

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

Jon Tate

Angelo Bernasconi

Tiago Moreira Candelaria

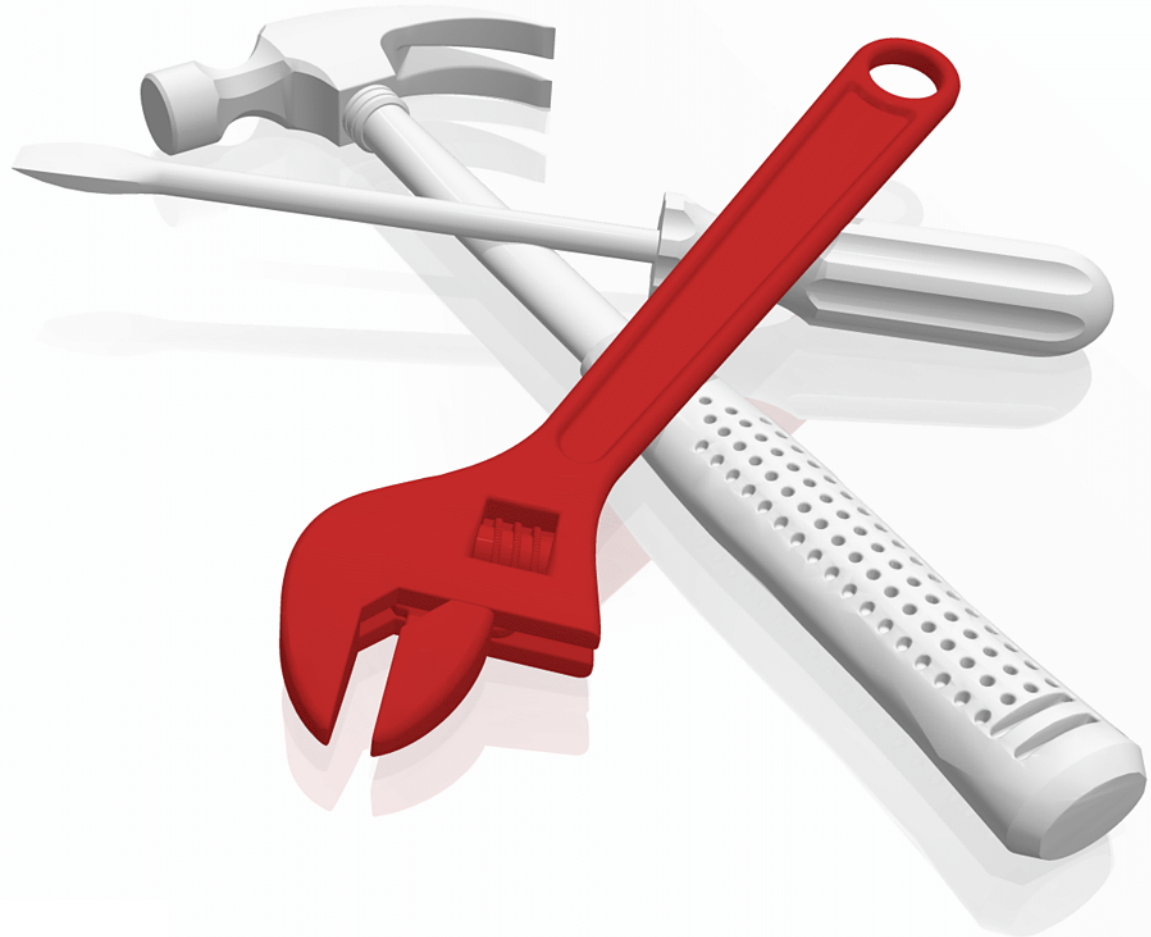
Bastos

Giulio Fiscella

Bruno Garcia Galle

Jana Jamsek

Antonio Rainero





International Technical Support Organization

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

December 2016

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

Fifth Edition (December 2016)

This edition applies to IBM Spectrum Virtualize V7.8, and the associated hardware and software detailed within. Note that screenshots may differ from the generally available (GA) version as parts of this book were written with pre-GA code.

This document was created or updated on February 15, 2017.

© Copyright International Business Machines Corporation 2016. 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!	xvi
Comments welcome	xvi
Stay connected to IBM Redbooks	xvi
Summary of changes	xix
December 2016, Fifth Edition	xix
Chapter 1. SAN topology	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	4
1.2.2 Basic core-edge topology	6
1.2.3 Edge-core-edge topology	7
1.2.4 Full mesh topology	8
1.2.5 IBM Spectrum Virtualize/Storwize as a SAN bridge	8
1.2.6 Device placement	9
1.3 SAN Volume controller ports	12
1.3.1 Slots and ports identification	12
1.3.2 Port naming and distribution	13
1.4 Zoning	15
1.4.1 Types of zoning	16
1.4.2 Prezoning tips and shortcuts	16
1.4.3 SAN Volume Controller internode communications zone	17
1.4.4 SAN Volume Controller/Storwize storage zones	18
1.4.5 SAN Volume Controller/Storwize host zones	27
1.4.6 Zoning with multiple SAN Volume Controller/Storwize clustered systems	32
1.4.7 Split storage subsystem configurations	33
1.5 Distance extension for remote copy services	33
1.5.1 Optical multiplexors	33
1.5.2 Long-distance SFPs or XFPs	33
1.5.3 Fibre Channel over IP	33
1.5.4 Native IP replication	35
1.6 Tape and disk traffic that share the SAN	35
1.7 Switch interoperability	35
Chapter 2. Back-end storage	37
2.1 Round Robin Path Selection	38
2.2 Considerations for DS88xx series	39
2.2.1 Balancing workload across DS88xx series controllers	39
2.2.2 DS88xx series ranks to extent pools mapping	42
2.2.3 Determining the number of controller ports for DS88xx series	43
2.2.4 LUN masking	43

