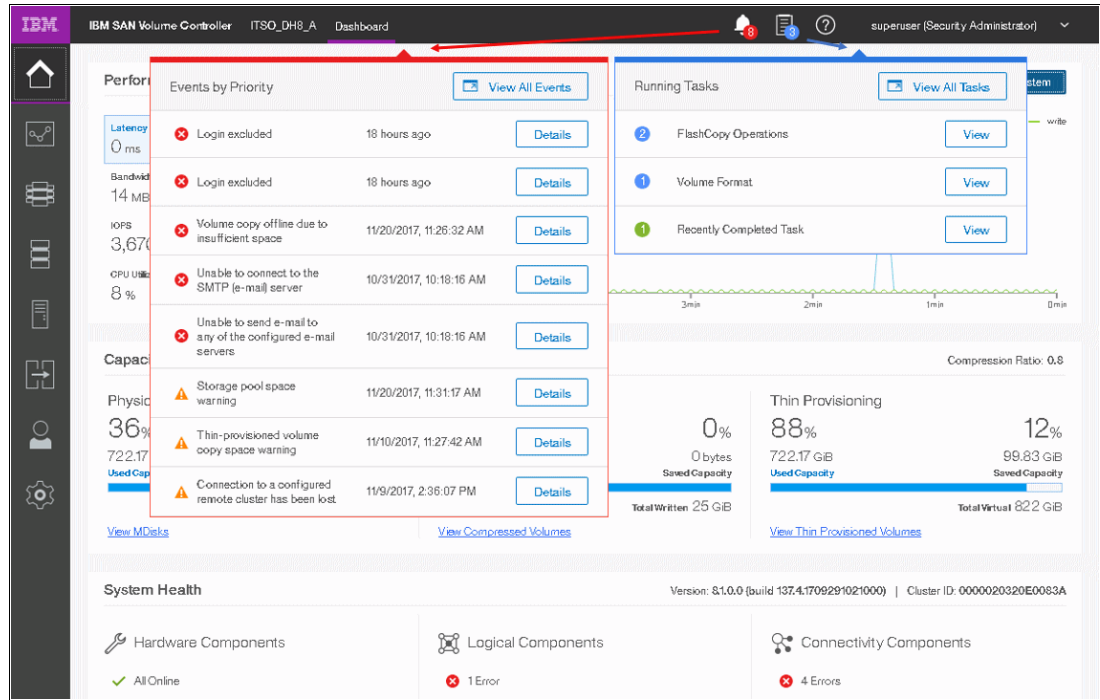


Figure 10-1 Events and tasks icons in GUI

The Dashboard provides an at-a-glance look into the condition of the system and notification of any critical issues that require immediate action. It contains sections for performance, capacity, and system health that provide an overall understanding of what is happening on the system.

Figure 10-2 on page 355 shows the Dashboard panel.

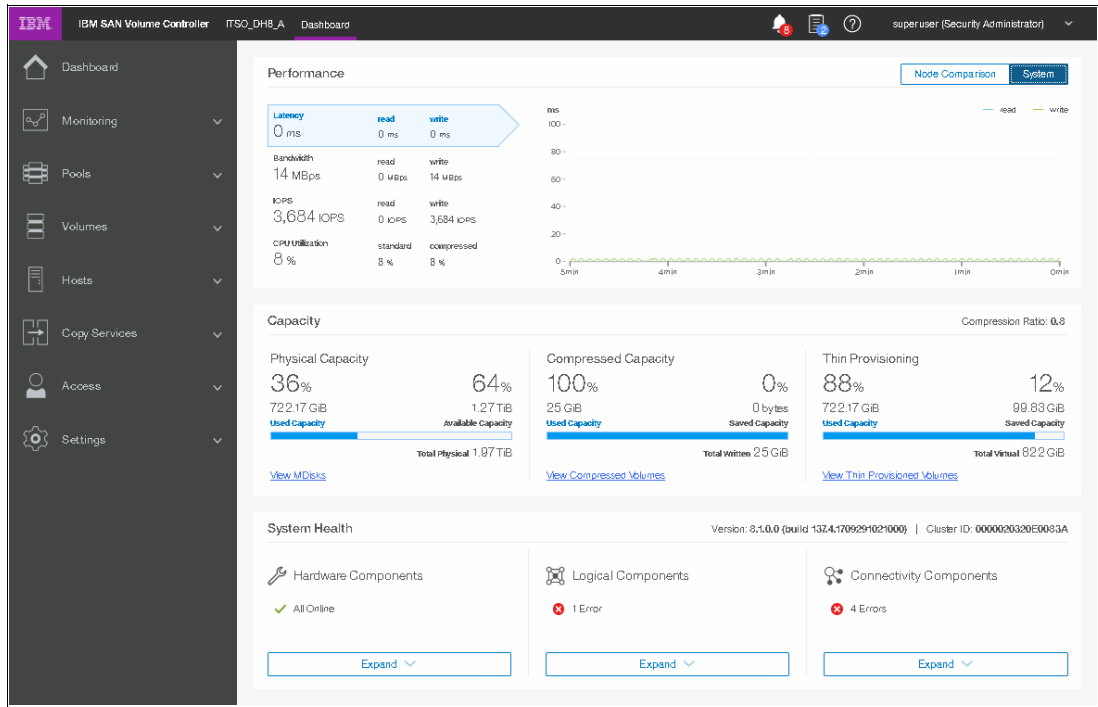


Figure 10-2 Dashboard

The System Health section in the bottom part of the Dashboard provides information on the health status of hardware, and logical and connectivity components. If you click **Expand** in each of these categories, the status of individual components is shown, as shown in the example in Figure 10-3. You can also go further and click **More Details**, which will take you to the panel related to that specific component, or will show you more information about it.

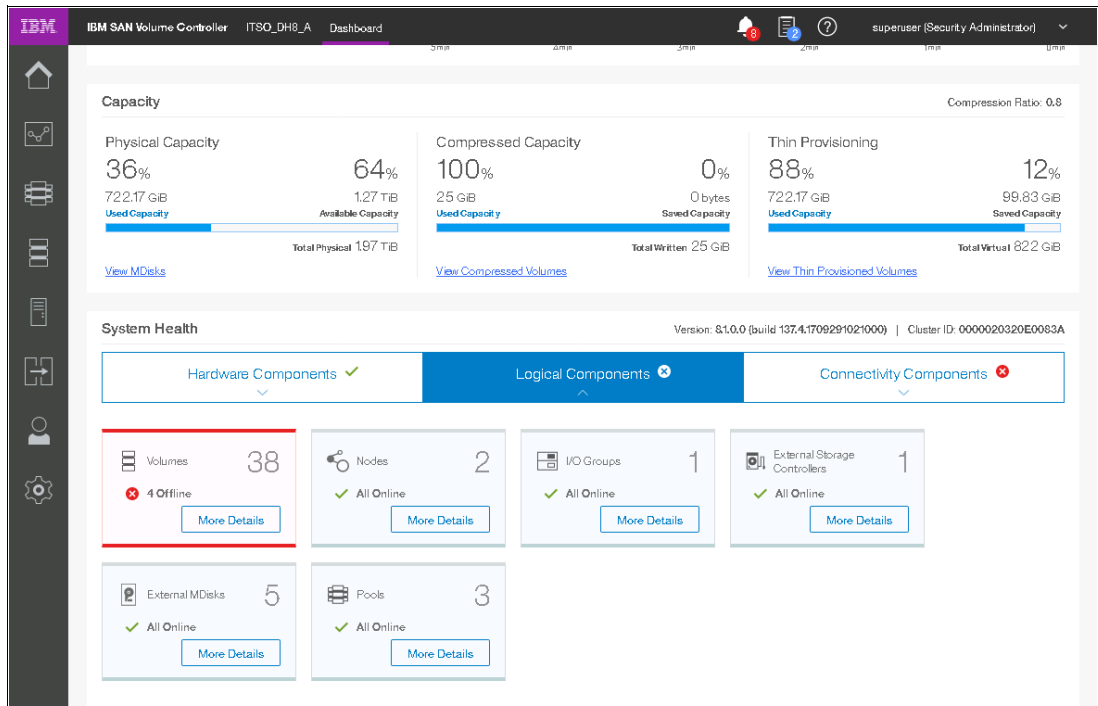


Figure 10-3 System Health section in Dashboard

The entire list of components in each category can be found in IBM Knowledge Center:

- ▶ IBM SAN Volume Controller
<https://ibm.biz/BdjLV8>
- ▶ IBM Storwize V7000
<https://ibm.biz/BdjLVJ>

More information about IBM Spectrum Virtualize troubleshooting can be found in IBM Knowledge Center:

- ▶ IBM SAN Volume Controller
<https://ibm.biz/BdjL7z>
- ▶ IBM Storwize V7000
<https://ibm.biz/BdjC2K>

10.1.1 Recommended actions and fix procedure

The **Monitoring** → **Events** panel shows information messages, warnings and issues on the IBM Spectrum Virtualize. So, this is a good place to check the current problems in the system.

Using the **Recommended Actions** filter, the most important events that need to be fixed are displayed.

If there is an important issue that needs to be fixed, the **Run Fix** button will be available in the top-left corner with an error message, indicating which event should be fixed as soon as possible. This fix procedure assists you to resolve problems in IBM Spectrum Virtualize. It analyzes the system, provides more information on the problem, suggest actions to be taken with steps to be followed, and finally checks to see if the problem is resolved.

So, if any error is reported by the system, such as system configuration problems and hardware failures, always use the fix procedures to resolve it.

Figure 10-4 on page 357 shows **Monitoring** → **Events** panel with the **Run Fix** button.

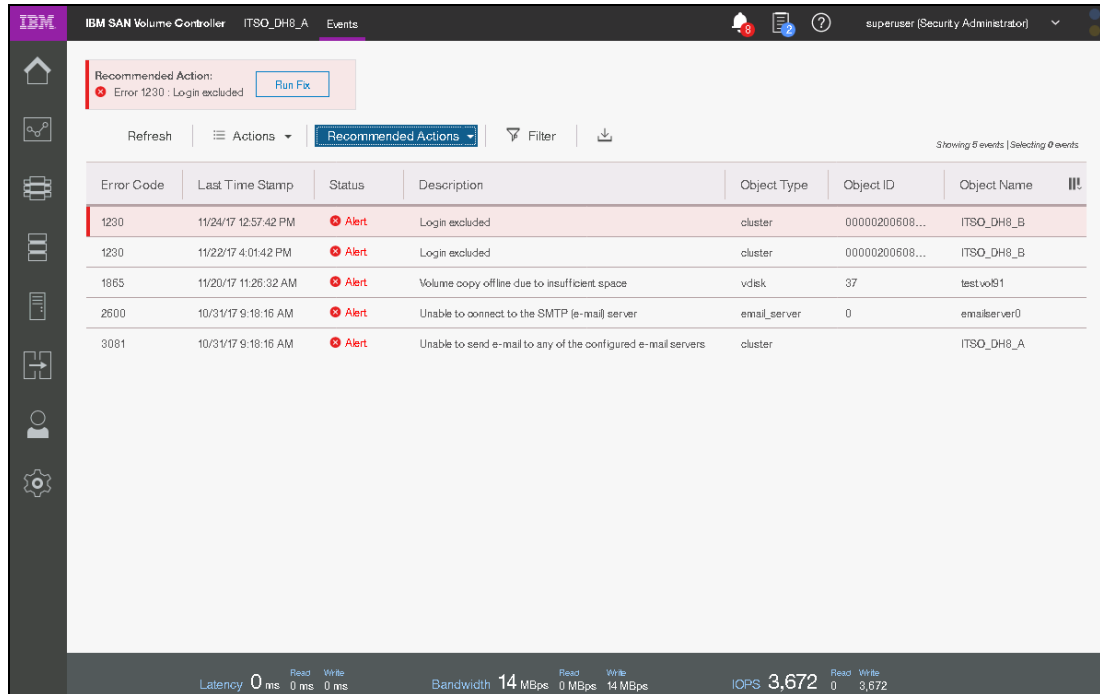


Figure 10-4 Monitoring → Events panel

Resolve alerts in a timely manner: When an issue or a potential issue is reported, resolve it as quick as possible to minimize its impact and potentially avoid more serious problems with your system.

10.2 Remote Support Assistance

Remote Support Assistance (RSA) enables IBM support to access the IBM Spectrum Virtualize device to perform troubleshooting and maintenance tasks. Support assistance can be configured to support personnel work on-site only, or to access the system both on-site and remotely. Both methods use secure connections to protect data in the communication between support center and system. Also, you can audit all actions that support personnel conduct on the system.

You can use just local support assistance if you have security restrictions that don't allow support to connect remotely to your systems. With Remote Support Assistance, support personnel can work both on-site or remotely through a secure connection from the support center. They can perform troubleshooting, upload support packages and download software to the system with your permission. When you configure remote support assistance in the GUI, local support assistance is enabled too.

With the remote support assistance method, you have two access types:

- ▶ At any time

Support center can start remote support sessions at any time.

- ▶ By permission only

Support center can start a remote support session only if permitted by an administrator. A time limit can be configured for the session.

Note: Clients who have purchased Enterprise Class Support (ECS) are entitled to IBM support using Remote Support Assistance to quickly connect and diagnose problems. However IBM support may choose to utilize this feature on non-ECS systems at their discretion, therefore we recommend configuring and testing the connection on all systems.

To configure remote support assistance, the following prerequisites should be met:

- ▶ Ensure that call home is configured with a valid email server.
- ▶ Ensure that a valid service IP address is configured on each node on the system.
- ▶ If your system is behind a firewall or if you want to route traffic from multiple storage systems to the same place, you must configure a Remote Support Proxy server. Before you configure remote support assistance, the proxy server must be installed and configured separately. The IP address and the port number for the proxy server needs to be set-up on when enabling remote support centers.
- ▶ If you do not have firewall restrictions and the storage nodes are directly connected to the Internet, request your network administrator to allow connections to 129.33.206.139 and 204.146.30.139 on port 22.
- ▶ Both uploading support packages and downloading software require direct connections to the Internet. A DNS server must be defined on your system for both of these functions to work.
- ▶ To ensure that support packages are uploaded correctly, configure the firewall to allow connections to the following IP addresses on port 443: 129.42.56.189, 129.42.54.189, and 129.42.60.189.
- ▶ To ensure that software is downloaded correctly, configure the firewall to allow connections to the following IP addresses on port 22: 170.225.15.105, 170.225.15.104, 170.225.15.107, 129.35.224.105, 129.35.224.104, and 129.35.224.107.

Remote support assistance can be configured both using GUI and CLI. The detailed steps to configure it can be found in the following publications:

- ▶ *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.1*, SG24-7933
- ▶ *Implementing the IBM Storwize V7000 with IBM Spectrum Virtualize V8.1*, SG24-7938

10.3 Common issues

SANs, storage subsystems and host systems can be complicated. They often consist of hundreds or thousands of disks, multiple redundant subsystem controllers, virtualization engines, and different types of SAN switches. All of these components must be configured, monitored, and managed properly. If issues occur, administrators must know what to look for and where to look.

IBM Spectrum Virtualize has useful error logging mechanisms. It keeps track of its internal events and informs the user about issues in the SAN or storage subsystem. It also helps to isolate problems with the attached host systems. So, with these functions, administrators can easily locate any issue areas and take the necessary steps to fix any events.

In many cases, IBM Spectrum Virtualize and its service and maintenance features guide administrators directly, provide help, and suggest remedial action. Furthermore, IBM Spectrum Virtualize determines whether the problem still persists or not.

Another feature that helps administrators to isolate and identify issues that might be related to IBM Spectrum Virtualize is the ability of their nodes to maintain a database of other devices that communicate with the IBM Spectrum Virtualize device. Devices, like hosts and back-end storages, are added or removed from the database as they start or stop communicating to IBM Spectrum Virtualize.

Although IBM Spectrum Virtualize node hardware and software events can be verified in the GUI or CLI, external events like failures in the SAN zoning configuration, hosts, and back-end storages are common. They need to have a troubleshooting performed out of IBM Spectrum Virtualize, too. As an example, a misconfiguration in the SAN zoning might lead to the IBM Spectrum Virtualize cluster not working properly.

This problem occurs because the IBM Spectrum Virtualize cluster nodes communicate with each other by using the Fibre Channel SAN fabrics.

In this case, check the following areas from an IBM Spectrum Virtualize perspective:

- ▶ The attached hosts. For more information, see 10.3.1, “Host problems” on page 359.
- ▶ The SAN. For more information, see 10.3.2, “SAN events” on page 361.
- ▶ The attached storage subsystem. For more information, see 10.3.3, “Storage subsystem issues” on page 361.
- ▶ The local FC port masking. For more information, see Chapter 6.1.5, “Port masking” on page 245.

10.3.1 Host problems

From the host perspective, you can experience various situations that range from performance degradation to inaccessible disks. To diagnose any host-related issue, you can start checking the hosts configuration on IBM Spectrum Virtualize side. The *Hosts* panel in the GUI or the following CLI commands should be used to start a verification in any possible hosts related issue:

▶ **lshost**

Check the host’s status. If status is online, it means the host ports are online in both nodes of an I/O group. If status is offline, it means the host ports are offline in both nodes of an I/O group. If status is inactive, it means that the host has volumes mapped to it, but all of its ports have no SCSI commands in the last 5 minutes. And, if status is degraded, it means at least one of the host ports but not all of them is not online in at least one node of an I/O group. Example 10-1 shows the **lshost** command output.

Example 10-1 lshost command

```
IBM_2145:ITS0_DH8_B:superuser>lshost
id name      port_count iogrp_count status  site_id site_name host_cluster_id
host_cluster_name
0 Win2K8     2          4        degraded
1 ESX_62_B  2          4        online
2 ESX_62_A  2          1        offline
```

▶ **lshost <host_id_or_name>**

It shows more details about a specific host, and it is very used in a case you need to identify which host port is not online in IBM Spectrum Virtualize node or IBM Storwize V7000 controller.

10.3.2 SAN events

Introducing IBM Spectrum Virtualize into your SAN environment and the use of its virtualization functions are not difficult tasks. However, before you can use IBM Spectrum Virtualize in your environment, you must follow some basic rules. These rules are not complicated, but you can make mistakes that lead to accessibility issues or a reduction in the performance experienced.