2.3 Considerations for IBM XIV Storage System	45
2.3.1 Connectivity considerations	45
2.3.2 Host options and settings for XIV systems	45
2.3.3 Managed disks considerations	46
2.3.4 Additional considerations	47
2.4 Considerations for IBM Storwize V7000/V5000/V3700	47
2.4.1 Cabling and zoning	48
2.4.2 Defining internal storage	48
2.5 Considerations for IBM FlashSystem 900	51
2.5.1 Physical FC port connection and zoning	52
2.5.2 Storage capacity and Arrays	52
2.5.3 Extent size	52
2.5.4 Storage pools	52
2.5.5 Volumes	54
2.6 Considerations for third-party storage with EMC VMAX and Hitachi Data Systems . . .	56
Chapter 3. Storage pools and managed disks	57
3.1 Availability considerations for storage pools	58
3.2 Selecting storage subsystems	59
3.3 Selecting the storage pool.	59
3.3.1 Selecting the number of arrays per storage pool	60
3.3.2 Selecting LUN attributes	64
3.3.3 Considerations for Storwize family systems	65
3.3.4 Considerations for the IBM XIV Storage System	66
3.3.5 IBM FlashSystem A9000/A9000R considerations	68
3.3.6 Considerations for the DS88XX family	72
3.4 Quorum disk considerations	73
3.5 Volume Consideration	77
3.6 Tiered storage	78
3.7 Adding MDisk to existing storage pools	80
3.7.1 Checking access to new MDisk	80
3.7.2 Persistent reserve	81
3.7.3 Renaming MDisk	81
3.8 Rebalancing extents across a storage pool	81
3.9 Removing MDisk from existing storage pools	82
3.9.1 Migrating extents from the MDisk to be deleted	82
3.9.2 Verifying the identity of an MDisk before removal	82
3.9.3 Correlating the back-end volume with the MDisk	83
3.10 Remapping managed MDisk	88
3.11 Controlling extent allocation order for volume creation	89
3.12 Considerations when using Encryption	89
3.12.1 Encryption at rest with USB Key	89
3.12.2 Encryption at rest with Key Server	91
3.12.3 Encryption on external storage	93
Chapter 4. Volumes	95
4.1 Overview of volumes	96
4.1.1 Striping compared to sequential type	96
4.1.2 Thin-provisioned volumes	96
4.1.3 Space allocation	97
4.1.4 Compressed volumes	98
4.1.5 Thin-provisioned volume	101
4.1.6 Limits on virtual capacity of Thin-provisioned volumes	105

4.2	Creating volumes	106
4.2.1	Changing the preferred node within an I/O group or cross I/O group	107
4.3	Volume migration	108
4.3.1	Image-type to striped-type migration	108
4.3.2	Migrating to image-type volume	109
4.3.3	Migrating with volume mirroring	110
4.4	VMware Virtual Volumes	111
4.5	IPreferred paths to a volume	112
4.5.1	Governing of volumes	113
4.6	Cache mode and cache-disabled volumes	118
4.6.1	Underlying controller remote copy with Spectrum Virtualize and Storwize cache-disabled volumes	119
4.6.2	Using underlying controller FlashCopy with Spectrum Virtualize and Storwize cache disabled volumes	120
4.6.3	Changing the cache mode of a volume	120
4.7	Using Spectrum Virtualize or Storwize with FlashSystem	121
4.8	FlashCopy services	123
4.8.1	FlashCopy rules summary	123
4.8.2	IBM Spectrum Protect Snapshot	124
4.8.3	IBM System Storage Support for Microsoft Volume Shadow Copy Service	124
4.9	Configuration Backup	125
Chapter 5.	Copy Services	131
5.1	Introduction to copy services	132
5.2	FlashCopy	133
5.2.1	FlashCopy Use Cases	133
5.2.2	Flash Copy capabilities overview	134
5.2.3	Flash Copy functional overview	140
5.2.4	FlashCopy planning considerations	145
5.3	Remote Copy services	152
5.3.1	Remote copy functional overview	153
5.3.2	Remote copy network planning	166
5.3.3	Remote copy services planning	175
5.3.4	Remote copy use cases	183
5.3.5	1920 error	188
5.4	Native IP Replication	199
5.4.1	Native IP replication technology	199
5.4.2	IP partnership limitations	200
5.4.3	VLAN support	202
5.4.4	IP Compression	203
5.4.5	Remote copy groups	204
5.4.6	Supported configurations	206
5.4.7	Native IP replication performance consideration	213
5.5	Volume Mirroring	215
5.5.1	Volume mirroring use cases	216
5.5.2	Mirrored volume components	218
5.5.3	Performance considerations of Volume Mirroring	219
5.5.4	Bitmap space for out-of-sync volume copies	221
5.5.5	Synchronization status of volume copies	223
Chapter 6.	Hosts	225
6.1	Configuration guidelines	226
6.1.1	Host levels and host object name	226

6.1.2	Host Cluster	226
6.1.3	The number of paths	227
6.1.4	Host ports	227
6.1.5	Port masking	228
6.1.6	Host to I/O group mapping	229
6.1.7	Volume size as opposed to quantity	230
6.1.8	Host volume mapping	230
6.1.9	Server adapter layout	234
6.2	N-Port Virtualization ID (NPIV)	234
6.3	Host pathing	237
6.3.1	Multipathing Software	237
6.3.2	Preferred path algorithm	238
6.3.3	Path selection	238
6.3.4	Path management	239
6.3.5	Non disruptive Volume migration between I/O groups	240
6.4	I/O queues	242
6.4.1	Queue depths	242
6.5	Host clustering and reserves	243
6.5.1	Clearing reserves	243
6.5.2	Spectrum Virtualize MDisk reserves	244
6.6	AIX hosts	244
6.6.1	HBA parameters for performance tuning	244
6.6.2	Configuring for fast fail and dynamic tracking	246
6.6.3	SDDPCM	246
6.7	Virtual I/O Server	247
6.7.1	Methods to identify a disk for use as a virtual SCSI disk	248
6.7.2	UDID method for MPIO	248
6.8	Windows hosts	249
6.8.1	Clustering and reserves	249
6.8.2	Tunable parameters	249
6.8.3	Guidelines for disk alignment by using Windows with Spectrum Virtualize volumes	249
6.9	Linux hosts	250
6.9.1	Tunable parameters	250
6.10	Solaris hosts	251
6.10.1	Solaris MPxIO	251
6.10.2	Symantec Veritas Volume Manager	252
6.10.3	DMP multipathing	252
6.10.4	Troubleshooting configuration issues	253
6.11	VMware server	253
6.11.1	Multipathing solutions supported	254
6.11.2	Multipathing configuration maximums	254
6.12	Monitoring	255
6.12.1	Load measurement and stress tools	256
Chapter 7. IBM Easy Tier function		257
7.1	Easy Tier	258
7.1.1	Easy Tier concepts	258
7.1.2	Four tiers Easy Tier and Read Intensive flash drive	260
7.1.3	SSD arrays and Flash MDisks	261
7.1.4	Disk tiers	263
7.1.5	Easy Tier process	266
7.1.6	Easy Tier operating modes	268

7.2 Easy Tier implementation considerations	269
7.2.1 Implementation rules	269
7.2.2 Limitations	270
7.2.3 Easy Tier settings	270
7.3 Monitoring tools	275
7.3.1 Offloading statistics	275
7.3.2 Interpreting the STAT tool output	276
7.3.3 IBM STAT Charting Utility	283
Chapter 8. Monitoring	287
8.1 Generic monitoring	288
8.1.1 Monitoring with GUI	288
8.1.2 Monitoring using quotas and alert	289
8.2 Performance Monitoring	289
8.2.1 Performance monitoring with the GUI	290
8.2.2 Performance monitoring with IBM Spectrum Control	291
8.2.3 Important metrics for debugging	294
8.2.4 Performance support package	295
8.3 Metro and Global Mirror monitoring with IBM Copy Services Manager and scripts . .	297
8.3.1 Monitoring MM and GM with scripts	298
8.4 Monitoring Tier1 SSD	299
Chapter 9. Maintenance	301
9.1 Documenting Spectrum Virtualize and SAN environment	302
9.1.1 Naming conventions	302
9.1.2 SAN fabrics documentation	305
9.1.3 Spectrum Virtualize documentation	307
9.1.4 Storage documentation	309
9.1.5 Technical Support information	309
9.1.6 Tracking incident and change tickets	310
9.1.7 Automated support data collection	311
9.1.8 Subscribing to Spectrum Virtualize support	311
9.2 Storage management users	311
9.3 Standard operating procedures	312
9.3.1 Allocating and deallocating volumes to hosts	312
9.3.2 Adding and removing hosts	313
9.4 Spectrum Virtualize code update	313
9.4.1 Current and target Spectrum Virtualize code level	314
9.4.2 Spectrum Virtualize Upgrade Test Utility	314
9.4.3 Spectrum Virtualize hardware considerations	316
9.4.4 Attached hosts preparation	317
9.4.5 Storage controllers preparation	317
9.4.6 SAN fabrics preparation	317
9.4.7 SAN components update sequence	318
9.4.8 Spectrum Virtualize participating in Metro Mirror or Global Mirror	319
9.4.9 Spectrum Virtualize update	319
9.4.10 Spectrum Virtualize disk drive update	319
9.5 SAN modifications	320
9.5.1 Cross-referencing HBA WWPNs	320
9.5.2 Cross-referencing LUN IDs	322
9.5.3 HBA replacement	322
9.6 Hardware upgrades for Spectrum Virtualize	323
9.6.1 Adding Spectrum Virtualize nodes to an existing cluster	323