Two types of SAN zones are needed to run IBM Spectrum Virtualize in your environment: A *host zone* and a *storage zone*. In addition, you must have an IBM Spectrum Virtualize zone that contains all of the IBM Spectrum Virtualize node ports of the IBM Spectrum Virtualize cluster. This IBM Spectrum Virtualize zone enables intracluster communication. For more information and important points about setting up IBM Spectrum Virtualize in a SAN fabric environment, see Chapter 1, “Storage area network” on page 1.

Because IBM Spectrum Virtualize is in the middle of the SAN and connects the host to the storage subsystem, check and monitor the SAN fabrics.

10.3.3 Storage subsystem issues

Today, various heterogeneous storage subsystems are available. All of these subsystems have different management tools, different setup strategies, and possible problem areas depending on the manufacturer. To support a stable environment, all subsystems must be correctly configured, following the respective preferred practices and with no existing issues.

Check the following areas if you experience a storage-subsystem-related issue:

- ▶ Storage subsystem configuration. Ensure that a valid configuration and preferred practices are applied to the subsystem.
- ▶ Storage subsystem controllers. Check the health and configurable settings on the controllers.
- ▶ Storage subsystem array. Check the state of the hardware, such as a disk drive module (DDM) failure or enclosure alerts.
- ▶ Storage volumes. Ensure that the logical unit number (LUN) masking is correct.
- ▶ Host attachment ports. Check the status, configuration and connectivity to SAN switches.
- ▶ Layout and size of RAID arrays and LUNs. Performance and redundancy are contributing factors.

IBM Spectrum Virtualize has several CLI commands that you can use to check the status of the system and attached storage subsystems too. Before you start a complete data collection or problem isolation on the SAN or subsystem level, use the following commands first and check the status from the IBM Spectrum Virtualize perspective:

▶ **lscontroller <controller_id_or_name>**

Check that multiple worldwide port names (WWPNs) that match the back-end storage subsystem controller ports are available.

Check that the *path_counts* are evenly distributed across each storage subsystem controller, or that they are distributed correctly based on the preferred controller. Use the *path_count* calculation that is described in 10.5.4, “Solving back-end storage issues” on page 375. The total of all *path_counts* must add up to the number of managed disks (MDisks) multiplied by the number of IBM Spectrum Virtualize nodes.

► **lsmdisk**

Check that all MDisks are online (not degraded or offline).

► **lsmdisk <mdiskid_id_or_name>**

Check several of the MDisks from each storage subsystem controller. Are they online? Do they all have path_count = number of backend ports in the zone to IBM Spectrum Virtualize x number of nodes? See Example 10-4 for an example of the output from this command.

Example 10-4 Issuing an lsmdisk command

```
IBM_2145:itsosvcc11:superuser>lsmdisk 0
id 0
name flash9h01_itsosvcc11_0
status online
mode managed
mdisk_grp_id 0
mdisk_grp_name Pool0
capacity 1.6TB
quorum_index
block_size 512
controller_name itsoflash9h01
ctrl_type 4
ctrl_WWNN 500507605E852080
controller_id 1
path_count 32
max_path_count 32
ctrl_LUN_# 0000000000000000
UID 6005076441b53004400000000000000100000000000000000000000000000000
preferred_WWPN
active_WWPN many
.
lines removed for brevity
.
IBM_2145:itsosvcc11:superuser>
```

Example 10-4 shows that the Flash900 has eight ports zoned to IBM Spectrum Virtualize, and IBM Spectrum Virtualize has four nodes, so 8 x 4 = 32.

► **lsvdisk**

Check that all volumes are online (not degraded or offline). If the volumes are degraded, are there stopped FlashCopy jobs? Restart any stopped FlashCopy jobs or seek IBM Spectrum Virtualize support guidance.

► **lsfabric**

Use this command with the various options, such as **-controller controllerid**. Also, check different parts of the IBM Spectrum Virtualize configuration to ensure that multiple paths are available from each IBM Spectrum Virtualize node port to an attached host or controller. Confirm that all IBM Spectrum Virtualize node port WWPNs are connected to the back-end storage consistently.

For more information about managing subsystems, see Chapter 2, “Back-end storage” on page 49.

Determining the correct number of paths to a storage subsystem

By using IBM Spectrum Virtualize CLI commands, it is possible to determine the total number of paths to a storage subsystem. To determine the proper value of the available paths, use the following formula:

Number of MDisks x Number of SVC nodes per Cluster = Number of paths
mdisk_link_count x Number of SVC nodes per Cluster = Sum of path_count

Example 10-5 shows how to obtain this information by using the **lscontroller** <controllerid> and **svcinfo lsnode** commands.

Example 10-5 Output of the svcinfo lscontroller command

```
IBM_2145:itsosvcc11:superuser>lscontroller 1
id 1
controller_name itsof9h01
WWNN 500507605E852080
mdisk_link_count 16
max_mdisk_link_count 16
degraded no
vendor_id IBM
product_id_low FlashSys
product_id_high tem-9840
product_revision 1430
ctrl_s/n 01106d4c0110-0000-0
allow_quorum yes
fabric_type fc
site_id
site_name
WWPN 500507605E8520B1
path_count 64
max_path_count 64
WWPN 500507605E8520A1
path_count 64
max_path_count 64
WWPN 500507605E852081
path_count 64
max_path_count 64
WWPN 500507605E852091
path_count 64
max_path_count 64
WWPN 500507605E8520B2
path_count 64
max_path_count 64
WWPN 500507605E8520A2
path_count 64
max_path_count 64
WWPN 500507605E852082
path_count 64
max_path_count 64
WWPN 500507605E852092
path_count 64
max_path_count 64
IBM_2145:itsosvcc11:superuser>svcinfo lsnode
id name UPS_serial_number WWNN status IO_group_id IO_group_name
config_node UPS_unique_id hardware iscsi_name
```

```

iscsi_alias panel_name enclosure_id canister_id enclosure_serial_number site_id
site_name
1 node1          500507680C003AE1 online 0          io_grp0    yes
DH8 iqn.1986-03.com.ibm:2145.itsosvcc11.node1 78CBFEA0
2 node2          500507680C003ACA online 0          io_grp0    no
DH8 iqn.1986-03.com.ibm:2145.itsosvcc11.node2 78CBB0
3 node3          500507680C003A9F online 1          io_grp1    no
DH8 iqn.1986-03.com.ibm:2145.itsosvcc11.node3 78CBLP0
4 node4          500507680C003DB6 online 1          io_grp1    no
DH8 iqn.1986-03.com.ibm:2145.itsosvcc11.node4 78CCAQ0
IBM_2145:itsosvcc11:superuser>

```

Example 10-5 shows that sixteen MDisks are present for the storage subsystem controller with ID 1, and four IBM Spectrum Virtualize nodes are in the IBM Spectrum Virtualize cluster. In this example, the path_count is 16 x 4 = 64.

10.3.4 Port masking issues

Some situations of performance degradation and buffer-to-buffer credit exhaustion can be caused by incorrect local FC port masking and remote FC port masking. To have a healthy operation in your IBM Spectrum Virtualize, configure both your local FC port masking and your remote FC port masking accordingly.

The ports intended to have only intracluster/node to node communication traffic must not have replication data or host/back-end data running on it. The ports intended to have only replication traffic must not have intracluster/node to node communication data or host/back-end data running on it.

10.3.5 Interoperability

When you experience events in the IBM Spectrum Virtualize environment, ensure that all components that comprise the storage infrastructure are interoperable. In an IBM Spectrum Virtualize environment, the IBM Spectrum Virtualize support matrix is the main source for this information. For the latest IBM Spectrum Virtualize support matrix, see *IBM System Storage Interoperation Center (SSIC) website*:

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

Although the latest IBM Spectrum Virtualize code level is supported to run on older host bus adapters (HBAs), storage subsystem drivers, and code levels, use the latest tested levels for best results.

10.4 Collecting data and isolating the problem

Data collection and problem isolation in an IT environment are sometimes difficult tasks. In the following section, the essential steps that are needed to collect debug data to find and isolate problems in an IBM Spectrum Virtualize environment are described.

10.4.1 Collecting data from IBM Spectrum Virtualize

When there is a problem with an IBM SAN Volume Controller or Storwize V7000 and you have to open a case with IBM support, you need to provide the support packages for the device. To collect and upload the support packages to IBM support center, you can do it automatically via IBM Spectrum Virtualize, or download the package from the device and manually upload to IBM. The easiest way is automatically upload the support packages from IBM Spectrum Virtualize. It can be done via both GUI and CLI.

The next sections show how to automatically upload the support package to IBM support center. More details of this procedure can be found in:

- ▶ *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.1, SG24-7933*
- ▶ *Implementing the IBM Storwize V7000 with IBM Spectrum Virtualize V8.1, SG24-7938*

Data collection using the GUI

To perform data collection using the GUI, complete the following steps:

1. In the panel **Settings** → **Support** → **Support Package**, both options to collect and upload support packages are available.
2. To automatically upload them, click **Upload Support Package** button.
3. In the pop-up screen, enter the PMR number and the type of support package to upload to the IBM support center. The **Snap Type 4** can be used to collect standard logs and generate a new statesave on each node of the system.

4. The *Upload Support Package* panel is shown in Figure 10-5.

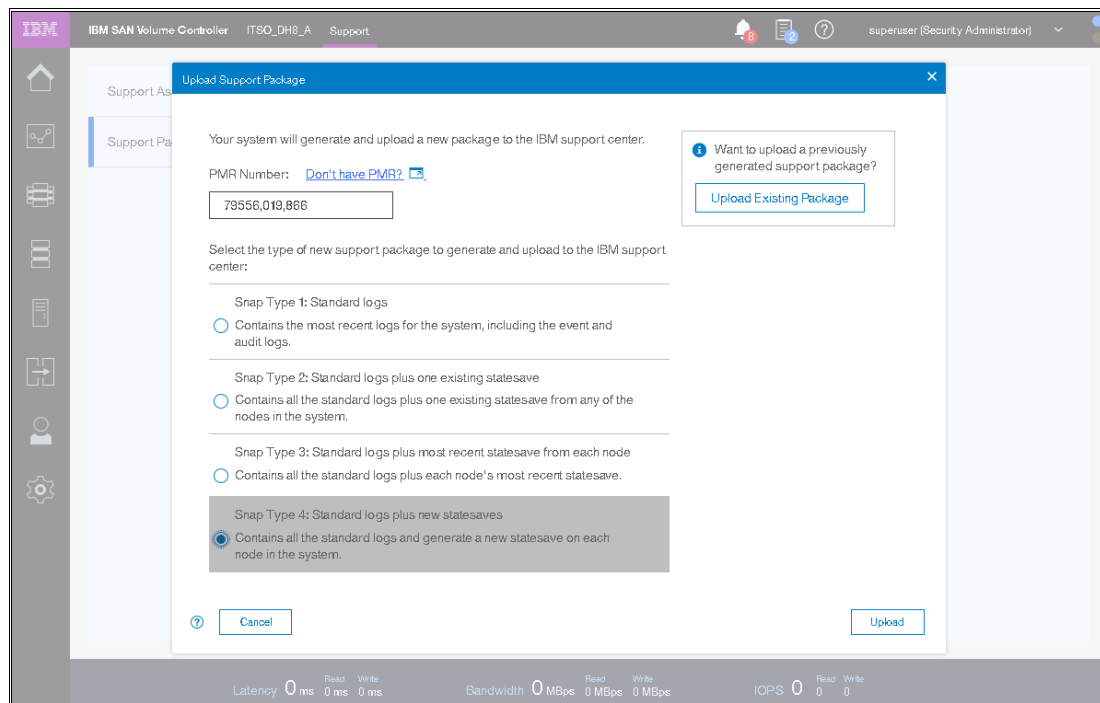


Figure 10-5 Upload Support Package panel

Data collection using the CLI

To collect the same type of support packages mentioned above using the CLI, you have to first generate a new livedump of the system using the `svc_livedump` command, and then upload the log files and new generated dumps using the `svc_snap` command, as shown in Example 10-6. To verify if the support package was successfully uploaded, use the `sainfo lscmdstatus` command.

Example 10-6 The `svc_livedump` command

```
IBM_2145:ITS0_DH8_B:superuser>svc_livedump -nodes all -yes
Livedump - Fetching Node Configuration
Livedump - Checking for dependent vdisks
Livedump - Check Node status
Livedump - Prepare specified nodes - this may take some time...
Livedump - Prepare node 1
Livedump - Prepare node 2
Livedump - Trigger specified nodes
Livedump - Triggering livedump on node 1
Livedump - Triggering livedump on node 2
Livedump - Waiting for livedumps to complete dumping on nodes 1,2
Livedump - Waiting for livedumps to complete dumping on nodes 2
Livedump - Successfully captured livedumps on nodes 1,2
IBM_2145:ITS0_DH8_B:superuser>svc_snap upload pmr=ppppp,bbb,ccc gui3
Collecting data
Packaging files
Snap data collected in /dumps/snap.ABCDEFG.171128.223133.tgz
IBM_2145:ITS0_DH8_B:superuser>sainfo lscmdstatus
last_command satask supportupload -pmr ppppp,bbb,ccc -filename
/dumps/snap.ABCDEFG.171128.223133.tgz
```

```
last_command_status CMMVC8044E Command completed successfully.
T3_status
T3_status_data
cpfiles_status Complete
cpfiles_status_data Copied 1 of 1
snap_status Complete
snap_filename /dumps/snap.ABCDEFG.171128.223133.tgz
installcanistersoftware_status
supportupload_status Complete
supportupload_status_data [PMR=ppppp,bbb,ccc] Upload complete
supportupload_progress_percent 0
supportupload_throughput_KBps 0
supportupload_filename /dumps/snap.ABCDEFG.171128.223133.tgz
downloadsoftware_status
downloadsoftware_status_data
downloadsoftware_progress_percent 0
downloadsoftware_throughput_KBps 0
downloadsoftware_size
IBM_2145:ITS0_DH8_B:superuser>
```

10.4.2 SDDPCM and SDDDSM data collection

If there is a problem related to host communication with IBM SAN Volume Controller or Storwize V7000, collecting data from hosts and multipath software is very useful.

SDDPCM for AIX provides the **sddpcmgetdata** script to collect information used for problem determination. This script creates a tar file at the current directory with the current date and time as a part of the file name. When you suspect you have an issue with SDDPCM, it is essential to run this script and send this tar file to IBM support.

SDDDSM for Windows hosts also contains a utility to collect information for problem determination. The **sddgetdata.bat** tool creates a CAB file in the installation directory with the current date and time as part of the file name. The CAB file includes the following information:

- ▶ SystemInfo
- ▶ HKLM \SYSTEM\CurrentControlSet, HKLM\HARDWARE\DEVICEMAP, and HKLM\Cluster output from registry
- ▶ SDDDSM directory contents
- ▶ HBA details
- ▶ Datapath outputs
- ▶ Pathtest trace
- ▶ SDDSRV logs
- ▶ Cluster logs
- ▶ System disks and paths

The execution of **sddgetdata.bat** tool is shown in Example 10-7.

Example 10-7 sddgetdata.bat tool

```
C:\Program Files\IBM\SDDDSM>sddgetdata.bat
Collecting SDD trace Data
```

```
Flushing SDD kernel logs
SDD logs flushed
Collecting datapath command outputs
Collecting System Information
Collecting SDD and SDDSRV logs
Collecting Most current driver trace
Please wait for 30 secs... Writing DETAILED driver trace to trace.out
Generating a CAB file for all the Logs
sdddata_WIN-IWG6VLJN3U3_20171129_151423.cab file generated
C:\Program Files\IBM\SDDDSM>
```

More information about diagnostics for IBM SDD can be found in the latest *Multipath Subsystem Device Driver User's Guide* at:

<https://ibm.biz/BdjCyy>

10.4.3 Additional data collection

Data collection methods vary by storage platform, SAN switch and operating system.

When there is an issue in a SAN environment and it is not clear where the problem is occurring, you might have to collect data from several devices in the SAN.

Bellow you can find basic information that should be collected for each type of device:

- ▶ Hosts
 - Operating system: Version and level
 - HBA: Driver and firmware level
 - Multipathing driver level
- ▶ SAN switches
 - Hardware model
 - Software version
- ▶ Storage subsystems
 - Hardware model
 - Software version

Regarding host, storage and SAN data collection, due to the dynamic changes that occur over time, follow this IBM w3 Connections community (only available to IBMers):

<https://ibm.biz/emea-coc>

Note: This community is IBM internal only. You must use your intranet ID and password. If you are not an IBM employee, contact your IBM representative or the vendor of your hardware and follow the specific procedures for data collection.

The w3 Connections community has up-to-date procedures for several kinds of devices, including hosts, storage, and SAN, as shown in Figure 10-6.

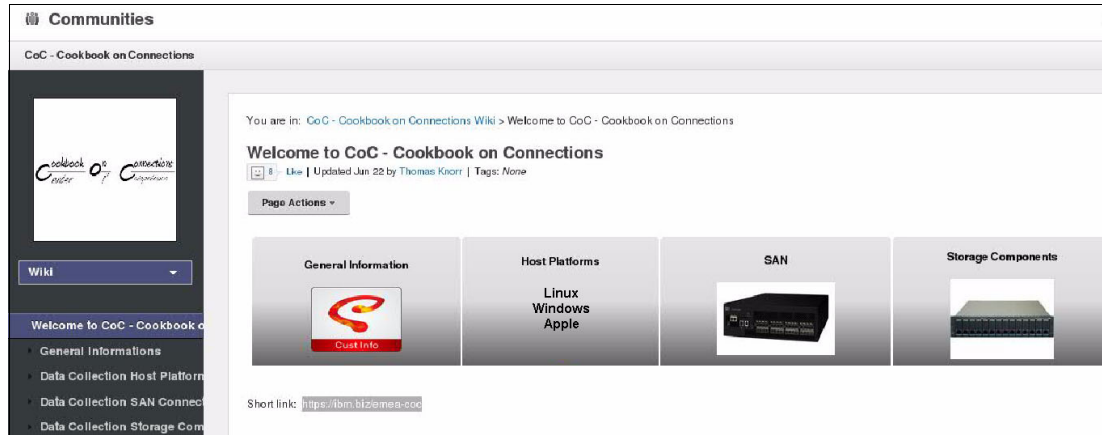


Figure 10-6 CoC - Cookbook on Connections internal wiki

10.5 Recovering from problems

You can recover from several of the more common events that you might encounter. In all cases, you must read and understand the current product limitations to verify the configuration and to determine whether you need to upgrade any components or install the latest fixes or patches.