9.6.2	Upgrading Spectrum Virtualize nodes in an existing cluster	325
9.6.3	Moving to a new Spectrum Virtualize cluster	325
9.6.4	Splitting a Spectrum Virtualize cluster.	327
9.7	Adding expansion enclosures	327
9.8	Spectrum Virtualize scripting.	328
9.8.1	Connecting to Spectrum Virtualize using PuTTY predefined connection	329
9.8.2	Run commands in the Spectrum Virtualize shell.	331
9.8.3	Scripting toolkit	331
Chapter 10.	Troubleshooting and diagnostics	333
10.1	Common issues	334
10.1.1	Host problems.	334
10.1.2	Spectrum Virtualize events	334
10.1.3	SAN events	337
10.1.4	Storage subsystem issues	337
10.1.5	Port masking issues	339
10.2	Collecting data and isolating the problem	339
10.2.1	Data collection	339
10.3	Recovering from problems	340
10.3.1	Solving host problems.	340
10.3.2	Solving Spectrum Virtualize events.	342
10.3.3	Solving SAN issues.	344
10.3.4	Solving back-end storage issues	345
10.3.5	Replacing a failed disk	346
10.4	Health status during upgrade and known errors	347
Chapter 11.	IBM Real-time Compression	349
11.1	Evaluate compression savings using Comprestimator	350
11.2	Evaluate workload using Disk Magic.	351
11.3	Verify available CPU resources	352
11.4	Configure a balanced system	353
11.5	Standard benchmark tools	353
11.6	Compression with FlashCopy	353
11.7	Compression with Easy Tier	354
11.8	Compression on the backend	354
11.9	Migrating generic volumes	355
11.10	Mixed volumes in the same MDisk group	355
Appendix A.	IBM i considerations	357
	IBM i Storage management	358
	Single level storage	358
A.1	IBM i response time	358
	Planning for IBM i capacity.	359
	Connecting SAN Volume Controller or Storwize to IBM i.	359
Native connection		359
Connection with VIOS_NPIV		360
Connection with VIOS virtual SCSI.		360
	Setting of attributes in VIOS.	361
FC adapter attributes		361
Disk device attributes		361
	Disk drives for IBM i	362
	Defining LUNs for IBM i	362
	Data layout	363
	Fibre Channel adapters in IBM i and VIOS	364

Zoning SAN switches	365
IBM i Multipath	365
Boot from SAN	366
IBM i mirroring	366
Copy services considerations	366
Appendix B. Business continuity	367
11.11 Business Continuity with Stretched Cluster	368
11.12 Business Continuity with Enhanced Stretched Cluster	368
11.13 Business Continuity with HyperSwap	368
11.14 IP Quorum Configuration	369
Related publications	371
IBM Redbooks	371
Other resources	372
Referenced websites	372
Help from IBM	373

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

The following terms are trademarks of other companies:

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 that are based on field experience and describes the performance gains that can be achieved by implementing the IBM System Storage® SAN Volume Controller and Storwize® V7000 powered by IBM Spectrum™ Virtualize V7.8.

It 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, this book 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. Next, it 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. Before reading this book, you must have 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 31 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).



Angelo Bernasconi is an Executive Certified Storage and SDS IT Specialist for IBM Systems Italy. He has 30 years of experience in the delivery of professional services and solutions for IBM Enterprise customers in open systems. He holds a degree in Electronics and his areas of expertise are the IBM Storage Portfolio and SDS, and design and implementation of storage solutions. He has written extensively about SAN and storage products in a variety of IBM publications. He is also a member of the Italy SNIA committee and a member of the Italy TEC.



Tiago Moreira Candelaria Bastos is a SAN and Storage Disk specialist for IBM Brazil. He has over 15 years experience in the IT arena, and is an IBM Certified Master IT Specialist. Certified for Storwize, he works on Storage as a Service implementation projects and his areas of expertise include planning, configuring and troubleshooting IBM DS8000®, Storwize V5000 and V7000.



Giulio Fiscella is a Software Engineer for IBM Systems at the IBM Manchester Lab. He works on the whole IBM Spectrum Virtualize™ portfolio developing and testing new functionality. He has worked on several projects since the V7.2.0 release and he was fully responsible for different features released in V7.7.1 and V7.8.0.



Bruno Garcia Galle joined IBM in 2007 as a SAN and Storage Support specialist for IBM Global Services in Brazil. Since 2009, he has been working as a SAN Storage Subject Matter Expert (SME) for many international customers supporting different customers and environments. Bruno's areas of expertise include Enterprise and Midrange Storage as well as storage virtualization and Storage Area Network (SAN) from different brands. Currently, he is a senior IT Specialist in project planning and implementation, and is engaged on SAN and storage related projects.



Jana Jamsek is an IT specialist for IBM Slovenia. She works in Storage Advanced Technical Skills for Europe as a specialist for IBM Storage Systems and IBM i systems. Jana has 8 years of experience in the 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 the University of Ljubljana, Slovenia. Jana works on complex customer cases that involve IBM i and Storage systems in different European and Middle East countries. She presents on 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.



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 15 years of experience in the delivery of storage services for Open Systems and 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.

Thanks to the authors of the previous edition:

Pawel Brodacki, Tilak Buneti, Christian Burns, Jana Jamsek, Erez Kirson, Marcin Tabinowski, Bosmat Tuv-El

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

Christopher Bulmer, Paul Cashman, Carlos Fuente, Katja Gebuhr, Warren Hawkins, Gareth Jones, Evelyn Perez, Mark Visser, Stephen Wright

IBM Hursley, UK

Nick Clayton

IBM Systems, UK

Navin Manohar

Terry Niemeyer

IBM Systems, US

Chris Saul

IBM Systems, US

Barry Whyte

IBM Systems, New Zealand

Bill Wiegand

IBM Systems, US

Da Lu

IBM Systems, China

Chuck Laing

IBM GTS, US

Ian MacQuarrie

IBM Systems, US

Torsten Rothenwald

IBM Systems, Germany

Rahul Fiske

IBM Systems, India

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

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 SG24-7521-04
for IBM System Storage SAN Volume Controller and Storwize V7000 Best Practices and
Performance Guidelines
as created or updated on February 15, 2017.

December 2016, Fifth Edition

This revision includes the new and changed information and is based on the latest hardware available from IBM and as described in the book. We have also based this on IBM Spectrum Virtualize V7.8, and have removed chapters and topics where that information can easily be found by your favorite search engine or in other IBM Redbooks.



SAN topology

The 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 the classic storage practices.

In the following chapter we do not describe how to design and build a flawless SAN from the beginning, but provide guidance to connect Spectrum Virtualize/Storwize in an existing SAN to achieve a stable, redundant, resilient, scalable and performance likely environment. However, the principles here can be used and taken into account when you are building your SAN.

This chapter includes the following sections:

- ▶ 1.1, “SAN topology general guidelines” on page 2
- ▶ 1.2, “SAN topology specific guidelines” on page 4
- ▶ 1.3, “SAN Volume controller ports” on page 12
- ▶ 1.4, “Zoning” on page 15
- ▶ 1.5, “Distance extension for remote copy services” on page 33
- ▶ 1.6, “Tape and disk traffic that share the SAN” on page 35
- ▶ 1.7, “Switch interoperability” on page 35

1.1 SAN topology general guidelines

The SAN topology requirements for the Spectrum Virtualize/Storwize do not differ too much from any other SAN. Keep in mind 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, it is necessary guarantee that your environment follows an actual SAN architecture and designs best practices for the storage industry.

For additional SAN design and best practices refer to the white paper on Brocade SAN Design and Best Practices, available at:

<https://ibm.biz/BdsFjG>

A topology is described in terms of how the switches are interconnected. There are some 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. Independently of the environment demands there are a few best practices that must be followed to keep your SAN healthy, performing good, redundant, and resilient.

1.1.1 SAN Performance and scalability

Regardless of the storage and the environment, planning and sizing of the SAN will make a difference when growing your environment, or even so during problems troubleshooting.

Once most SAN installations will continue to grow over the years, the main SAN industry-lead companies design their products in a way to support a certain growth, so keep in mind that your SAN must be designed considering a short/medium time growth.

From the performance stand point the following topics must be evaluated and considered:

- ▶ Host-to-storage fan-in fan-out ratios.
- ▶ Host to 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 stand point consider 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 refers to 1.1.2, "ISL Considerations".

Consider the use of SAN director-class switches, they reduce the number of switches in a SAN and provide the best scalability available. The SAN b-type directors can be easily cascaded using Inter-chassis Link (ICL) ports, that are dedicated exclusively to inter-connect SAN Directors. For more about SAN b-type directors and Inter-chassis Link (ICL) and SAN director scalability refers to *IBM b-type Gen 5 16 Gbps Switches and Network Advisor*, SG24-8186.

So, if possible, plan for the maximum size configuration that you expect your SAN Volume Controller/Storwize installation to reach. Planning for the maximum size does not mean that you must purchase all of the SAN hardware initially. It requires you to only design the SAN taking into consideration the expected maximum size.

1.1.2 ISL Considerations

Inter switch links (ISL) are responsible for interconnecting the SAN switches allowing SAN flexibility and scalability. For this reason they can be considered as the core of a SAN topology and, consequently, they are sometimes the main cause of issues that can affect a SAN. So 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 inter-switch link (ISL) oversubscription ratio.
 - The standard recommendation is up to 7:1 (7 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. On the other hand, high performance SAN's starts at 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 Inter Switch Link 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 inter-switch links (ISL) 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.

Once the Fibre Channel switches have reached 4 Gbps and beyond, the number of issues related to traffic saturation has dramatically decreased. 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 ISL on you SAN. Using an extra ISL avoids congestion if an ISL fails because of issues such as a SAN switch line card or port blade failure.
- ▶ Use the trunking license to obtain better performance.
- ▶ 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 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 on page 4.