To obtain support for any IBM product, see the following IBM Support website:

<https://www.ibm.com/support/home/>

For more information about the latest flashes, concurrent code upgrades, code levels, and compatibility matrix, see the following IBM SAN Volume Controller and Storwize V7000 support websites:

- ▶ IBM SAN Volume Controller
<https://ibm.biz/Bdj9ka>
- ▶ IBM Storwize V7000
<https://ibm.biz/Bdj9kn>

10.5.1 Solving IBM Spectrum Virtualize events

For any events in the IBM SAN Volume Controller or Storwize V7000, before you try to fix the problem anywhere else, use the **Recommended Actions** functionality in **Monitoring** → **Events** panel. This is shown in 10.1.1, “Recommended actions and fix procedure” on page 356.

The Events panel with the **Recommended Actions** filter shows event conditions that require actions and the procedures to diagnose and fix them. The highest-priority event is indicated in the top-left corner and it will also appear highlighted in the table of events.

If for any reason you need to run the **Fix Procedure** in another event before fixing the highest-priority event, you must select the event, click in **Actions** menu and select Run Fix Procedure, as shown in the Figure 10-7.

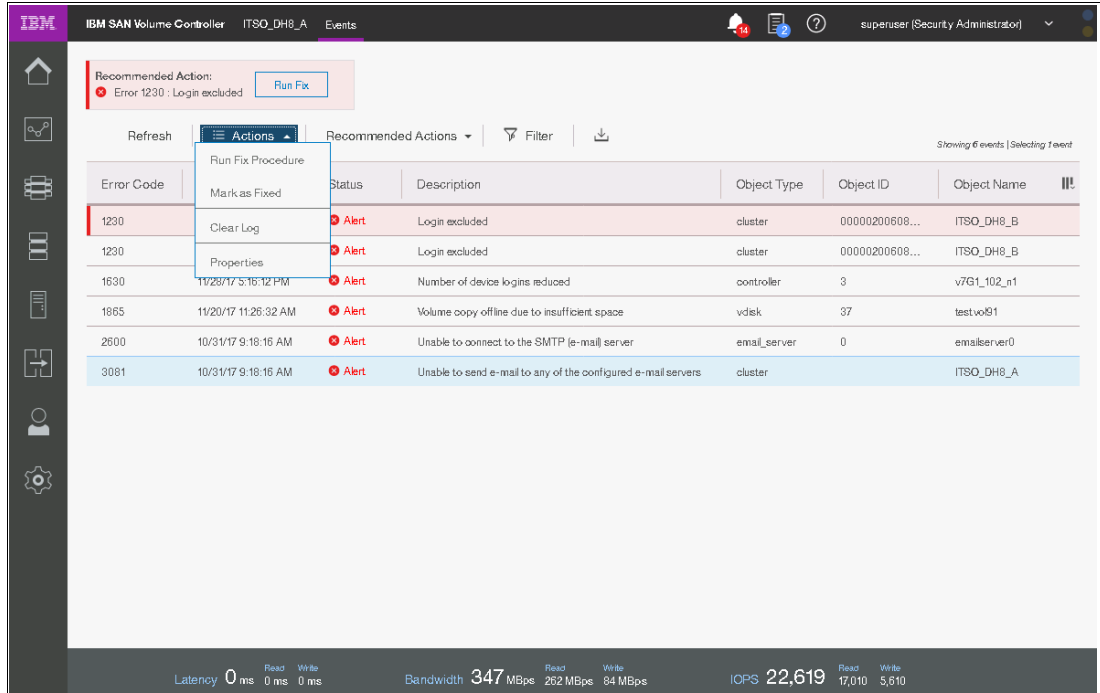


Figure 10-7 Action menu in events table

To obtain more information about any event, select an event in the table, and click **Properties** in the **Actions** menu, as shown in Figure 10-7.

Tip: You can also get access to **Run Fix Procedure** and **Properties** by right-clicking an event.

In the *Properties and Sense Data* window for the specific event, as shown in Figure 10-8 on page 371, additional information about it is displayed. You can review and also click **Run Fix** to run the **Fix Procedure**.

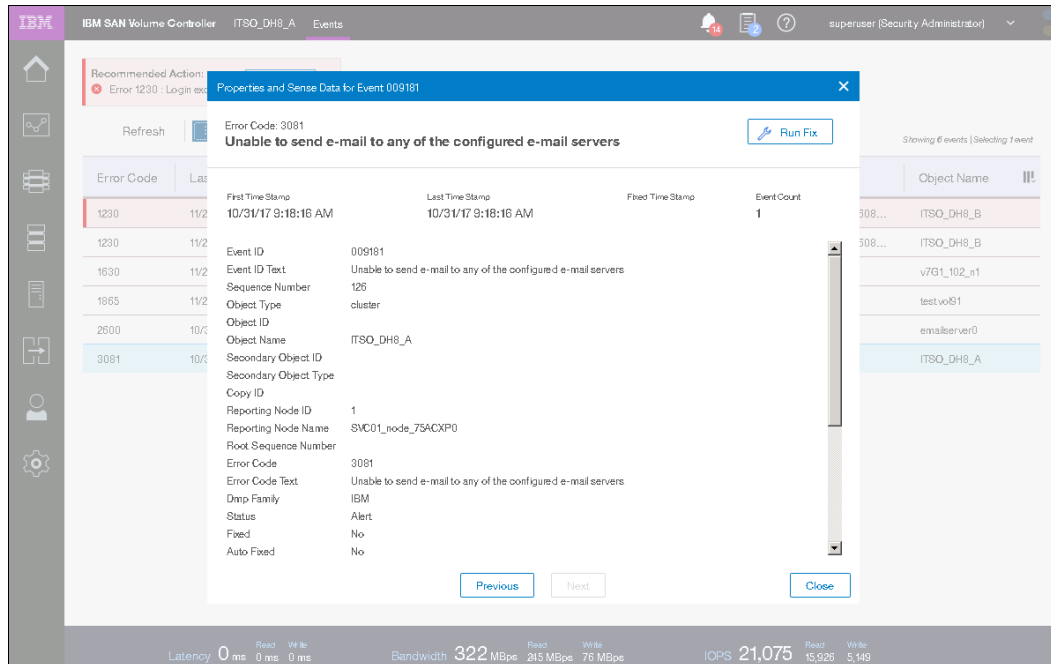


Figure 10-8 Properties and sense data for event window

Tip: From the Properties and Sense Data for Event Window, you can use the **Previous** and **Next** buttons to move between events.

Another common practice is to use the IBM Spectrum Virtualize CLI to find issues and resolve them. The following list of commands scan back-end storage subsystems changes and provides information about the status of your environment:

- ▶ **lseventlog**
Display a list of events and more detailed information about an event.
- ▶ **detectmdisk**
Discovers changes in the back-end storage configuration.
- ▶ **lspartnership**
Checks the IBM Spectrum Virtualize local and remote clusters status.
- ▶ **lssystem**
Displays detailed information about IBM Spectrum Virtualize cluster.
- ▶ **svcinfo lsnode <node_id_or_name>**
Checks the IBM Spectrum Virtualize nodes and port status.
- ▶ **lscontroller <controller_id_or_name>**
Checks the back-end storage status.
- ▶ **lsmdisk <mdisk_id_or_name>**
Provides a status of MDisk.
- ▶ **lsmdiskgrp <mdiskgrp_id_or_name>**
Provides a status for storage pools.
- ▶ **lsvdisk <vdisk_id_or_name>**
Checks whether volumes are online and working correctly.

Locating issues: Although IBM Spectrum Virtualize raises error messages, most events are not caused by IBM Spectrum Virtualize. Most issues are introduced by the storage subsystems or the SAN.

If the problem is caused by IBM Spectrum Virtualize and you are unable to fix it by using the Recommended Action feature or the event log, collect the IBM Spectrum Virtualize support package as described in 10.4.1, “Collecting data from IBM Spectrum Virtualize” on page 365. To identify and fix other issues outside of IBM Spectrum Virtualize, consider the guidance in the other sections in this chapter that are not related to IBM Spectrum Virtualize.

Replacing a failed disk drive

When IBM Spectrum Virtualize detects a failed disk drive, it automatically generates an error in the *Events* panel. To replace the failed disk drive, always run **Fix Procedure** for this event in **Monitoring** → **Events** panel.

The **Fix Procedure** will help you to identify the enclosure and slot where the bad drive is located, and will guide you to the correct steps to follow in order to replace it. When a disk drive fails, it is removed from the array. If a suitable spare drive is available, it is taken into the array and the rebuild process starts on this drive.

After the failed disk drive is replaced and the system detects the replacement, it reconfigures the new drive as spare. So, the failed disk drive is removed from the configuration, and the new drive is then used to fulfill the array membership goals of the system.

10.5.2 Solving host problems

Apart from hardware-related situations, problems can exist in such areas as the operating system or the software that is used on the host. These problems normally are handled by the host administrator or the service provider of the host system. However, the multipathing driver that is installed on the host and its features can help to determine possible issues.

Example 10-8 shows a volume path issue reported by SDD output on the host by using the **datapath query adapter** and **datapath query device** commands. The adapter in degraded state means that specific HBA on the server side can't reach all the nodes in the I/O group which the volumes are associated. Also we can notice in the **datapath query device** command output that each device (or volume) has only three paths, when it is expected to have four paths.

Example 10-8 SDD output on a host with faulty paths

```
C:\Program Files\IBM\SDDDSM>datapath query adapter
```

```
Active Adapters :2
```

Adpt#	Name	Interface	State	Mode	Select	Errors	Paths	Active
0	Scsi Port2 Bus0	FC	DEGRAD	ACTIVE	1860589	293	4	4
1	Scsi Port3 Bus0	FC	NORMAL	ACTIVE	1979793	259	8	8

```
C:\Program Files\IBM\SDDDSM>datapath query device
```

```
Total Devices : 8
```

```
DEV#: 0 DEVICE NAME: Disk2 Part0 TYPE: 2145 POLICY: LEAST I/O AND WEIGHT  
SERIAL: 600507680C838020E80000000000001B Reserved: No LUN SIZE: 5.0GB
```


HOST INTERFACE: FC
PREFERRED PATH SET :None

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
2	Scsi Port2 Bus0/Disk2 Part0	OPEN	NORMAL	2859569	0
5 *	Scsi Port3 Bus0/Disk2 Part0	OPEN	NORMAL	0	0
6	Scsi Port3 Bus0/Disk2 Part0	OPEN	NORMAL	2466689	0

DEV#: 1 DEVICE NAME: Disk3 Part0 TYPE: 2145 POLICY: LEAST I/O AND WEIGHT
SERIAL: 600507680C838020E80000000000001E Reserved: No LUN SIZE: 10.0GB
HOST INTERFACE: FC
PREFERRED PATH SET :None

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
2	Scsi Port2 Bus0/Disk3 Part0	OPEN	NORMAL	35	0
5 *	Scsi Port3 Bus0/Disk3 Part0	OPEN	NORMAL	0	0
6	Scsi Port3 Bus0/Disk3 Part0	OPEN	NORMAL	22	0

.
lines removed for brevity

Faulty paths can be caused by hardware and software problems, such as the following examples:

- ▶ Hardware
 - Faulty Small Form-factor Pluggable transceiver (SFP) on the host or SAN switch
 - Faulty fiber optic cables
 - Faulty HBAs
- ▶ Software
 - A back-level multipathing driver
 - Obsolete HBA firmware or driver
 - Wrong zoning
 - Incorrect host-to-VDisk mapping

Based on field experience, complete the following hardware checks first:

- ▶ Check whether any connection error indicators are lit on the host or SAN switch.
- ▶ Check whether all of the parts are seated correctly. For example, cables are securely plugged in to the SFPs and the SFPs are plugged all the way into the switch port sockets.
- ▶ Ensure that no fiber optic cables are broken. If possible, swap the cables with cables that are known to work.

After the hardware check, continue to check the following aspects of software setup:

- ▶ Check that the HBA driver level and firmware level are at the preferred and supported levels.
- ▶ Check the multipathing driver level, and make sure that it is at the preferred and supported level.
- ▶ Check for link layer errors that are reported by the host or the SAN switch, which can indicate a cabling or SFP failure.
- ▶ Verify your SAN zoning configuration.
- ▶ Check the general SAN switch status and health for all switches in the fabric.

The output of SDD commands also helps to troubleshoot possible connectivity issues to the IBM Spectrum Virtualize device. Example 10-9 shows that one of the HBAs reported errors, and the cause can be any of the hardware or software examples mentioned above.

Example 10-9 Output from datapath query adapter and datapath query device

C:\Program Files\IBM\SDDSM>**datapath query adapter**

Active Adapters :2

Adpt#	Name	Interface	State	Mode	Select	Errors	Paths	Active
0	Scsi Port2 Bus0	FC	NORMAL	ACTIVE	1755262	12	8	8
1	Scsi Port3 Bus0	FC	NORMAL	ACTIVE	1658236	0	8	8

C:\Program Files\IBM\SDDSM>**datapath query device**

Total Devices : 8

DEV#: 0 DEVICE NAME: Disk2 Part0 TYPE: 2145 POLICY: LEAST I/O AND WEIGHT
 SERIAL: 600507680C838020E8000000000001B Reserved: No LUN SIZE: 5.0GB
 HOST INTERFACE: FC
 PREFERRED PATH SET :None

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
1 *	Scsi Port2 Bus0/Disk2 Part0	OPEN	NORMAL	0	0
2	Scsi Port2 Bus0/Disk2 Part0	OPEN	NORMAL	1599542	10
5 *	Scsi Port3 Bus0/Disk2 Part0	OPEN	NORMAL	0	0
6	Scsi Port3 Bus0/Disk2 Part0	OPEN	NORMAL	1492830	0

DEV#: 1 DEVICE NAME: Disk3 Part0 TYPE: 2145 POLICY: LEAST I/O AND WEIGHT
 SERIAL: 600507680C838020E8000000000001E Reserved: No LUN SIZE: 10.0GB
 HOST INTERFACE: FC
 PREFERRED PATH SET :None

```
=====
```

Path#	Adapter/Hard Disk	State	Mode	Select	Errors
1 *	Scsi Port2 Bus0/Disk3 Part0	OPEN	NORMAL	0	0
2	Scsi Port2 Bus0/Disk3 Part0	OPEN	NORMAL	9	0
5 *	Scsi Port3 Bus0/Disk3 Part0	OPEN	NORMAL	0	0
6	Scsi Port3 Bus0/Disk3 Part0	OPEN	NORMAL	10	0

.
 lines removed for brevity

10.5.3 Solving SAN issues

Some situations can cause issues in the SAN fabric and SAN switches. Problems can be related to a hardware fault or to a software problem on the switch. The following hardware defects are normally the easiest problems to find:

- ▶ Switch power, fan, or cooling units
- ▶ Installed SFP modules
- ▶ Fiber optic cables

Software failures are more difficult to analyze. In most cases, you must collect data and involve IBM Support. But before you take any other steps, check the installed code level for any known issues. Also, check whether a new code level is available that resolves the problem that you are experiencing.

The most common SAN issues often are related to zoning. For example, perhaps you chose the wrong WWPN for a host zone, such as when two IBM Spectrum Virtualize node ports must be zoned to one HBA with one port from each IBM Spectrum Virtualize node. However, as shown in Example 10-10, two ports are zoned that belong to the same node. Therefore, the result is that the host and its multipathing driver do not see all of the necessary paths.

Example 10-10 Incorrect WWPN zoning

```
zone: Senegal_Win2k3_itsosvcc11_iogrp0_Zone
      50:05:07:68:01:20:37:dc
      50:05:07:68:01:40:37:dc
      20:00:00:e0:8b:89:cc:c2
```

The correct zoning must look like the zoning that is shown in Example 10-11.

Example 10-11 Correct WWPN zoning

```
zone: Senegal_Win2k3_itsosvcc11_iogrp0_Zone
      50:05:07:68:01:40:37:e5
      50:05:07:68:01:40:37:dc
      20:00:00:e0:8b:89:cc:c2
```

The following IBM Spectrum Virtualize error codes are related to the SAN environment:

- ▶ Error 1060 Fibre Channel ports are not operational.
- ▶ Error 1220 A remote port is excluded.

Bottleneck is another common issue related to SAN switches. The bottleneck can be presented in a port where a host, storage subsystem or IBM Spectrum Virtualize device is connected, or in Inter-Switch Link (ISL) ports. The bottleneck can occur in some cases, like when a device connected to the fabric is slow to process received frames or if a SAN switch port is unable to transmit frames at a rate that is required by a device connected to the fabric.

These cases can slow down communication between devices in your SAN. To resolve this type of issue, you have to refer to the SAN switch documentation or open a case with the vendor to investigate and identify what is causing the bottleneck and fix it.

If you cannot fix the issue with these actions, use the method that is described in 10.4, “Collecting data and isolating the problem” on page 365, collect the SAN switch debugging data, and then contact the vendor for assistance.

10.5.4 Solving back-end storage issues

IBM Spectrum Virtualize has useful tools for finding and analyzing back-end storage subsystem issues because it has a monitoring and logging mechanism.

Typical events for storage subsystem controllers include incorrect configuration, which results in a *1625 - Incorrect disk controller configuration* error code. Other issues related to the storage subsystem include failures pointing to the managed disk I/O (error code 1310), disk media (error code 1320), and error recovery procedure (error code 1370).

10.5.5 Common error recovery using IBM Spectrum Virtualize CLI

For SAN or back-end storage issues, you can use the IBM Spectrum Virtualize CLI to perform common error recovery steps. Although the maintenance procedures perform these steps, it is sometimes faster to run these commands directly through the CLI.

Run these commands any time that you have the following issues:

- ▶ You experience a back-end storage issue (for example, error code 1370 or error code 1630).
- ▶ You performed maintenance on the back-end storage subsystems.

Important: Run these commands when back-end storage is just configured, a zoning change occurs or any other type of changes related to the communication between IBM Spectrum Virtualize and back-end storage subsystem happens. This ensures that IBM Spectrum Virtualize has recognized the changes.