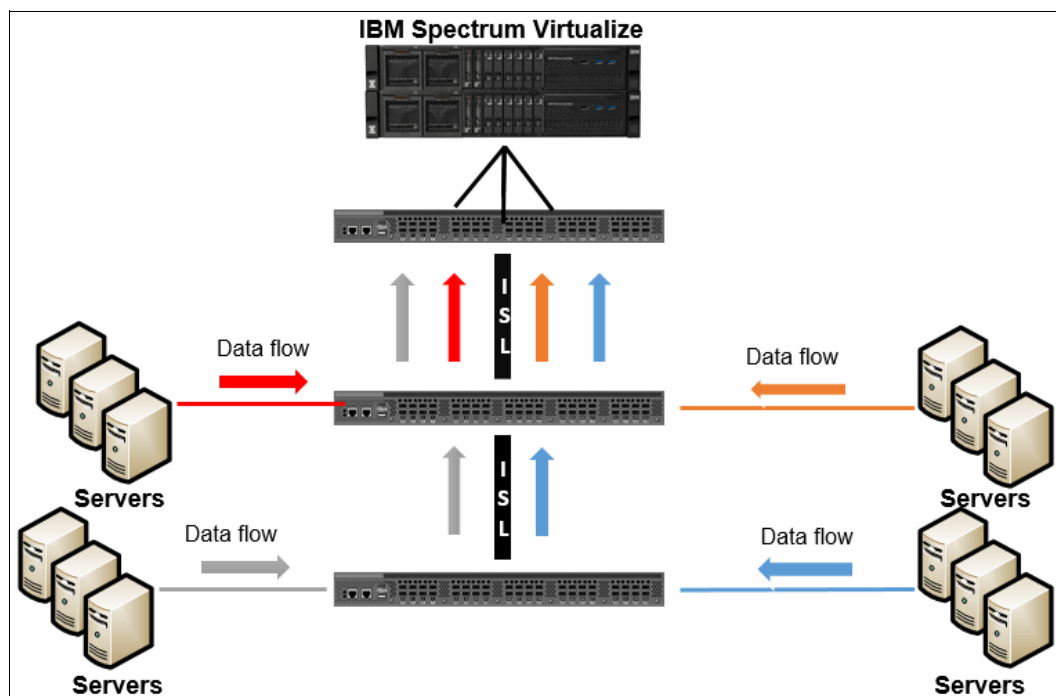


Figure 1-1 ISL Data flow

- ▶ 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 via the links connected to other blades.
- ▶ Plan for peak load, not for average.

1.2 SAN topology specific guidelines

Some best practices, as mentioned in 1.1, “SAN topology general guidelines” on page 2 apply to all SANs. However there are specific best 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.

In this chapter we will cover 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

having 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, once all servers and storages are connected to the same switches, there is no hop.

Note: To meet redundancy and resiliency requisites, a single switch solution needs at least two SAN switches/Directors, one per different fabric.

The preferred practice is to use a multislot director-class single switch over setting up a core-edge fabric that is made up solely of lower-end switches, as is 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.