Common error recovery involves the following IBM Spectrum Virtualize CLI commands:

- ▶ **lscontroller** and **lsmdisk**
Provides current status of all controllers and MDisks.
- ▶ **detectmdisk**
Discovers the changes in the back end.
- ▶ **lscontroller <controller_id_or_name>**
Checks the controller that was causing the issue and verifies that all the WWPNs are listed as you expect. It also checks that the path_counts are distributed evenly across the WWPNs.
- ▶ **lsmdisk**
Determines whether all MDisks are now online.

Note: When an issue is resolved using the CLI, check if the error has disappeared from **Monitoring** → **Events** panel. If not, make sure the error has been really fixed, and if so, manually mark the error as fixed.

10.6 Health status during upgrade and known error

It's important to understand that during the software upgrade process, alerts indicating the system is not healthy are reported. This is a normal behavior because the IBM SAN Volume Controller nodes or IBM Storwize V7000 controllers go offline during this process, so the system triggers these alerts.

Known error

While trying to upgrade an IBM Spectrum Virtualize, you might get a message such as Error in verifying the signature of the update package.

This message does not mean that you have an issue in your system. Sometimes this happens because there is not enough space on the system to copy the file, or the package is incomplete or contains errors. In this case, open a PMR with IBM support and follow their instructions.

10.7 Call Home Web and Health Checker feature

Call Home Web is an IBM tool to view Call Home information on the web.

Call Home is a functionality present in several IBM systems, including IBM SAN Volume Controller and Storwize V7000, which allow them to automatically report problems and status to IBM.

Call Home Web provides the following information about IBM systems:

- ▶ Automated tickets
- ▶ Warranty and contract status
- ▶ Health check alerts and recommendations
- ▶ System connectivity heartbeat
- ▶ Recommended software levels
- ▶ Inventory
- ▶ Security bulletins

To access the Call Home Web, go to IBM support website at:

<http://support.ibm.com>

In the IBM support website, Call Home Web is available at **My support** → **Call Home Web** as shown in Figure 10-9.

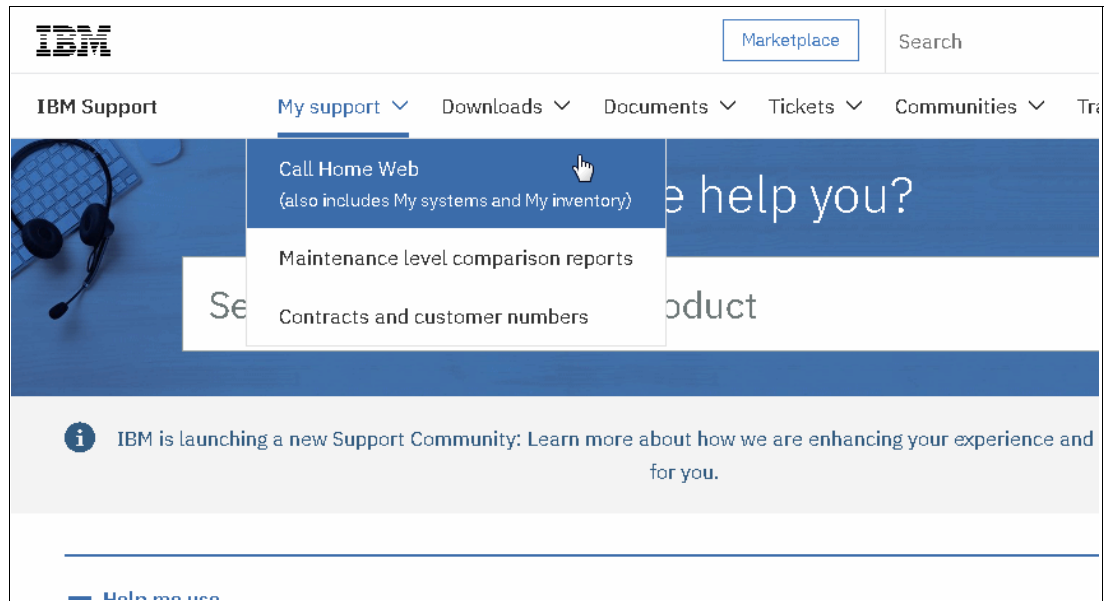


Figure 10-9 Call Home Web

To allow Call Home Web analyze data of IBM Spectrum Virtualize systems and provide useful information about them, the devices need to be added to the tool. The machine type, model and serial number are required to register the product in Call Home Web. Also, it is required that IBM SAN Volume Controller or Storwize V7000 have call home and inventory notification enabled and operational.

For a video guide on how to setup and use IBM Call Home Web see:

<https://www.youtube.com/watch?v=7G9rqk8NXPA>

10.7.1 Health Checker

A new feature of Call Home Web is the Health Checker, a tool that runs in the IBM Cloud™.

It analyzes call home and inventory data of systems registered in Call Home Web and validates their configuration. Then, it displays alerts and provide recommendations in the Call Home Web tool.

Note: Call Home Web should be used because it provides useful information about your systems, and with the Health Checker feature it helps you to monitor the system, and proactively provides alerts and creates recommendations related to them.



IBM Real-time Compression

This chapter highlights the preferred practices for IBM Real-time Compression that uses IBM Spectrum Virtualize software installed on IBM SAN Volume Controller, IBM Storwize family, and IBM FlashSystem V9000. The main goal is to provide compression users with guidelines and factors to consider to achieve the best performance results and enjoy the compression savings that the Real-time Compression technology offers.

This chapter assumes that the reader is already familiar with IBM Spectrum Virtualize Real-time Compression technology. Information on this technology can be found in many sources, including the following publications:

- ▶ *IBM Real-time Compression in IBM SAN Volume Controller and IBM Storwize V7000*, REDP-4859
- ▶ *Implementing IBM Real-time Compression in SAN Volume Controller and IBM Storwize V7000*, TIPS1083

This chapter includes the following sections:

- ▶ Evaluate compression savings using Comprestimator
- ▶ Evaluate workload using Disk Magic
- ▶ Verify available CPU resources
- ▶ Configure a balanced system
- ▶ Standard benchmark tools
- ▶ Compression with FlashCopy
- ▶ Compression with Easy Tier
- ▶ Compression on the backend
- ▶ Migrating generic volumes
- ▶ Mixed volumes in the same MDisk group

11.1 Evaluate compression savings using Comprestimator

Before you use Real-time Compression technology, it is important to understand the typical workloads you have in your environment. You need to determine whether these workloads are a good candidate for compression. You should then plan to implement workloads that are suitable for compression.

To determine the compression savings you are likely to achieve for the workload type, IBM has developed an easy-to-use utility called IBM Comprestimator. The utility uses advanced mathematical and statistical algorithms to perform the sampling and analysis process in a short and efficient way. The utility also displays its accuracy level by showing the maximum error range of the results based on the internal formulas. The utility performs only read operations, so it has no effect on the data that is stored on the device.

From IBM Spectrum Virtualize version 7.6, the Comprestimator utility can be used directly from the IBM Spectrum Virtualize shell. Example 11-1 shows the CLI commands to use the utility.

Example 11-1 Estimating compression savings from the CLI

```
IBM_Storwize:Spectrum_Virtualize_Cluster:user>analyzevdisk 0
IBM_Storwize:Spectrum_Virtualize_Cluster:user>lsvdiskanalysisprogress
vdisk_count pending_analysis estimated_completion_time
1           1                161014214700
IBM_Storwize:Spectrum_Virtualize_Cluster:user>lsvdiskanalysis -nohdr
0 vdisk0 sparse 161014214659 100.00GB 0.00MB 0.00MB 0 0.00MB 0.00MB 0 0.00MB 0 0
```

From IBM Spectrum Virtualize V7.7, the Comprestimator utility can be used directly from the IBM Spectrum Virtualize GUI. Figure 11-1 on page 383 shows how to start a system-wide analysis of compression estimates by clicking **Volumes** → **Actions** → **Space Savings** → **Estimate Compression Savings**.

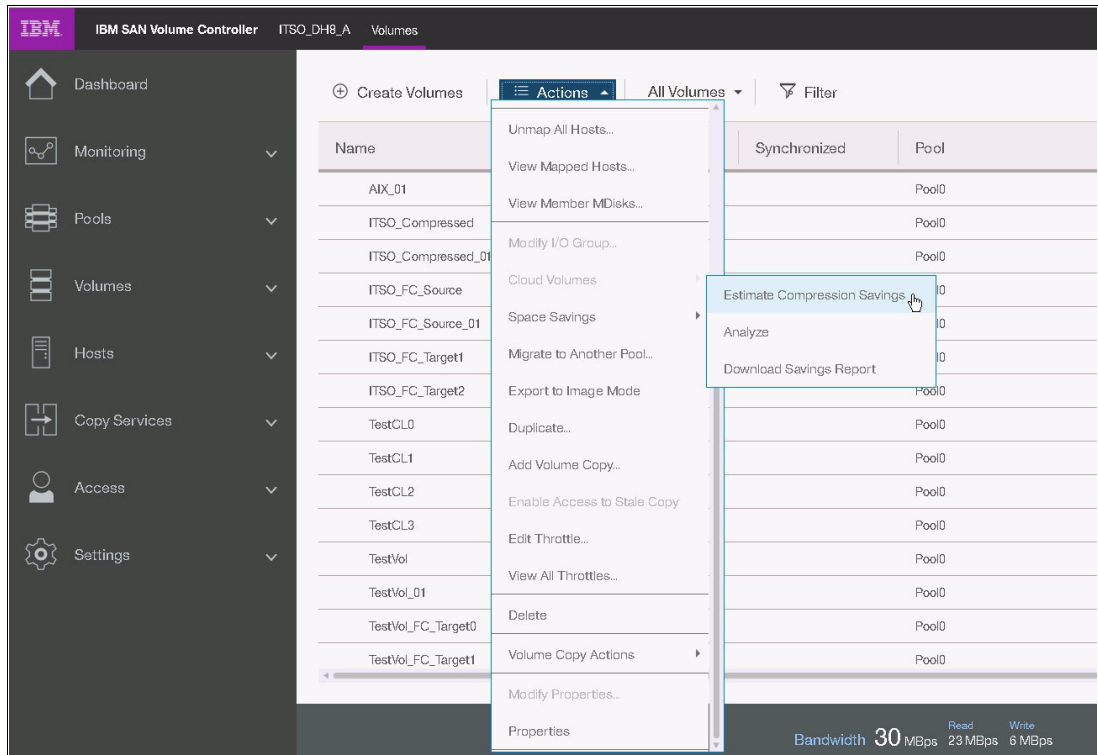


Figure 11-1 Estimating compression savings from the GUI

If using an older IBM Spectrum Virtualize version or if you want to estimate the compression savings of a different storage system before changing to IBM Spectrum Virtualize, the Comprestimator utility can be installed on a host that has access to the devices that are analyzed. More information together with the latest version can be found at this website:

<http://www.ibm.com/support/docview.wss?uid=ssg1S4001012>

These are the preferred practices for using Comprestimator:

- ▶ Run the Comprestimator utility before you implement an IBM Spectrum Virtualize solution and before you implement the Real-time Compression technology.
- ▶ Download the latest version of the utility from IBM if you are not using the version included with IBM Spectrum Virtualize.
- ▶ Use Comprestimator to analyze volumes that contain as much active data as possible rather than volumes that are mostly empty. This technique increases the accuracy level and reduces the risk of analyzing old data that is deleted but might still have traces on the device.

Note: Comprestimator can run for a long period (a few hours) when it is scanning a relatively empty device. The utility randomly selects and reads 256 KB samples from the device. If the sample is empty (that is, full of null values), it is skipped. A minimum number of samples with actual data are required to provide an accurate estimation.

When a device is mostly empty, many random samples are empty. As a result, the utility runs for a longer time as it tries to gather enough non-empty samples that are required for an accurate estimate. If the number of empty samples is over 95%, the scan is stopped.

- Use Table 11-1 thresholds for volume compressibility to determine whether to compress a volume.

Table 11-1 Thresholds for Real-time Compression implementation

	Data Compression Rate	Recommendation
On products that have Quick Assist compression acceleration cards installed and are on version 7.4 and later	> 40% compression savings	Use compression
	< 40% compression savings	Evaluate workload
On all other products	> 25% compression savings	Use compression
	< 25% compression savings	Evaluate workload

11.2 Evaluate workload using Disk Magic

Proper initial sizing greatly helps to avoid future sizing problems. Disk Magic is one such tool that is used for sizing and modeling storage subsystems for various open systems environments and various IBM platforms. It provides accurate performance and capacity analysis and planning for IBM Spectrum Virtualize products, other IBM storage solutions, and other vendors' storage subsystems. Disk Magic provides in-depth environment analysis, and is an excellent tool to estimate the performance of a system that is running Real-time Compression.

If you are an IBM Business Partner, more information about Disk Magic, and the latest version, can be found at this website:

<http://www.ibm.com/partnerworld/wps/servlet/ContentHandler/SSPQ048068H83479I86>

If you are an IBM customer, ask an IBM representative to evaluate the workload of your storage environment when implementing an IBM Spectrum Virtualize Real-time Compression solution.

11.3 Verify available CPU resources

Before compression is enabled on IBM Spectrum Virtualize systems, measure the current system utilization to ensure that the system has the CPU resources that are required for compression.

Compression is recommended for an I/O Group if the sustained CPU utilization is *below* the per-node values that are listed in Table 11-2. For node types for which the value listed is N/A, Real-time Compression can be implemented with no consideration regarding CPU utilization. This is because these node types have dedicated CPU resources for Real-time Compression.

Table 11-2 CPU resources recommendations

SAN Volume Controller					Storwize			IBM Spectrum Virtualize Software
CF8 & CG8 (4 core)	CG8 (6 core)	CG8 (12 core)	DH8 (Dual CPU)	SV1	V5030	V7000 Gen1	V7000 Gen2/Gen2+	
25%	30%	N/A	N/A	N/A	30%	25%	50%	30%

If any node in a particular I/O Group already has sustained processor utilization greater than the values in Table 11-2, do not create compressed volumes in this I/O Group. Doing so might affect existing non-compressed volumes that are owned by this I/O Group. If it is an option, add more I/O groups. If you have any questions, speak to your IBM representative.

Customers who are planning to use Real-time Compression on 6-core SAN Volume Controller CG8 nodes should enhance their system with more CPU and cache memory resources that are dedicated to Real-time Compression. This upgrade preserves full performance and resources for non-compressed workloads. Information about upgrading to SAN Volume Controller CG8 dual CPU model is available with RPQ #8S1296.

Customers who are planning to use Real-time Compression on V7000 Gen2/Gen2+ should install the extra Quick Assist compression acceleration card per node canister for better performance.

Note: To use the Real-time Compression feature on SAN Volume Controller DH8 and SV1 nodes, at least one Quick Assist compression acceleration card is required. To use the IBM Real-time Compression feature on the V9000 system, both Quick Assist compression acceleration cards are required.

11.4 Configure a balanced system

In a system with more than one I/O group, it is important to balance the compression workload. Consider a four-node (two I/O groups) IBM Spectrum Virtualize system with the following configuration:

- ▶ iogrp0: nodes 1 and 2 with 18 compressed volumes
- ▶ iogrp1: nodes 3 and 4 with two compressed volumes

This setup is not ideal, because CPU and memory resources are dedicated for compression use in all four nodes. However, in nodes 3 and 4, this allocation is used only for serving two volumes out of a total of 20 compressed volumes. The following preferred practices in this scenario should be used:

- ▶ Alternative 1: Migrate all compressed volumes from iogrp1 to iogrp0 when there are only a few compressed volumes (that is, 10 - 20).
- ▶ Alternative 2: Migrate compressed volumes from iogrp0 to iogrp1 and balance the load across nodes when there are many compressed volumes (that is more than 20).

Table 11-3 shows the load distribution for each alternative.

Table 11-3 Load distribution

	node1 volumes	node2 volumes	node3 volumes	node4 volumes
Original setup	9 compressed X non-compressed	9 compressed X non-compressed	1 compressed X non-compressed	1 compressed X non-compressed
Alternative 1	10 compressed X non-compressed	10 compressed X non-compressed	X non-compressed	X non-compressed
Alternative 2	5 compressed X non-compressed	5 compressed X non-compressed	5 compressed X non-compressed	5 compressed X non-compressed

11.5 Standard benchmark tools

Traditional block and file-based benchmark tools (such as IOmeter, IOzone, dbench, and fio) that generate truly random but not realistic I/O patterns do not run well with Real-time Compression.

These tools generate synthetic workloads that do not have any temporal locality. Data is not read back in the same (or similar) order in which it was written. Therefore, it is not useful to estimate what your performance looks like for an application with these tools. Consider what data a benchmark application uses. If the data is already compressed or it is all binary zero data, the differences that are measured are artificially bad or good, based on the compressibility of the data. The more compressible the data, the better the performance.

11.6 Compression with FlashCopy

By using the FlashCopy function of IBM Storage Systems, you can create a point-in-time copy of one or more volumes. You can use FlashCopy to solve critical and challenging business needs that require duplication of data on your source volume. Volumes can remain online and active while you create consistent copies of the data sets.

Follow these general guidelines:

- ▶ Consider configuring FlashCopy targets as non-compressed volumes. In some cases, the savings are not worth the other resources that are required because the FlashCopy target holds only the “split” grains that are backing the grains that were changed in the source. Therefore, total FlashCopy target capacity is a fraction of the source volume size.
- ▶ FlashCopy default grain size is 256 KB for non-compressed volumes and 64 KB for compressed volumes (new defaults from version 6.4.1.5 and 7.1.0.1 and later). Use the default grain size for FlashCopy with compressed volumes (64 KB) because this size reduces the performance effect when compressed FlashCopy targets are used.
- ▶ Consider the use of the background copy method. There are two ways to use FlashCopy: With or without background copy. When it is used without background copy, the host I/O is pending until the split event is finished (copy-on-write process). For example, if the host sends a 4 KB write, this I/O waits until the corresponding grain (64 KB or 256 KB) is read and decompressed.

It is then written to FlashCopy target copy. This configuration adds latency to every I/O. When background copy is used, all the grains are copied to the FlashCopy target right after the FlashCopy mapping is created. Although the configuration adds latency during the copy, it eliminates latency after the copy is complete.

11.7 Compression with Easy Tier

IBM Easy Tier is a performance function that automatically and non disruptively migrates frequently accessed data from magnetic media to solid-state drives (SSDs). In that way, the most frequently accessed data is stored on the fastest storage tier and the overall performance is improved.

Beginning with version 7.1, Easy Tier supports compressed volumes. A new algorithm is implemented to monitor read operations on compressed volumes instead of reads and writes. The extents with the most read operations that are smaller than 64 KB are migrated to SSD MDisks.

As a result, frequently read areas of the compressed volumes are serviced from SSDs. Easy Tier on non-compressed volumes operates as before and it is based on read and write operations that are smaller than 64 KB.

For more information about implementing IBM Easy Tier with IBM Real-time Compression, see *Implementing IBM Easy Tier with IBM Real-time Compression*, TIPS1072.

11.8 Compression on the backend

If you have an IBM Spectrum Virtualize system setup with some backend storage that supports compression (such as a Storwize product) and you plan to implement compression, configure compression volumes on the IBM Spectrum Virtualize system, not on the backend storage. This configuration minimizes I/O to the backend storage.

From version 7.3, the existence of a lower-level write cache below the Real-time Compression component in the software stack allows for the coalescing of compressed writes. As a result, an even bigger reduction in back-end I/Os is achieved because of the ability to perform full-stride writes for compressed data.

Note: These recommendations do not apply with backend controller with inherent compression mechanisms like IBM FlashSystem A9000 series systems.

In any case do not enable compression on both IBM Spectrum Virtualize and backend storage systems.

11.9 Migrating generic volumes

It is possible to migrate non-compressed volumes, both generic (fully allocated) or thin-provisioned, to compressed volumes by using volume mirroring. When migrating generic volumes that are created without initial zero formatting, extra considerations need to be taken into account. These volumes might contain traces of old data at the block device level. Such data is not accessible or viewable in the file system level. However, it might affect compression ratios and system resources during and after migration.

When using the Comprestimator utility to analyze such volumes, the expected compression results reflect the compression rate for all the data in the block device level. This data includes the old data. This block device behavior is limited to generic volumes, and does not occur when using Comprestimator to analyze thin-provisioned volumes.

The second issue is that old data is also compressed. Therefore, system resources and system storage space are wasted on compression of old data that is effectively inaccessible to users and applications.

Note: Regardless of the type of block device that is analyzed or migrated, it is also important to understand a few characteristics of common file systems space management.

When data is deleted from a file system, the space that it occupied before it was deleted is freed and available to the file system. It is available even though the data at block device level was not deleted. When using Comprestimator to analyze a block device or when migrating a volume that is used by a file system, all underlying data in the device is analyzed or migrated regardless of whether this data belongs to files that were deleted from the file system. This process affects even thin-provisioned volumes.

There is not a solution for existing generic volumes that were created without initial zero formatting. Migrating these volumes to compressed volumes might still be a good option and should not be discarded.

As a preferred practice, always format new volumes during creation. This process zeros all blocks in the disks and eliminates traces of old data. This is the default behavior from version 7.7.

11.10 Mixed volumes in the same MDisk group

Note: IBM Spectrum Virtualize version 7.3 onwards include a new cache architecture that is not affected by mixing compressed and non-compressed volumes in the same MDisk group. The following recommendation only applies to version 7.2 and earlier.

Consider a scenario in which hosts are sending write I/Os. If the response time from the backend storage increases above a certain level, the cache destaging to the entire pool is throttled down and the cache partition becomes full. This situation occurs under the following circumstances:

- ▶ In Storwize V7000: If the backend is HDD and latency is greater than 300 ms.
- ▶ In Storwize V7000: If the backend is SSD and latency is greater than 30 ms.
- ▶ In SAN Volume Controller: If the latency is greater than 30 ms.

From version 6.4.1.5 to 7.2, the following thresholds changed for both Storwize V7000 and SAN Volume Controller:

- ▶ For pools containing only compressed volumes, the threshold is 600 ms.
- ▶ For mixed pools, issue the following command to change to 600 ms system-wide:
`chsystem -compressiondestagemode on`

To check the current value, issue these commands:

```
lssystem | grep compression_destage  
compression_destage_mode on
```

With the new threshold, the compression module receives more I/O from cache, which improves the overall situation.

With V7.1 and later, performance improvements were made that reduce the probability of a cache throttling situation. However, in heavy sequential write scenarios, this behavior of full cache can still occur and the parameter that is described in this section can help to solve this situation. If none of these options help, separate compressed and non-compressed volumes to different storage pools. The compressed and non-compressed volumes do not share the cache partition, and so the non-compressed volumes are not affected.



A

IBM i considerations

IBM Storwize Family is an excellent storage solution for midrange and high-end IBM i customers. IBM SAN Volume Controller provides virtualization of different storage systems to an IBM i customer. SAN Volume Controller and Storwize enable IBM i installations for business continuity solutions that are extensively used.

This appendix provides preferred practice and guidelines for implementing the Storwize family and SAN Volume Controller with IBM i.

This appendix includes the following sections:

- ▶ IBM i Storage management
- ▶ Single-level storage
- ▶ IBM i response time
- ▶ Planning for IBM i capacity
- ▶ Connecting SAN Volume Controller or Storwize to IBM i
- ▶ Setting of attributes in VIOS
- ▶ Disk drives for IBM i
- ▶ Defining LUNs for IBM i
- ▶ Data layout
- ▶ Fibre Channel adapters in IBM i and VIOS
- ▶ Zoning SAN switches
- ▶ IBM i Multipath
- ▶ Boot from SAN
- ▶ IBM i mirroring
- ▶ Copy services considerations
- ▶ HyperSwap considerations

IBM i Storage management

When you are planning and implementing SAN Volume Controller and Storwize for an IBM i host, you must consider the way IBM i manages the available disk storage. Therefore, this section provides a short description of IBM i Storage management.

Many host systems require you to take responsibility for how information is stored and retrieved from the disk units. You must also provide the management environment to balance disk usage, enable disk protection, and maintain balanced data that is spread for optimum performance.

The IBM i host is different in that it takes responsibility for managing the information in IBM i disk pools, which are also called auxiliary storage pools (ASPs). When you create a file, you do not assign it to a storage location. Instead, the IBM i system places the file in the location that ensures the best performance from an IBM i perspective.

IBM i Storage management function normally spreads the data in the file across multiple disk units (LUNs when external storage is used). When you add more records to the file, the system automatically assigns more space on one or more disk units or LUNs.

Single-level storage

IBM i uses a single-level storage, object-orientated architecture. It sees all disk space and the main memory as one storage area, and uses the same set of virtual addresses to cover main memory and disk space. Paging of the objects in this virtual address space is performed in 4 KB pages.

Single-level storage makes main memory work as a large cache. Reads are done from pages in main memory, and requests to disk are done only when the needed page is not there. Writes are done to main memory, and write operations to disk are performed only as a result of swap or file close. Therefore, application response time depends not only on disk response time, but on many other factors. These factors include how large the IBM i storage pool is for the application, how frequently the application closes files, and whether it uses journaling.

IBM i response time

IBM i IT Centers are usually concerned about the following types of performance:

- ▶ **Application response time:** The response time of an application transaction. This time is usually critical for the customer.
- ▶ **Duration of batch job:** Batch jobs are usually run during the night. The duration of a batch job is critical for the customer because it must be finished before regular daily transactions start.
- ▶ **Disk response time:** Disk response time is the time that is needed for a disk I/O operation to complete. It includes the service time for actual I/O processing and the wait time for potential I/O queuing on the IBM i host. Disk response time significantly influences both application response time and the duration of a batch job.

Planning for IBM i capacity

To correctly plan the disk capacity virtualized by SVC or Storwize disk capacity for IBM i, you must be aware of IBM i block translation for external storage formatted in 512-byte blocks. IBM i disks have a block size of 520 bytes. SVC and Storwize are formatted with a block size of 512 bytes, so a translation or mapping is required to attach these to IBM i. IBM i performs the following change of the data layout to support 512-byte blocks (sectors) in external storage: For every page (8 * 520 byte sectors), it uses an extra ninth sector.

The page stores the 8-byte headers of the 520-byte sectors in the ninth sector, and therefore changes the previous 8* 520-byte blocks to 9* 512-byte blocks. The data that was previously stored in 8 * sectors is now spread across 9 * sectors, so the required disk capacity on SVC or Storwize is 9/8 of the IBM i usable capacity. Similarly, the usable capacity in IBM i is 8/9 of the allocated capacity in these storage systems.

Therefore, when attaching an SVC or Storwize to IBM i, you should have extra capacity on the storage subsystem so that the 8/9ths of the effective storage capacity that is available to IBM i covers the needs of the IBM i workload.

The performance impact of block translation in IBM i is very small or negligible.

Connecting SAN Volume Controller or Storwize to IBM i

SAN Volume Controller or Storwize V7000 can be attached to IBM i in the following ways:

- ▶ Native connection without the use of Virtual I/O Server (VIOS)
- ▶ Connection with VIOS in NPIV mode
- ▶ Connection with VIOS in virtual SCSI mode

This section describes the guidelines and preferred practices for each type of connection.

Note: For updated and detailed information about the current requirements, see the IBM System Storage Interoperation Center (SSIC):

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

Additionally, see the IBM i POWER® External Storage Support Matrix Summary:

<https://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS4563>

Native connection

Native connection requires that IBM i logical partition (LPAR) resides in POWER7® or later. It also requires IBM i level V7.1, Technology Release (TR) 7 or later when implemented in POWER7, and it requires IBM i level V7.1 TR 8 or later when in POWER8®.

Native connection *with SAN switches* can be done with:

- ▶ 4 Gb Fibre Channel (FC) adapters feature number #5774 or #5276
- ▶ 8 Gb FC adapters feature number #5735 or #5273
- ▶ 16 Gb FC adapters feature number EN0A or EN0B

Direct native connection *without SAN switches* can be done with these adapters:

- ▶ 4 Gb FC adapters in IBM i connected to 8 Gb adapters in FlashSystem V9000
- ▶ 16 Gb adapters in IBM i connected to 16 Gb adapters in FlashSystem V9000

For both resiliency and performance reasons, connect SVC or Storwize to IBM i with Multipath using two or more FC adapters:

- ▶ You can define a maximum of 127 active paths (127 LUNs) to 16 Gb ports in IBM i, with IBM i V7.2 Technology Refresh (TR) 7 or later, and with IBM i V7.3 TR3 or later.
- ▶ You can define a maximum of 64 active paths (64 LUNs) to 16 Gb ports with IBM i release and TR lower than V7.2 TR7 and V7.3 TR3.
- ▶ You can define a maximum of 64 active paths (64 LUNs) to 4 Gb port or 8 Gb port, regardless of the IBM i level.

The LUNs will report in IBM i as disk units with type 2145.

IBM i enables SCSI command tag queuing in the LUNs from natively connected SVC or Storwize V7000. The queue depth on a LUN with this type of connection is 16.

Connection with VIOS_NPIV

Connection with VIOS_NPIV requires that IBM i partition is in POWER6® server or later. It requires IBM i V7.1 TR6 or later when implemented in POWER6 or in POWER7, and it requires IBM i level V7.1 TR8 or later when in POWER8. This type of connection requires switches that must be NPIV enabled.

For both resiliency and performance reasons, connect SVC or Storwize to IBM i in Multipath using two or more VIOS.

Observe the following rules for mapping server virtual FC adapters to the ports in VIOS when implementing NPIV connection:

- ▶ Map a maximum of one virtual FC adapter from an IBM i LPAR to a port in VIOS.
- ▶ You can map up to 64 virtual FC adapters each from another IBM i LPAR to the same port in VIOS.
- ▶ You can use the same port in VIOS for both NPIV mapping and connection with VIOS virtual SCSI (VSCSI).
- ▶ If PowerHA solutions of IBM i independent auxiliary storage pool (IASP) is implemented, you need to map the virtual FC adapter of the system disk pool to a different port than the virtual FC adapter of the IASP.

You can define a maximum of 127 active paths (127 LUNs) to a virtual FC adapter with IBM i release 7.2 TR7 or later, and with IBM i release 7.3 TR3 or later.

You can define a maximum of 64 active paths (64 LUNs) to a virtual FC adapter with IBM i release and TR lower than 7.2 TR7 and 7.3 TR3.

The LUNs report in IBM i as disk units with type 2145.

IBM i enables SCSI command tag queuing in the LUNs from VIOS_NPIV connected to FlashSystem V9000. The queue depth on a LUN with this type of connection is 16.

Connection with VIOS virtual SCSI

Connection in VIOS VSCSI mode requires that IBM i partition is in POWER6 server or later. This type of connection requires IBM i V6.1.1 or later when IBM i is in POWER6 or POWER7, and it requires IBM i V7.1 TR8 or later when in POWER8.

Using Multipath with two or more VIOS improves resiliency and performance. When implementing Multipath with this type of connection, remember the following considerations:

- ▶ IBM i Multipath is performed with two or more VSCSI adapters, each of them assigned to a server VSCSI adapter in different VIOS. An hdisk from each VIOS is assigned to the relevant server VSCSI adapters. The hdisk in each VIOS represents the same SVC/Storwize LUN.
- ▶ In addition to IBM i Multipath, also implement Multipath in each VIOS by using one of the Multipath drives, preferably SDDPCM driver. The paths that connect adapters in VIOS to the LUNs in SVC/Storwize are managed by VIOS Multipath driver.

It is possible to connect up to 4095 LUNs per target, and up to 510 targets per port in a physical adapter in VIOS. With IBM i release 7.2 and later, you can attach a maximum of 32 LUNs to a port in the virtual SCSI adapter in IBM i. With IBM i releases before 7.2, a maximum of 16 LUNs can be attached to a port in the IBM i virtual SCSI adapter. The LUNs report in IBM i as disk units of type 6B22.

IBM i enables SCSI command tag queuing in the LUNs from VIOS VSCSI connected to FlashSystem V9000. The queue depth on a LUN with this type of connection is 32.

Setting of attributes in VIOS

This section describes the values of certain attributes in VIOS that must be set up for Multipath, or should be set up for best performance.

FC adapter attributes

With either VIOS Virtual SCSI connection or NPIV connection, specify the following attributes for each SCSI I/O Controller Protocol Device (fscsi) device that connects an SVC or Storwize LUN for IBM i:

- ▶ The attribute `fc_err_recov` should be set to `fast_fail`
- ▶ The attribute `dyntrk` should be set to `yes`

The specified values for the two attributes are related to how AIX FC adapter driver or AIX disk driver handle a certain type of fabric-related errors. Without setting these values for the two attributes, the way to handle the errors is different, and will cause unnecessary retries.

Disk device attributes

With VIOS Virtual SCSI connection, specify the following attributes for each hdisk device that represents an SVC or Storwize LUN connected to IBM i:

- ▶ If Multipath with two or more VIOS is used, the attribute `reserve_policy` should be set to `no_reserve`.
- ▶ The attribute `queue_depth` should be set to 32.
- ▶ The attribute `algorithm` should be set to `load_balance`.

Setting `reserve_policy` to `no_reserve` is required to be set in each VIOS if Multipath with two or more VIOS is implemented, to remove SCSI reservation on the hdisk device.

Set `queue_depth` to 32 for performance reasons. Setting this value ensures that the maximum number of I/O requests that can be outstanding on a hdisk in the VIOS at a time matches the maximum number of 32 I/O operations that IBM i operating system allows at a time to one VIOS VSCSI-connected LUN.

Set `algorithm` to `load_balance` for performance reasons. Setting this value ensures that the SDDPCM driver in VIOS balances the I/O across available paths to Storwize or SVC.

Disk drives for IBM i

This section describes how to implement internal disk drives in Storwize, and background storage of SVC and Storwize for IBM i host. These suggestions are based on the characteristics of a typical IBM i workload, such as relatively high write ratio, and small degree of skew due to spreading the objects by IBM i storage management.

We expect that most of the storage configurations for IBM i will include both hard disk drives (HDDs) and solid-state drives (SSDs) or Flash cards, either as Storwize internal drives or in the storage systems attached to SVC or Storwize.

Make sure that a sufficiently large part of disk capacity resides on SSD or Flash cards. Generally, have at least 20% of IBM i capacity on Flash technology; for good performance we recommend 50% of capacity on SSD or Flash cards.

Even if certain parts of IBM i capacity resides on SSD or Flash cards, it is important that you provide a sufficient number of HDDs with high rotation. We recommend to use 15 K rpm HDDs of 300 GB or 600 GB capacity, along with Flash technology.

The IBM i workload usually achieves the best performance when using disk capacity entirely from SSD or Flash cards.

Exploitation of SSDs and Flash cards with SVC or Storwize is achieved through Easy Tier. Even if you do not plan to install Flash storage or SSD, you can still use Easy Tier to evaluate your workload and provide information on the benefit you might gain by adding Flash technology in the future.

Use Disk Magic modeling before implementing a certain disk configuration for IBM i. In Disk Magic, enter the current performance data of the IBM i workload, then enter the planned configuration. When modeling Easy Tier, specify the lowest skew level for IBM i workload.

Disk Magic provides the predicted disk response time of IBM i on the planned disk configuration and the response time at workload growth.

Defining LUNs for IBM i

LUNs for IBM i host are defined from block-based storage. Create them the same way as for open hosts. The minimum size of an IBM i LUN is 180 MB. This setting provides 160 MB to IBM i due to block translation. The maximum size is up to 2.25 TB (excluding 2.25 TB itself), which provides up to 2 TB to IBM i.

In general, the more LUNs that are available to IBM i, the better the performance. The following are the reasons for this:

- ▶ If more LUNs are attached to IBM i, the storage management uses more threads and therefore enables better performance.
- ▶ The wait time component of disk response time is lower when more LUNs are used, resulting in lower latency of disk I/O operations.

However, the higher number of LUNs drives the requirement for more FC adapters on IBM i due to the addressing restrictions of IBM i if you are using native attachment. With VIOS attached IBM i, a larger number of LUNs brings extra complexity in implementing and management.