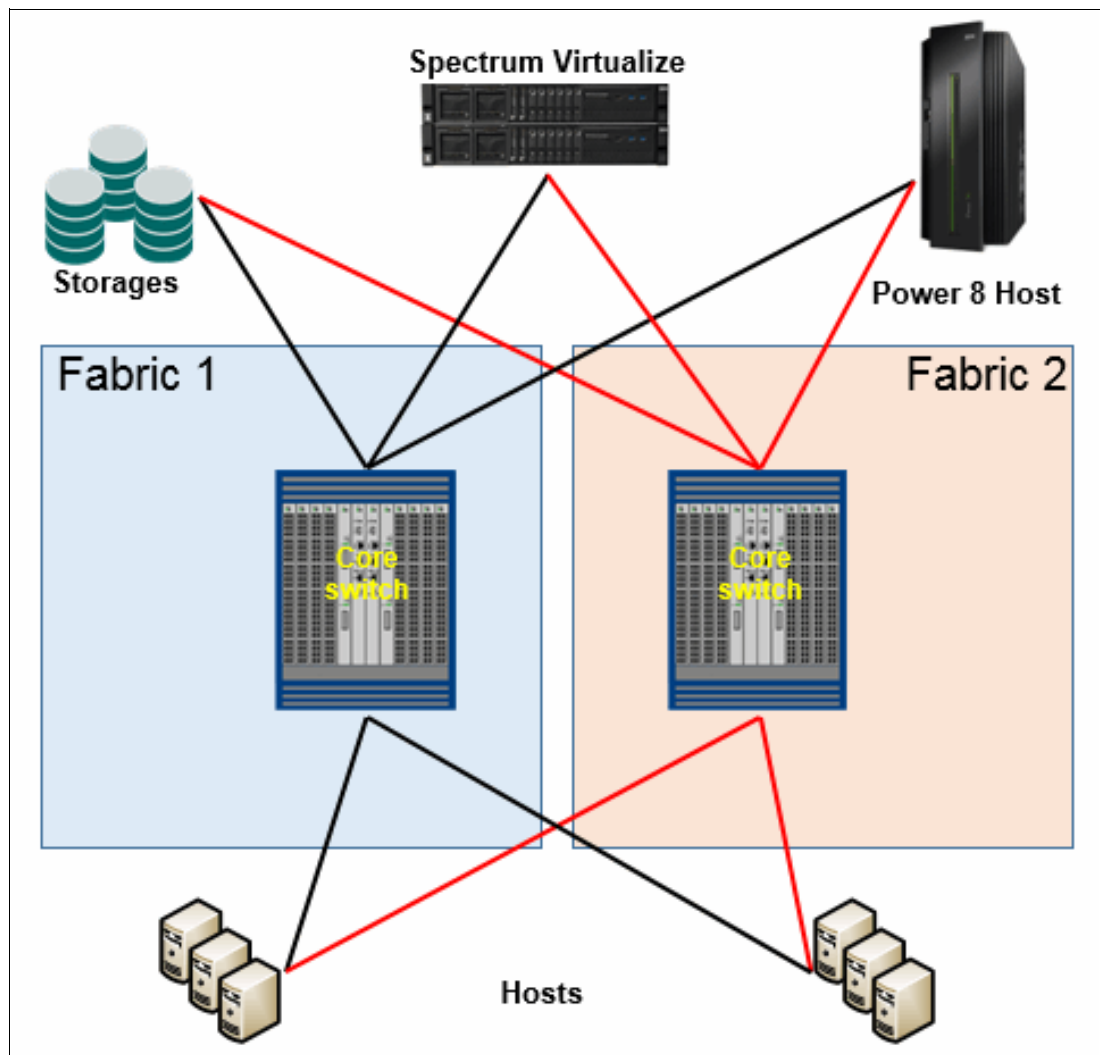


Figure 1-2 Single switch SAN

1.2.2 Basic core-edge topology

The core-edge topology (as shown in Figure 1-3 on page 6) 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/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.