The sizing process determines the optimal number of LUNs required to access the needed capacity while meeting performance objectives. Regarding both these aspects and the preferred practices, our guidelines are as follows:

- ▶ For any IBM i disk pool (ASP), define all the LUNs as the same size.
- ▶ 45 GB is the preferred minimum LUN size.
- ▶ You should not define LUNs larger than 200 GB.
- ▶ A minimum of 8 * LUNs for each ASP or LPAR is preferred.

When defining LUNs for IBM i, take into account the minimum capacity for load source (boot disk) LUN:

- ▶ With IBM i release 7.1, the minimum capacity is 20 GB
- ▶ With IBM i release 7.2 before TR1, the minimum capacity is 80 GB in IBM i
- ▶ With IBM i release 7.2 TR1 and later, the minimum capacity is 40 GB in IBM i

Data layout

Spreading workloads across all Storwize or SVC components maximizes the utilization of the hardware resources in the storage subsystem. However, it is always possible when sharing resources that performance problems might arise due to contention on these resources. Isolation of workloads is most easily accomplished where each ASP or LPAR has its own managed storage pool. This configuration ensures that you can place data where you intend. I/O activity should be balanced between the two nodes or controllers on the SVC or Storwize.

Regarding this, use the following data layout:

- ▶ In Storwize with HDD, make sure that you isolate critical IBM i workloads in separate disk pools.
- ▶ In Storwize with Easy Tier on mixed HDD and SSD or Flash disk, you can share the disk pool among IBM i workloads. Only very large critical workloads should be in isolated disk pools.
- ▶ In Storwize using entirely SSD or Flash storage, you can share the disk pool among IBM i workloads.
- ▶ Avoid mixing IBM i LUNs and non-IBM i LUNs in the same disk pool.

There is also an option to create a disk pool of SSD in Storwize or SVC, and create an IBM i ASP that uses disk capacity from the SSD pool. The applications that run in that ASP will experience a performance boost.

IBM i data relocation methods, such as ASP balancing and Media preference, are not available to use with SSDs in Storwize or SVC.

Fibre Channel adapters in IBM i and VIOS

The following Fibre Channel adapters are used in IBM i when connecting Storwize or SVC in native mode:

- ▶ 16 Gb PCIe2 Dual Port FC adapter feature number EN0A, or feature number EN0B (Low Profile)
- ▶ 8 Gb PCIe Dual Port Fibre Channel Adapter feature number 5735, or feature number 5273 (Low Profile)

For VIOS_NPIV connection, use the following FC adapters in VIOS:

- ▶ 16 Gb PCIe2 Dual Port FC adapter feature number EN0A, or feature number EN0B (Low Profile)
- ▶ 8 Gb PCIe Dual Port Fibre Channel Adapter feature number 5735 or feature number 5273 (Low Profile)
- ▶ 8 Gb PCIe2 2-Port Fibre Channel Adapter feature number EN0G, or feature number EN0F (Low Profile)
- ▶ 8 Gb PCIe2 4-Port Fibre Channel Adapter feature number 5729
- ▶ 8 Gb PCIe2 4-port Fibre Channel Adapter feature number EN12
- ▶ 8 Gb PCIe2 4-port Fibre Channel Adapter feature number EN0Y (Low Profile)

Note: For updated and detailed information about the current requirements, see the IBM System Storage Interoperation Center (SSIC):

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

Additionally, see the IBM i POWER External Storage Support Matrix Summary:

<https://www.ibm.com/support/techdocs/atmastr.nsf/WebIndex/PRS4563>

When you size the number of FC adapters for an IBM i workload for native or VIOS_NPIV connection, take into account the maximum I/O rate (IOPS) and data rate (MBps) that a port in a particular adapter can sustain at 70% utilization. Also take into account the I/O rate and data rate of the IBM i workload.

If multiple IBM i partitions connect through the same port in VIOS_NPIV, take into account the maximum rate at the port at 70% utilization and the sum of I/O rates and data rates of all connected LPARs.

For sizing, you might consider the throughput specified in Table A-1 that shows the throughput of a port in a particular adapter at 70% utilization.

Table A-1 Throughput of Fibre Channel adapters

Maximal I/O rate per port	16 Gb 2-port adapter	8 Gb 2-port adapter
IOPS per port	52,500 IOPS	23,100 IOPS
Sequential throughput per port	1,330 MBps	770 MBps
Transaction throughput per port	840 MBps	371 MBps

Zoning SAN switches

With native connection and the connection in VIOS_NPIV, zone the switches so that one worldwide port name (WWPN) of one IBM i port is in a zone with two ports of Storwize or SVC, each port from one node canister. This technique ensures resiliency for the I/O to and from a LUN assigned to that WWPN. If the preferred node for that LUN fails, the I/O rate continues using the non-preferred node.

Note: In an SVC Split Cluster configuration, you might need to create two zones, each containing IBM i port and one port from SVC, that overlap on the IBM i port.

When connecting with VIOS virtual SCSI, zone one physical port in VIOS with all available ports in SVC or Storwize, or with as many ports as possible to allow load balancing. Keep in mind that there are a maximum of eight paths available from VIOS to SVC or Storwize. SVC or Storwize ports that are zoned with one VIOS port should be evenly spread between the node canisters.

IBM i Multipath

Multipath provides greater resiliency for SAN-attached storage. IBM i supports up to eight paths to each LUN. In addition to the availability considerations, lab performance testing has shown that two or more paths provide performance improvements when compared to a single path.

Typically two paths to a LUN is the ideal balance of price and performance. However, you can implement more than two paths for workloads where high I/O rates are expected to LUNs, or where high access density is expected, such as all SSD Storwize or SVC with attached FlashSystem as background storage. As a preferred practice, four paths are a good solution for such configurations.

Multipath for a LUN is achieved by connecting the LUN to two or more ports that belong to different adapters in IBM i partition. With native connection to Storwize or SVC, the ports for Multipath must be in different physical adapters in IBM i. With VIOS_NPIV, the virtual Fibre Channel adapters for Multipath must be assigned to different VIOS. If more than two paths are used, you can use two VIOS and split the paths among them. With VIOS VSCSI connection, the virtual SCSI adapters for Multipath must be assigned to different VIOS.

Every LUN in Storwize or SVC uses one node as the preferred node. The I/O traffic to and from the particular LUN normally goes through the preferred node. If that node fails, the I/O operations are transferred to the remaining node. With IBM i Multipath, all the paths to a LUN through the preferred node are active, and the paths through the non-preferred node are passive. Multipath employs the load balancing among the paths to a LUN that go through the node that is preferred for that LUN.

Boot from SAN

All connection options, Native, VIOS_NPIV, and VIOS Virtual SCSI, support IBM i Boot from SAN. The IBM i boot disk (LoadSource) is on a Storwize or SVC LUN that is connected the same way as the other LUNs. There are not any special requirements for LoadSource connection. When installing IBM i operating system with disk capacity on Storwize or SVC, the installation prompts you to select one of the available LUNs for the LoadSource.

IBM i mirroring

Some clients prefer to have additional resiliency with IBM i mirroring functions. For example, they use mirroring between two Storwize or SVC systems, each connected with one VIOS.

When starting mirroring with VIOS connected Storwize or SVC, you should add the LUNs to the mirrored ASP in steps:

1. Add the LUNs from two virtual adapters with each adapter connecting one to-be mirrored half of LUNs.
2. After mirroring is started for those LUNs add the LUNs from the two new virtual adapters, each adapter connecting one to-be mirrored half, and so on. This way, you ensure that the mirroring is started between the two SVC or Storwize and not among the LUNs in the same SVC.

Copy services considerations

Storwize or SVC supports both synchronous replication (Metro Mirror) and asynchronous replication (Global Mirror). It provides two options for Global Mirror: Standard Global Mirror, and the Change Volumes enhancement that allows for a flexible and configurable RPO that allows GM to be maintained during peak periods of bandwidth constraint.

You must size the bandwidth of Metro Mirror or Global Mirror links to accommodate the peaks of IBM i workload to avoid affecting production performance.

The current zoning guidelines for mirroring installations advise that a maximum of two ports on each SVC node/Storwize V7000 node canister be used for mirroring. The remaining two ports on the node/canister should not have any visibility to any other cluster. If you have been experiencing performance issues when mirroring is in operation, implementing zoning in this fashion might help to alleviate this situation.

When planning for FlashCopy for IBM i, make sure that enough disk drives are available to the FlashCopy target LUNs to keep good performance of production IBM i while FlashCopy relationships are active. This guideline is valid for both FlashCopy with background copying and without background copying. When using FlashCopy with Thin provisioned target LUNs, make sure that there is sufficient capacity available for their growth. This amount depends on the amount of write operations to source or target LUNs.

HyperSwap considerations

SVC or Storwize HyperSwap is supported by IBM i release 7.2 TR3 or later. It is supported in native and VIOS_NPIV connection.

The HyperSwap solution with IBM i and IASP is supported by IBM i release 7.2 TR5 or later and by IBM i release 7.3 TR1 or later. With this solution you need to install PowerHA IBM SystemMirror® for i that enables LUN switching to site 2. It is supported in native and VIOS_NPIV mode.

We highly advise combining HyperSwap with IBM i multipath for the best resiliency. Zone the switches so that WWPNs from IBM i are zoned with both Storwize nodes on each site, and that switches from both fabrics are evenly used.

An example of zoning for VIOS_NPIV connection is shown on Figure 11-2.

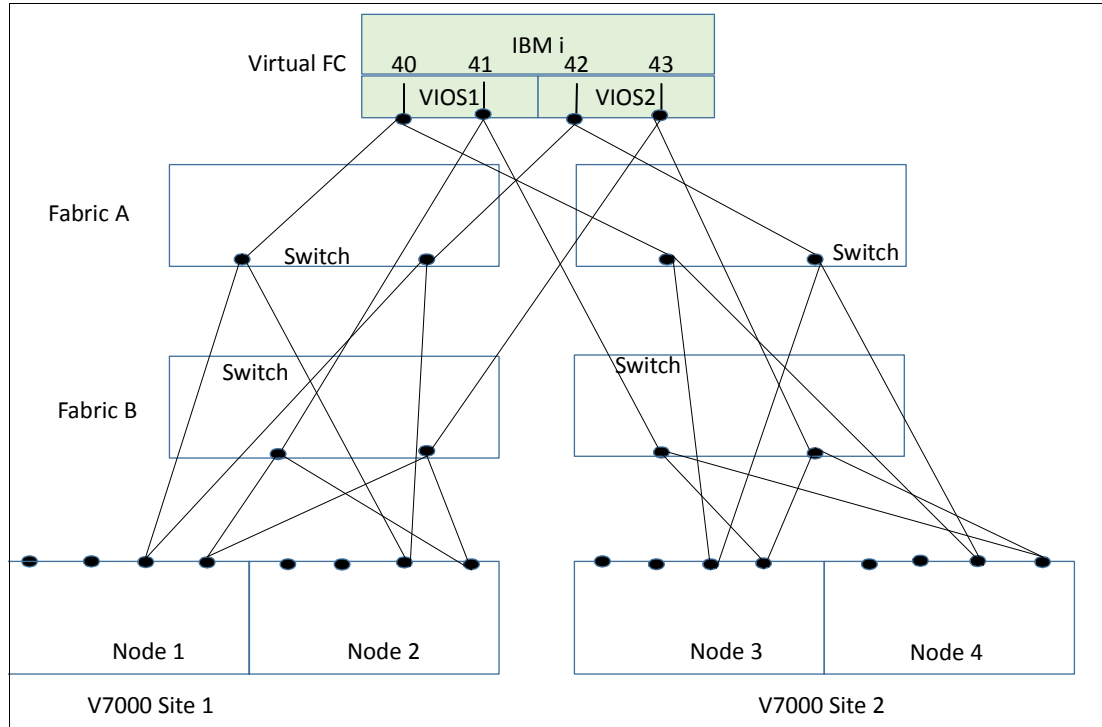


Figure 11-2 HyperSwap connection example

Next, we briefly describe some high availability solutions with HyperSwap and IBM i.

Outage of Storwize I/O group at site 1

In this scenario the entire IBM i capacity resides on HyperSwap LUNs.

After the outage occurs the I/O rate automatically transfers to the SVC or Storwize nodes at site 2. IBM i workload keeps running, and there are no relevant messages in IBM i message queues.

When the outage has finished, the IBM i I/O rate automatically transfers to nodes on site 1. IBM i workload keeps running without interruption.

Disaster at site 1

In this scenario we use a prepared IBM i logical partition in the system at site 2. The entire IBM i disk capacity is on HyperSwap LUNs. Two hosts are defined in Storwize: one host with the WWPNs of IBM i at site 1, and one with WWPNs of site 2.

After the failure of site 1 the IBM i LUNs are still available from SVC or Storwize nodes at site 2. In the HyperSwap cluster unmap the HyperSwap LUNs from the host of IBM i at site 1, map the LUNs to the host of IBM i at Site 2, and IPL IBM i at site 2. After the IPL is finished you may resume the workload on site 2.

Once the outage of site 1 is finished, power-down IBM i at site 2, unmap the IBM i LUNs from the host of site 2 and map them to the host at site 1. IPL IBM i at site 1 and resume the workload. The I/O rate will be transferred to the Storwize nodes at site 1.

Planned outage with Live Partition Mobility and Storwize HyperSwap

IBM PowerVM® Live Partition Mobility (LPM) allows you to move a running logical partition, including its operating system and running applications, from one system to another without any shutdown or without disrupting the operation of that logical partition. For more information about LPM refer to *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

In this solution we combine Live Partition Mobility with HyperSwap, to transfer the workload onto site 2 during the planned outage of site 1. This scenario requires connection with VIOS_NPIV and all IBM i LUNs as HyperSwap LUNs. For more information about LPM requirements, refer to *IBM PowerVM Virtualization Introduction and Configuration*, SG24-7940.

Before starting LPM you have to define the host in Storwize with the WWPNs of the second port of the virtual FC adapters. Specify *site 2* for this host. This way you enable the I/O rate to transfer to the nodes at site 2 when LPM is transferring the IBM i partition.

After the outage is finished, use LPM to transfer the IBM i partition to site 1. During the transfer, the I/O rate will automatically move to the nodes at site 1.

Implementation with IASP: disaster at site 1

This scenario requires PowerHA for i software installed, and the corresponding IBM i setup which consists of two IBM partitions in the cluster, IASP on the IBM i at site 1, cluster resource group, and PowerHA copy description. The workload is running in the IASP. For more information about PowerHA for i setup refer to *IBM PowerHA SystemMirror for i: Preparation (Volume 1 of 4)*, SG24-8400.

In this scenario, ensure that all IBM i LUNs, not just the IASP LUNs, are HyperSwap volumes.

If there is a disaster at site 1, PowerHA for i automatically switches the IASP to the system at site 2, and the workload can be resumed at site 2.

After the failure at site 1 is fixed, use PowerHA for i to switch IASP back to site 1 and resume the workload at this site.



Business continuity

Business continuity (BC) and continuous application availability are among the most important requirements for many organizations. Advances in virtualization, storage, and networking have made enhanced business continuity possible. Information technology solutions can now manage both planned and unplanned outages, and provide the flexibility and cost efficiencies that are available from cloud-computing models.

This chapter briefly describes the Stretched Cluster, Enhanced Stretched Cluster, and HyperSwap solutions for IBM Spectrum Virtualize, and HyperSwap configurations specific for Storwize. Technical details or implementation guidelines are not presented in this chapter because they are described in separate publications.

This appendix includes the following sections:

- ▶ Business continuity with Stretched Cluster
- ▶ Business continuity with Enhanced Stretched Cluster
- ▶ Business continuity with HyperSwap
- ▶ Third site and IP quorum
- ▶ Comparison of business continuity solutions

Business continuity with Stretched Cluster

Within standard implementations of IBM Spectrum Virtualize, all the I/O Group nodes are physically installed in the same location. To supply the different high availability (HA) needs that customers have, the *Stretched Cluster* configuration was introduced, where each node (from the same I/O Group) on the system is physically at a different site.

When implemented with mirroring technologies, such as volume mirroring, these configurations can be used to maintain access to data on the system if there are power failures or site-wide outages at different levels, SAN, back-end storage, or IBM Spectrum Virtualize nodes.

Stretched Clusters are considered high availability (BC and HA) solutions because both sites work as instances of the production environment (there is no standby location). Combined with application and infrastructure layers of redundancy, Stretched Clusters can provide enough protection for data that requires availability and resiliency.

When IBM Spectrum Virtualize was first introduced, the maximum supported distance between nodes within an I/O Group was 100 meters. With the evolution of code and introduction of new features, stretched cluster configurations have been enhanced to support distances up to 300 km. These geographically dispersed solutions leverage on specific configurations that use Fibre Channel (FC) or Fibre Channel over IP (FC/IP) switch, or Multiprotocol Router (MPR) inter-switch links (ISLs) between different locations.

Business continuity with Enhanced Stretched Cluster

IBM Spectrum Virtualize V7.2 introduced the *Enhanced Stretched Cluster* (ESC) feature that further improved the Stretched Cluster configurations. The Enhanced Stretched Cluster introduced the *site awareness* concept for nodes and external storage, and the disaster recovery (DR) feature that enables you to manage effectively rolling disaster scenarios.

Within IBM Spectrum Virtualize V7.5, the site awareness concept has been extended to hosts. This extension enables more efficiency for host I/O traffic through the SAN, and easier host path management.

Stretched Cluster and Enhanced Stretched Cluster solutions are currently the only solutions that support a three-site configuration for high availability and disaster recovery solutions.

Business continuity with HyperSwap

The *HyperSwap* high availability feature in IBM Spectrum Virtualize and Storwize enables business continuity during a hardware failure, power failure, connectivity failure, or disasters, such as fire or flooding. The HyperSwap feature is available on the IBM Spectrum Virtualize and Storwize family.

The HyperSwap feature provides highly available volumes accessible through two sites at up to 300 km apart. A fully independent copy of the data is maintained at each site. When data is written by hosts at either site, both copies are synchronously updated before the write operation is completed. The HyperSwap feature automatically optimizes itself to minimize data that is transmitted between sites, and to minimize host read and write latency.

HyperSwap has the following key features:

- ▶ Works with IBM Spectrum Virtualize and IBM Storwize V7000, V5000, and V7000 unified hardware.
- ▶ Uses intra-cluster synchronous remote copy (named Active-Active Metro Mirror with change volumes) capabilities along with existing change volume and access I/O group technologies.
- ▶ Makes a host's volumes accessible across two IBM StorwizeV7000/V5000 or IBM Spectrum Virtualize I/O groups in a clustered system by using the Active-Active Metro Mirror relationship. The volumes appear as a single volume to the host.
- ▶ Works with the standard multipathing drivers that are available on various host types, with no additional host support required to access the highly available volume.

Third site and IP quorum

In stretched cluster or HyperSwap configurations, you must use a third, independent site to house a quorum device acting as the tie-breaker in case of split-brain scenarios. To use a quorum disk as the quorum device, this third site must use Fibre Channel connectivity together with an external storage system. Sometimes, the third site requirement turns out to be expensive in terms of infrastructure and network costs. For this reason a less demanding tie-breaking solution, based on Java application, has been introduced with V7.6.

To use an IP-based quorum application as the quorum device for the third site, no Fibre Channel connectivity is used. Java applications are run on hosts at the third site. However, there are strict requirements on the IP network, with using IP quorum applications.

For stable quorum resolutions, an IP network must provide the following requirements:

- ▶ Connectivity from the hosts to the service IP addresses of all nodes. If IP quorum is configured incorrectly, the network must also deal with possible security implications of exposing the service IP addresses, because this connectivity can also be used to access the service GUI.
- ▶ Port 1260 is used by IP quorum applications to communicate from the hosts to all nodes.
- ▶ The maximum round-trip delay must not exceed 80 ms, which means 40 ms each direction.
- ▶ A minimum bandwidth of 2 MBps is ensured for node-to-quorum traffic.

Even with IP quorum applications at the third site, quorum disks at site one and site two are required because they are used to store metadata. The maximum number of applications that can be deployed is five. Currently, supported Java runtime environments (JREs) can be found at the following website:

<https://ibm.biz/BdjAsP>

For more information about IP Quorum requirements and installation, see the IP Quorum configuration section in IBM Knowledge Center:

<https://ibm.biz/BdjAr7>

Note: The IP Quorum configuration process has been integrated into the IBM Spectrum Virtualize GUI from V7.7.1 and later.

Comparison of business continuity solutions

The business continuity solutions described in this section have different characteristics both in terms of implementation and features. Table 11-4 provides a comparison of these business continuity solutions that can help to identify the most fitting solution to a specific environment and needs.

Table 11-4 Business continuity solutions comparison

	Standard Stretched Cluster	Enhanced Stretched Cluster	HyperSwap
The function is available on these products	IBM Spectrum Virtualize only	IBM Spectrum Virtualize only	<ul style="list-style-type: none"> ▶ IBM Spectrum Virtualize with two or more I/O Groups ▶ Storwize V7000 ▶ Storwize V5000 ▶ FlashSystem V9000
Complexity of configuration	Command-line interface (CLI) or graphical user interface (GUI) on a single system; simple object creation	Command-line interface (CLI) or graphical user interface (GUI) on a single system; simple object creation	Command-line interface (CLI) or graphical user interface (GUI) on a single system; simple object creation
The number of sites on which data is stored	Two	Two	Two
Distance between sites	Up to 300 km (186.4 miles)	Up to 300 km (186.4 miles)	Up to 300 km (186.4 miles)
Independent copies, which are maintained, of data	Two	Two	Two (four if you use additional Volume Mirroring to two pools in each site)
Technology for host to access multiple copies and automatically fail over	Standard host multipathing driver	Standard host multipathing driver	Standard host multipathing driver
Cache that is retained if only one site is online?	Yes, if spare node is used, no otherwise	Yes, if spare node is used, no otherwise	Yes
Host-to-storage-system path optimization	Manual configuration of preferred node	Manual configuration of preferred node for each volume before version 7.5; automatic configuration that is based on host site as HyperSwap from V7.5	Automatic configuration based on host site (requires Asymmetric Logical Unit Access (ALUA)/Target Port Group Support (TPGS) support from the multipathing driver)
Synchronization and resynchronization of copies	Automatic	Automatic	Automatic
Stale consistent data is retained during resynchronization for DR?	No	No	Yes
Scope of failure and resynchronization	Single volume	Single volume	One or more volumes; the scope is user-configurable

	Standard Stretched Cluster	Enhanced Stretched Cluster	HyperSwap
Ability to use FlashCopy with an HA solution	Yes (although no awareness of the site locality of the data)	Yes (although no awareness of the site locality of the data)	Limited: You can use FlashCopy maps with a HyperSwap Volume as a source; avoids sending data across link between sites
Ability to use Metro Mirror, Global Mirror, or Global Mirror Change Volume with an HA solution	One remote copy; it can maintain current copies on up to four sites	One remote copy; it can maintain current copies on up to four sites	No
Maximum number of highly available volumes	5,000	5,000	1,250
Minimum required paths for each logical unit (LUN) for each host port	Two	Two	Four
Minimum number of I/O Groups	One	One I/O Group is supported, but it is not recommended	Two
Rolling disaster support	No	Yes	Yes
Licensing	Included in base product	Included in base product	Requires Remote Mirroring license for volumes. Exact license requirements might vary by product.

Business continuity solutions implementation requires special considerations in the infrastructure and network setup. Throughout this book, many recommendations have been made regarding specific topics, but for a complete coverage of the implementation guidelines see *IBM Spectrum Virtualize and SAN Volume Controller Enhanced Stretched Cluster with VMware*, SG24-8211 for Enhanced Stretched Cluster and *IBM Storwize V7000, Spectrum Virtualize, HyperSwap, and VMware Implementation*, SG24-8317 for HyperSwap configurations.



Scripting examples

This topic describes some universal ways of accessing IBM Spectrum Virtualize system by using scripts. These methods can be easily applied to reporting and checking tasks. Though most of the current operating systems and almost any modern programming language have resources to accomplish this, this section concentrates mostly on UNIX/Linux environments with Bourne-again shell (bash) for simple examples, and on Python version 3 for advanced cases.

Important: You use and modify these scripts at your own risk.

This section demonstrates basic usage of the following protocols, standards, and APIs:

- ▶ Secure Shell (SSH)
- ▶ SMI-S
- ▶ HTTPS and RESTful API on IBM Spectrum Control
- ▶ HTTPS on IBM Spectrum Virtualize
- ▶ Conclusions

Secure Shell (SSH)

Secure Shell (SSH) is a network protocol for secure operations on remote services over an insecure network. IBM Spectrum Virtualize has a powerful command-line interface (CLI) which is accessible via the SSH protocol. Together with SSH, the IBM Spectrum Virtualize CLI can be used for interactive system configuration, storage administration, reporting, and automating these tasks with scripts.

Security: SSH protocol allows you to authenticate users by passwords or by asymmetric cryptography methods: SSH keys. It is strongly advised to configure and use SSH keys for security reasons.

Bash

Typically, under UNIX-like environments bash scripts are written for the default SSH client (**ssh**) installed in the system.

Most often, the SSH client is used in the form of running a specified command, but not a login shell on a remote system. Simply put, instead of opening an interactive session, SSH runs a command on the remote system, forwards its output to the local computer and finally exits. This mode allows us to use SSH in shell command substitutions or in conveyors for piping its standard output to any external program for further filtering, parsing, and data processing.

The first line of Example C-1 shows how to use SSH to run a command (**lsystemstats**) on the remote IBM SVC/Storwize system (*itso_storage*) with privileges of *itso_user*, and store its output in a local shell variable (*lsstats*) by using the shell command substitution mechanism.

The second line of the example demonstrates further data processing: **printf** command pipes contents of the *lsstats* variable through the **grep** filter to print out only VDisk-related lines.

Example: C-1 Using ssh to extract data from the storage system and save it in a variable

```
lsstats=$(ssh itso_user@itso_storage lssystemstats)
printf "$lsstats" | grep vdisk
```

If there is no need to store data for further processing, the previous code can be simplified using pipes only (Example C-2).

Example: C-2 Extracting data from the storage system to process it with a conveyor of commands

```
ssh itso_user@itso_storage lssystemstats | grep vdisk
```

For simple cases on Windows, the **plink.exe** SSH command-line client from PuTTY family utilities combined with the **findstr** command can be used to achieve similar results (Example C-3).

Example: C-3 The plink.exe client usage

```
C:\Program Files\PuTTY>plink itso_user@itso_storage lssystemstats | findstr vdisk
```

Note that another good reason to use SSH key-based authentication in scripts is automatic key submission. The SSH client takes care of it during the authentication process on the

remote system. Otherwise, with password-based methods, a user will have to enter passwords manually on every SSH connection.

However, if key-based authentication is not available in a particular environment, it is still possible to modify the bash script to wrap SSH commands by a special utility that supplies the user password automatically, this method can be potentially insecure and it is out of the intended scope of this chapter.

In general, the method of command execution remotely via SSH opens a route to write advanced scripts for fetching data from IBM Spectrum Virtualize systems and processing it on a local computer with tools and utilities available to a user in a particular environment.

Example C-4 shows a short script which endlessly (with an interval of 1 second) collects current system-level performance statistics and saves the acquired data into a CSV-formatted file.

Example: C-4 Collecting system-level performance statistic in a file

```
#!/bin/bash

user="itso_user"
target="itso_storage"
fs=","
cmd="lssystemstats -nohdr -delim "${fs}
header="Date"${fs}"Name"${fs}"Current value"
outfile="sysstats.csv"
frequency=1

printf '%s\n' "$header" > $outfile
while true; do
    ssh ${user}@${target} ${cmd} |
        gawk -F${fs} 'BEGIN {OFS=FS} {print strftime("%F %T"), $1, $2}' >> $outfile
    sleep $frequency
done
```

In the first part of the script, various variables are defined to be used later:

- ▶ `user='itso_user'`: specifies a username on a storage system
- ▶ `target='itso_storage'`: the IBM Spectrum Virtualize storage system itself
- ▶ `fs=','`: is a variable to hold a field separator, and comma sign is a good choice for formatted textual data
- ▶ `cmd="lssystemstats -nohdr -delim "${fs}`: the command to be run on the storage system
- ▶ `header="Name"${fs}"Current value"`: custom header to be printed instead of default one
- ▶ `outfile='sysstats.txt'`: the file name for output data
- ▶ `frequency=1`: a statistic collection interval in seconds

Also note, most of IBM Spectrum Virtualize CLI commands have these useful options:

- ▶ `-delim` to specify a custom field separator
- ▶ `-nohdr` to suppress headers in the output

The second part of the script is an infinite **while** loop that collects and processes data. In Example C-4, GNU implementation of awk language (`gawk`) is used for this purpose as a tool to parse data by columns and add the current timestamp into each row.

The main disadvantage of Example C-4 script is that a new SSH connection with the storage system is established on every loop iteration. It becomes painful for password-based SSH authentication as the user has to type a valid password continuously, but even with key-based authentication this is not optimal as it takes additional time and system resources to reopen connections.

Fortunately, IBM Spectrum Virtualize itself is powerful enough to run sophisticated scripts directly in the shell of the storage system (actually this shell is bash, running in restricted mode: rbash).

A slightly improved version of a system-level performance collecting script in Example C-5 shows the way to offload SSH loop from a local computer to the shell on the remote IBM Spectrum Virtualize system while post-processing performance data with gawk stays on the local computer.

Example: C-5 Moving the performance collection loop to the storage system

```
#!/bin/bash

user="itso_user"
target="itso_storage"
fs=","
cmd="lsystemstats -nohdr -delim "${fs}"
header="Date"${fs}"Name"${fs}"Current value"
outfile="sysstats.csv"
frequency=1

printf '%s\n' "$header" > $outfile
ssh ${user}@${target} "while true; do $cmd; sleep $frequency; done" |
  gawk -F${fs} 'BEGIN {OFS=FS} {print strftime("%F-%"), $1, $2}' >> $outfile
```

The advantage of this method is that there is only a single connection established with IBM Spectrum Virtualize as the **while** loop runs in the shell of the remote storage system.

Python

In Python an external module is required to connect with IBM Spectrum Virtualize via the SSH protocol. Nowadays, the most popular one is open-source paramiko, and it can be installed using pip.

Example C-6 illustrates the simplest template of paramiko usage in a Python script.

Example: C-6 Basic paramiko usage

```
import paramiko

target = 'itso_storage'
user = 'itso_user'
command = 'lsdisk'

client = paramiko.SSHClient()
client.set_missing_host_key_policy(paramiko.AutoAddPolicy())
client.connect(hostname=target, username=user)

stdin, stdout, stderr = client.exec_command(command)
data = stdout.read()
```

```

errors = stderr.read()

if data:
    print(data.decode('US-ASCII'))

if errors:
    print(errors.decode('US-ASCII'))

```

Here, the `paramiko` module is loaded in the first line of the script, and the next three lines define variables:

target	A hostname/IP address of IBM Spectrum Virtualize system. In this script it is <code>itso_storage</code> .
user	A username (<code>itso_user</code>) to log into the storage system.
command	A command (<code>lsvdisk</code>) to run on the target.

Further, `paramiko.SSHClient()` represents a session with the SSH server.

The `set_missing_host_key_policy(paramiko.AutoAddPolicy())` defines what to do if the remote system fingerprints are not known locally. Actually there are two policies: **AutoAddPolicy** and **RejectPolicy**. We chose the first one to simplify the example and to connect with the remote storage system anyway, but for security reasons in production and in advanced scripts, `load_system_host_keys()` can be used to load system host keys.

Next, `connect(hostname=target, username=user)` connects with the storage system. If keys-based authentication is configured properly, keys are checked automatically and a session with SSH server is established. There are several options available with `client.connect()`. For example, it's possible to specify certificates using `pkey` or `key_filename` arguments, and to set a user password with the `password` argument, if it is not possible to engage better authentication methods.

Therefore, in the worst, most insecure, and strongly advised-against way, the code (as it is shown in Example C-7) can even contain passwords in plain text. This is most certainly *not* recommended.

Example: C-7 Dangerous, not recommended client.connect() method invocation

```
client.connect(hostname='itso_storage', username='itso_user', password='MySecReT1')
```

To run a command on the remote system `exec_command()` method is used. It returns data from `stdin`, `stdout`, and `stderr` streams of the executed command. In the last lines of the script, the data is read from both `stdout` and `stderr` streams, and decoded into convenient US-ASCII character set to be printed by `print()`.

SMI-S

IBM Spectrum Virtualize supports the Storage Management Initiative Specification (SMI-S). It is an ISO-approved storage standard that provides access to common management functions and features of various storage systems. It is developed and maintained by the Storage Networking Industry Association (SNIA).

SMI-S is based on three main components: the Common Information Model (CIM), Web-Based Enterprise Management standards (WBEM), and Service Location Protocol (SLP).

There are a number of applications and tools that can be used to query SMI-S, but this chapter shows a simple way to reach CIM/WBEM services of the IBM SAN Volume Controller system using Python and the open-source PyWBEM module, which can be installed using `pip`.

Example C-8 demonstrates the very basics of PyWBEM module usage with SVC.

Example: C-8 Basic PyWBEM usage

```
import pywbem
import getpass

target = 'itso_storage'
url = 'https://' + target
username = 'itso_user'
password = getpass.getpass()

wbemc = pywbem.WBEMConnection(url, (username, password), 'root/ibm', no_verification=True)

cluster = wbemc.EnumerateInstances('IBMTSSVC_Cluster')
print(cluster[0].items())
```

In Example C-8, `WBEMConnection()` establishes HTTPS connection with WBEM services of IBM SAN Volume Controller. Here, the target storage system URL is specified by the `url` argument. `Username` and `password` as well as the CIM namespace (`root/ibm`) to query are also provided in the next lines.

Note, the `getpass` module is not necessary to work with SMI-S, its purpose is to securely read passwords from standard input with terminal echo function switched off to hide what is typed in. Next, `no_verification=True` argument disables SSL certificate verification in this code. In other words it forces the script to trust any certificate provided by the WBEM server.

Warning: Disabling the verification of the server SSL certificate is dangerous. It is insecure and should be avoided in production.

After connection is successfully established, instances of a given CIM class can be enumerated with `EnumerateInstances()` method, which returns a complex data structure: a list of `CIMInstance()` classes. In Example C-8, it is done with the `IBMTSSVC_Cluster` class, which represents system-level information comparable with the results of running the `lssystem` command.

There are different CIM classes available for comprehensive management of SAN Volume Controller system:

- ▶ `IBMTSSVC_Cluster`: System-level information
- ▶ `IBMTSSVC_Node`: Information about nodes
- ▶ `IBMTSSVC_ConcreteStoragePool`: MDisk groups
- ▶ `IBMTSSVC_BackendVolume`: MDisks themselves
- ▶ `IBMTSSVC_StorageVolume`: VDisk information

This document touches upon a small amount of them to illustrate SMI-S capabilities, but it does not provide a full list of these classes or their descriptions. Refer to the documentation on IBM SAN Volume Controller WBEM/CIM classes, their purposes, and relationship diagrams in IBM Knowledge Center:

<https://ibm.biz/Bdj6EW>

The last line of the script parses and prints out the data. But it is not the only way to run the job, Python is a flexible language, it allows us to do such work in different ways. Several approaches of processing the data acquired by `EnumerateInstances()` for a number of CIM classes are listed in Example C-9.

Example: C-9 Parsing EnumerateInstances() output for various classes representing Cluster, Nodes, and Storage pools

```
print('Cluster information')
cluster = wbemc.EnumerateInstances('IBMTSSVC_Cluster')
print(cluster[0]['ElementName'])
for c_prop in cluster[0]:
    print('\t{prop}: "{val}"'.format(prop=c_prop,
val=cluster[0].properties[c_prop].value))

print('Nodes information')
nodes = wbemc.EnumerateInstances('IBMTSSVC_Node')
for node in nodes:
    print(node['ElementName'])
    for n_prop in node:
        print('\t{prop}: "{val}"'.format(prop=n_prop, val=node[n_prop]))

print('Pools information')
pools = wbemc.EnumerateInstances('IBMTSSVC_ConcreteStoragePool')
print('PoolID', 'NumberOfBackendVolumes', 'ExtentSize', 'UsedCapacity',
'RealCapacity', 'VirtualCapacity', 'TotalManagedSpace', sep=',')
for pool in pools:
    print(
        pool['ElementName'], pool['NumberOfBackendVolumes'], pool['ExtentSize'],
        pool['UsedCapacity'], pool['RealCapacity'], pool['VirtualCapacity'],
        pool['TotalManagedSpace'], sep=','
    )
```

Using similar, yet different approaches, *Cluster information* and *Nodes information* sections of the example parse data in key/value pairs to show all acquired data. Alternatively, *Pools information* part filters data to print selected fields only, and it wastefully ignores all other fields.

For some classes, like `IBMTSSVC_StorageVolume`, full enumeration of all the instances can be quite slow. It can generate large amounts of unnecessary data which must be prepared by the storage system, transmitted over the network, and finally parsed by the script. Fortunately, it is possible to significantly reduce such data-flows by requesting a limited amount of necessary information only.

There is the `ExecQuery()` method, which allows us to request the WBEM server in a convenient query language, similar to SQL. Two dialects are recognized by PyWBEM:

- ▶ CIM Query Language (DMTF:CQL)
- ▶ WBEM Query Language (WQL)

Both dialects can be used with IBM SAN Volume Controller, but this chapter uses the CQL syntax in examples. DMTF specification (DSP0202) for CQL can be found here:

https://www.dmtf.org/sites/default/files/standards/documents/DSP0202_1.0.0.pdf

Example C-10 illustrates the flexibility of the `ExecQuery()` method.

Example: C-10 Querying required only data with ExecQuery() method

```
print('Vdisks:')
vdisks = wbemc.ExecQuery(
    'DMTF:CQL',
    "SELECT VolumeId, VolumeName, NumberOfBlocks FROM IBMTSSVC_StorageVolume"
    " WHERE VolumeName LIKE 'vdisk.'"
)
for vd in vdisks:
    print(vd['VolumeId'], vd['VolumeName'], vd['NumberOfBlocks'], sep=',')
```

One of the advantages of SMI-S on IBM SVC is its capability to collect performance data of various storage system components using *Statistic* family CIM classes. For example:

- ▶ IBMTSSVC_BackendVolumeStatistics
- ▶ IBMTSSVC_FCPortStatistics
- ▶ IBMTSSVC_NodeStatistics
- ▶ IBMTSSVC_StorageVolumeStatistics

A quite detailed, with commentaries, example of performance data collecting and processing script is shown in Example C-11. It works with `IBMTSSVC_StorageVolumeStatistics` to retrieve vdisk statistics.

Example: C-11 Accessing performance metrics with PyWBEM module

```
import pywbem
import getpass
import time

target = 'itso_storage'
user = 'itso_user'
password = getpass.getpass()

url = 'https://' + target
# Output field separator:
ofs = ','
header = ['InstanceID', 'ReadIOs', 'WriteIOs', 'TotalIOs', 'KBytesRead',
          'KBytesWritten', 'KBytesTransferred']
# Performance collection interval in minutes
frequency = 5

def vdisks_perf(wbem_connection, hdr):
    """Get performance statistics for vdisks"""

    # Form "select" request string
    request = "SELECT " + ','.join(hdr) + " FROM IBMTSSVC_StorageVolumeStatistics"
    result = []

    # Request WBEM
    vd_stats = wbem_connection.ExecQuery('DMTF:CQL', request)
    # parse reply and form a table
    for vds in vd_stats:
        # Handle 'InstanceID' in a specific way
        vde = [int(vds.properties[hdr[0]].value.split()[1])]
```

```

        # Collect the rest of numeric performance fields
        for fld in header[1:]:
            vde.append(int(vds.properties[fld].value))
        result.append(vde)

    return result

def count_perf(new, old, interval):
    """Calculate performance delta divided by interval to get per second values"""

    result = []
    for r in range(0, len(new)):
        row = [new[r][0]] # InstanceID
        for c in range(1, len(new[0])):
            row.append(round(float(new[r][c] - old[r][c]) / interval, 2))
        result.append(row)
    return result

def print_perf(stats, hdr):
    """Printout performance data matrix"""

    # Print header
    print(ofs.join(str(fld) for fld in hdr))

    # Print performance table
    for ln in stats:
        print('{}{}{}'.format(ln[0], ofs, ofs.join(str(fld) for fld in ln[1:])))

# Connect with WBEM/CIM services
wbemc = pywbem.WBEMConnection(url, (user, password), 'root/ibm',
no_verification=True)

new_perf = vdisks_perf(wbemc, header)
# Infinite performance processing loop
while True:
    old_perf = new_perf
    new_perf = vdisks_perf(wbemc, header)
    delta_perf = count_perf(new_perf, old_perf, frequency * 60)

    print_perf(delta_perf, header)
    time.sleep(frequency * 60)

```

Note that statistic collection must be already configured and running on the storage system. To check it, run the `lssystem | grep statistics` command.

To set the appropriate statistic generation interval, run the command `startstats -interval N`.

Where *N* is the interval in minutes. For Example C-11 to work properly, the interval must be 5 minutes or less.

SMI-S services can be used not only to report, but to configure storage systems. Algorithms of such operations on the storage system can be accessed using the following links:

<https://ibm.biz/Bdj6XZ>
<https://ibm.biz/Bdj6Xf>

HTTPS and RESTful API on IBM Spectrum Control

A different approach to access storage resources is to use Hypertext Transfer Protocol Secure (HTTPS) protocol and Representational State Transfer (REST or RESTful) API of IBM Spectrum Control server. The main advantage of this method is that it allows us to get information about the entire SAN/Storage infrastructure managed by the IBM Spectrum Control server.

IBM Spectrum Control REST API documentation is available on the following website:

<https://ibm.biz/Bdj6Xy>

The basic idea of it is to send an **HTTP GET** request for a specific URL to retrieve structured information about the storage resource.

For example, requesting `https://spectrumcontrolhost:9569/srm/REST/api/v1/` in the browser shows the root of the entire REST tree: access points for all sections of all possible resources registered in the Spectrum Control server (Figure C-1).

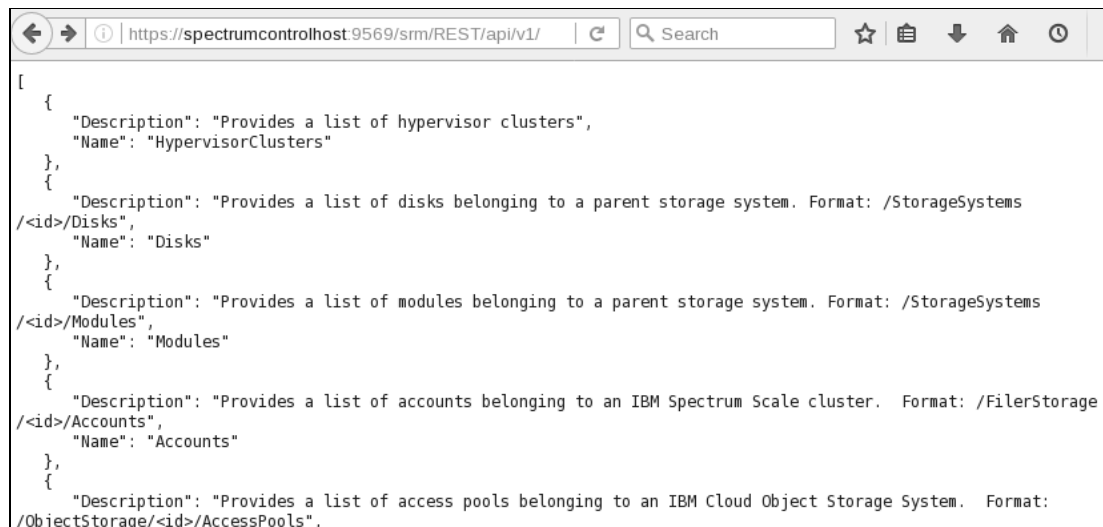


Figure C-1 Accessing Spectrum Control RESTful API using HTTP GET request

Branches and leaves (some of which are nested one in another) of IBM Spectrum Control RESTful API tree are listed in Table 5 on page 417.

Table 5 Branches and leaves of IBM Spectrum Control RESTful API tree

▶ Accessers	▶ NASFileSystems
▶ AccessPools	▶ NASNodes
▶ Accounts	▶ NASNSDs
▶ Applications	▶ NASPools
▶ Blades	▶ NASSnapshots
▶ ClusterNodes	▶ Nodes
▶ Clusters	▶ ObjectStorage
▶ Containers	▶ Performance
▶ Departments	▶ Pools
▶ Disks	▶ RAIDArrays
▶ Fabric	▶ Relationships
▶ FilerStorage	▶ RemoteReplication
▶ GeneralGroups	▶ Servers
▶ HostConnections	▶ Sites
▶ HypervisorClusters	▶ Slicestors
▶ Hypervisors	▶ StoragePools
▶ InterSwitchConnections	▶ StorageSystems
▶ IOGroups	▶ Switches
▶ ManagedDisks	▶ SwitchPorts
▶ Mirrors	▶ Vaults
▶ Modules	▶ Volumes
▶ NASFileSets	▶ ZoneSets

It is possible to access any of these sections manually with a browser, command-line **curl** or **wget** web-retrieving utilities. An example of using **wget** is shown here:

<https://ibm.biz/Bdj6XX>

It is also easy to automate the same operations using Python, because the output is produced in the well-known JavaScript Object Notation (JSON) format. Example C-12 demonstrates the code to fetch information about storage systems. To simplify the work with HTTPS, the open-source module `requests` is used to perform the task.

Example: C-12 Accessing Spectrum Control RESTful API using Python and requests module

```
import requests
import getpass

username = 'scuser'
password = getpass.getpass()

url = 'https://spectrumcontrolhost:9569/srm/'

sess = requests.Session()
sess.verify = False

resp = sess.post(url + 'j_security_check',
                 data={'j_username': username, 'j_password': password})
resp.raise_for_status()

resp = sess.get(url + 'REST/api/v1/' + 'StorageSystems')
resp.raise_for_status()

print(resp.json())
```

To request the password interactively and to hide the users typing, the `getpass` module is engaged.

To log a user in the IBM Spectrum Control authentication procedure requires retrieving a security token from the server and storing it to use during the session. Therefore, to preserve cookies (including the security token) and reuse the same TCP connection with the server (for performance benefits) `requests.Session()` is initialized in the following script:

```
sess = requests.Session()
```

As in the examples for WBEM/CIM (see Example C-8 on page 412 and Example C-11 on page 414) and because of the same reasons, SSL certificate verification here is also disabled: `sess.verify = False`.

Actual requests are sent by `sess.post()` and `sess.get()` within the same session. In the first case, **HTTP POST** is used to submit username and password via the following login form:

```
https://spectrumcontrolhost:9569/srm/j_security_check
```

The second one is an ordinary **HTTP GET** to request RESTful API for storage systems information:

```
https://spectrumcontrolhost:9569/srm/REST/api/v1/StorageSystems
```

In the script, `raise_for_status()` is run on every response. It raises an exception if the server returns 4xx or 5xx error status code but passes transparently if the request was successful.

The last line of the code in Example C-12, decodes JSON data of the response into convenient Python structure of nested lists and dictionaries. In this step any appropriate actions can be applied to parse, filter, and process acquired data, but for the sake of simplicity, the script just prints this structure to standard output.

HTTPS on IBM Spectrum Virtualize

Though it is less documented, it is possible to interact with the Hypertext Transfer Protocol Secure (HTTPS) server of IBM Spectrum Virtualize directly in the very same way typical browsers do. Also, as it was shown in the previous examples for IBM Spectrum Control, Python with open-source module `requests` can greatly simplify the task.

Example C-13 contains basic code for establishing an HTTPS connection with an IBM Spectrum Virtualize system to request and process information about virtual disks.

Example: C-13 Using HTTPS to query IBM Spectrum Virtualize systems

```
import getpass
import requests
import time

target = 'itso_storage'
user = 'itso_user'
password = getpass.getpass()

target_url = 'https://' + target + '{}'
```

```
sess = requests.Session()
sess.verify = False
```

```

print('Accessing the system')
resp = sess.get(target_url.format('/'))
resp.raise_for_status()

time.sleep(1)
print('Trying to login')
resp = sess.post(target_url.format('/login'),
                 data={'login': user, 'password': password})
resp.raise_for_status()

print('Requesting data')
resp = sess.post(target_url.format('/VDiskGridDataHandler'),
                 data={'sort': 'vdiskUid'})
resp.raise_for_status()

vdisks = resp.json()
for vdisk in vdisks['items']:
    print(vdisk['name'], vdisk['capacity'], vdisk['vdiskUid'],
          vdisk['mdiskGrpName'], vdisk['status'])

```

Besides requests, two other modules are imported in this code: `getpass` is needed to hide the user password, and `time` is used to insert a short pause between requests.

A session with disabled SSL certificate verification is established and is described in Example C-12 on page 417.

Before requesting the server for storage related information, such as VDisk data in this example, two steps must be completed in advance:

1. A valid session cookie must be acquired from the server. In Example C-13 it is performed with a simple **GET** request to the root of the server: `sess.get(target_url.format('/'))`.
Note, a small pause `time.sleep(1)` after this step might be required for the server to set up an environment for the new session.
2. A user must be authenticated by the storage system. This is done by filling in the login-form with a **POST** request: `sess.post(target_url.format('/login'), data={'login': user, 'password': password})`

Only after successfully accomplishing both of these stages can a request to query actual storage related information be done.

As it was shown earlier in Example C-12 on page 417, `resp.raise_for_status()` is called to catch HTTP errors if they occur.

The code accesses the `/VDiskGridDataHandler` resource, but actually any valid page available via WEBUI can be requested from the server, for example: `/FlashCopyTreeDataHandler`, `/HostGridDataHandler`. Another interesting URL is `/RPCAdapter` which can be used to obtain performance statistics from the storage system.

The **HTTP POST** method to query such resources is needed to submit additional parameters in JSON format or various form options, such as the sorting order `{'sort': 'vdiskUid'}` shown in Example C-13 on page 418.

Because the server also responds in JSON it is useful to process it with the built-in requests module JSON parser: `vdisks = resp.json()`.

To keep the example simple, the last lines of the script define just a for-loop to go over the acquired vdisk information and print some of the valuable items.

Conclusions

This chapter briefly showed several unified methods of accessing IBM SVC/Storwize systems with scripts. Modern and open protocols, APIs and standards allow the ability to write versatile code for storage management, reporting, and configuration tasks using programming languages convenient for different environments and operating systems.

The SNMP protocol on IBM Spectrum Virtualize is primarily used for system monitoring tasks. It is not touched by the scripting topic because SNMP-agents are configured to work in send-traps only mode, and it is not possible to access them with SNMP GetRequest/SetRequest/etc requests.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document (note that some publications referenced in this list might be available in softcopy only):

- ▶ *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V8.1*, SG24-7933
- ▶ *Implementing the IBM Storwize V7000 with IBM Spectrum Virtualize V8.1*, SG24-7938
- ▶ *Implementing the IBM Storwize V5000 Gen2 (including the Storwize V5010, V5020, and V5030) with IBM Spectrum Virtualize V8.1*, SG24-8162
- ▶ *IBM b-type Gen 5 16 Gbps Switches and Network Advisor*, SG24-8186
- ▶ *IBM Spectrum Virtualize: Hot Spare Node and NPIV target ports*, REDP-5477

You can search for, view, download, or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

The following Redbooks domains related to this book are also useful resources:

- ▶ IBM Storage Networking Redbooks
<http://www.redbooks.ibm.com/Redbooks.nsf/domains/san>
- ▶ IBM Flash Storage Redbooks
<http://www.redbooks.ibm.com/Redbooks.nsf/domains/flash>
- ▶ IBM Software Defined Storage Redbooks
<http://www.redbooks.ibm.com/Redbooks.nsf/domains/sds>
- ▶ IBM Disk Storage Redbooks
<http://www.redbooks.ibm.com/Redbooks.nsf/domains/disk>
- ▶ IBM Storage Solutions Redbooks
<http://www.redbooks.ibm.com/Redbooks.nsf/domains/storagesolutions>
- ▶ IBM Tape storage Redbooks
<http://www.redbooks.ibm.com/Redbooks.nsf/domains/tape>

Other publications

These publications are also relevant as further information sources:

- ▶ *IBM System Storage Master Console: Installation and User's Guide*, GC30-4090
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: CIM Agent Developers Reference*, SC26-7545
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: Command-Line Interface User's Guide*, SC26-7544
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: Configuration Guide*, SC26-7543
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: Host Attachment Guide*, SC26-7563
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: Installation Guide*, SC26-7541
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: Planning Guide*, GA22-1052
- ▶ *IBM System Storage Open Software Family SAN Volume Controller: Service Guide*, SC26-7542
- ▶ *IBM System Storage SAN Volume Controller - Software Installation and Configuration Guide*, SC23-6628
- ▶ *IBM System Storage SAN Volume Controller V6.2.0 - Software Installation and Configuration Guide*, GC27-2286
- ▶ *IBM System Storage SAN Volume Controller 6.2.0 Configuration Limits and Restrictions*, S1003799
- ▶ *IBM TotalStorage Multipath Subsystem Device Driver User's Guide*, SC30-4096
- ▶ *IBM XIV and SVC Best Practices Implementation Guide*
<http://ibm.co/1bk64gW>
- ▶ *Considerations and Comparisons between IBM SDD for Linux and DM-MPIO*
<http://ibm.co/1CD1gxG>

Online resources

These websites are also relevant as further information sources:

- ▶ IBM Storage home page
<http://www.ibm.com/systems/storage>
- ▶ SAN Volume Controller supported platform
<http://ibm.co/1FNjddm>
- ▶ SAN Volume Controller IBM Knowledge Center
<http://www.ibm.com/support/knowledgecenter/STPVGU/welcome>
- ▶ Cygwin Linux-like environment for Windows
<http://www.cygwin.com>

- ▶ Open source site for SSH for Windows and Mac
<https://www.ssh.com/ssh/download/>
- ▶ Windows Sysinternals home page
<http://www.sysinternals.com>
- ▶ Download site for Windows PuTTY SSH and Telnet client
<http://www.chiark.greenend.org.uk/~sgtatham/putty>

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Redbooks

IBM System Storage SVC and Storwize V7000 Best Practices and

SG24-7521-05
ISBN 073844300X



(0.5" spine)
0.475" x 0.873"
250 <-> 459 pages



SG24-7521-05

ISBN 073844300X

Printed in U.S.A.

Get connected