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

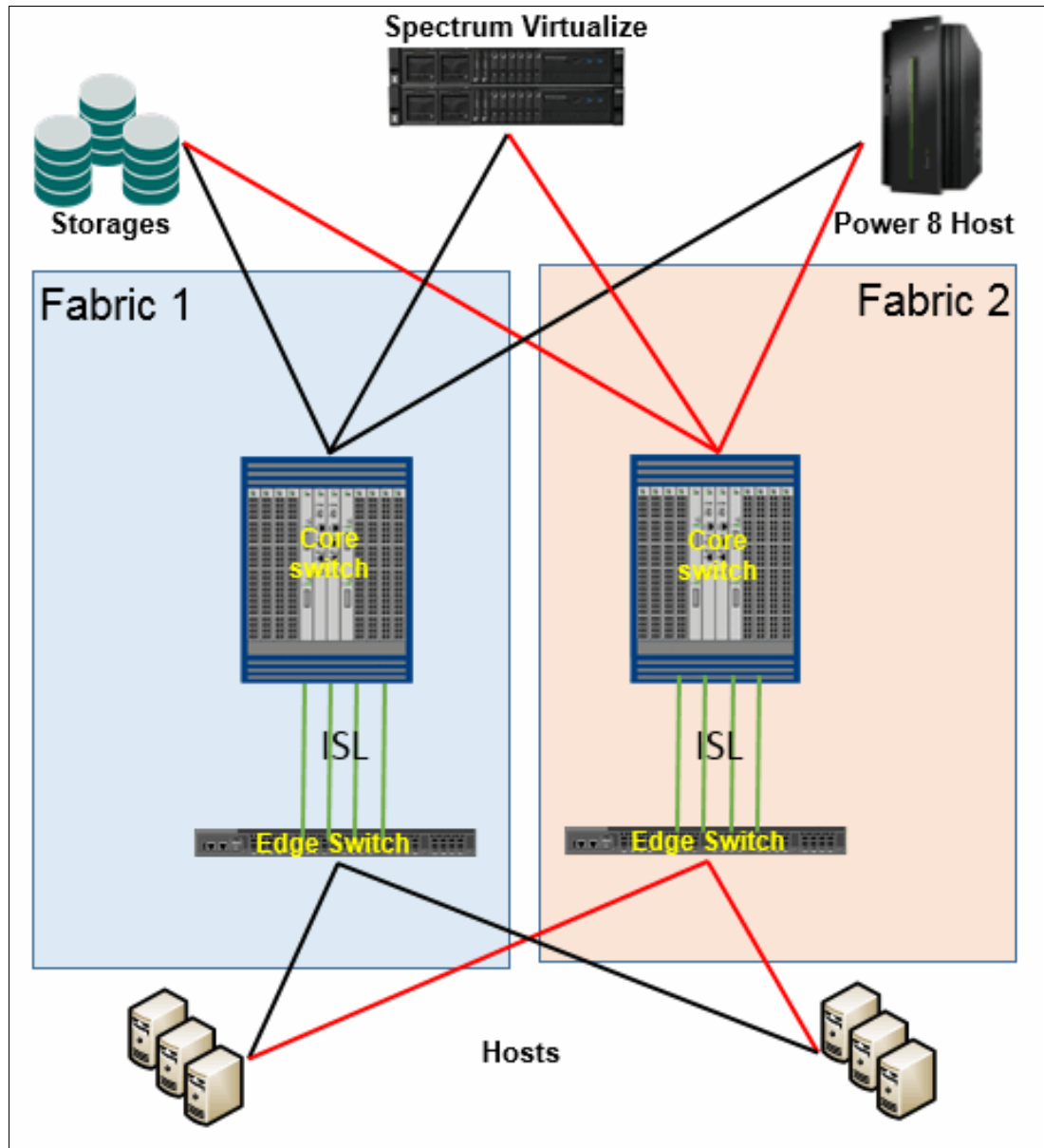


Figure 1-3 Core-edge topology

1.2.3 Edge-core-edge topology

Edge-core-edge is a topology 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 varying.

Figure 1-4 on page 7 shows an edge-core-edge topology showing two different edges, one exclusive for the storage, SAN Volume Controller and high bandwidth servers and the other pair exclusively for servers.

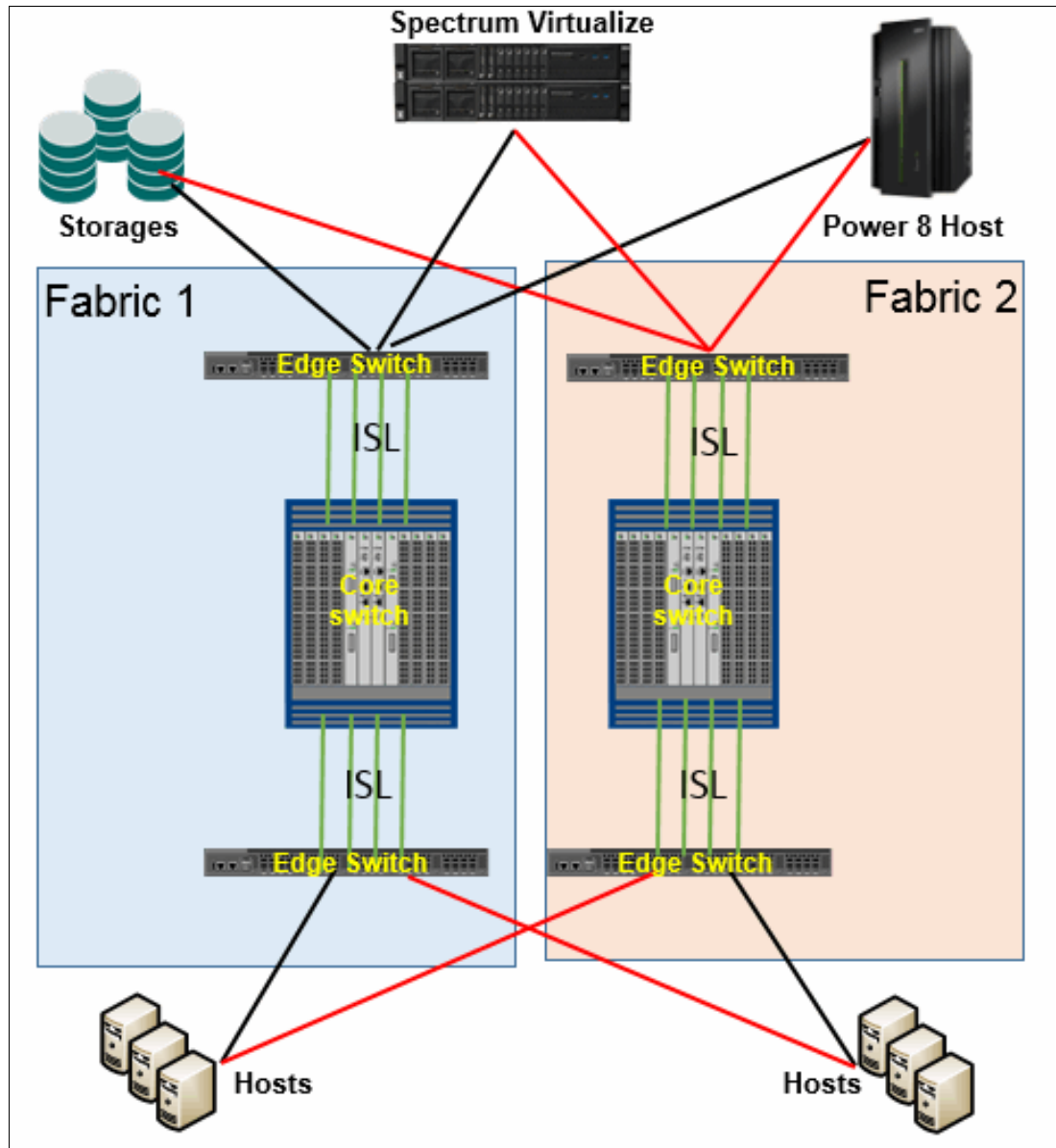


Figure 1-4 Edge-core-edge topology

Edge-core-edge fabrics allow a 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 once the number of hops will be no more than one hop. Figure 1-5 on page 8 shows a full mesh topology.

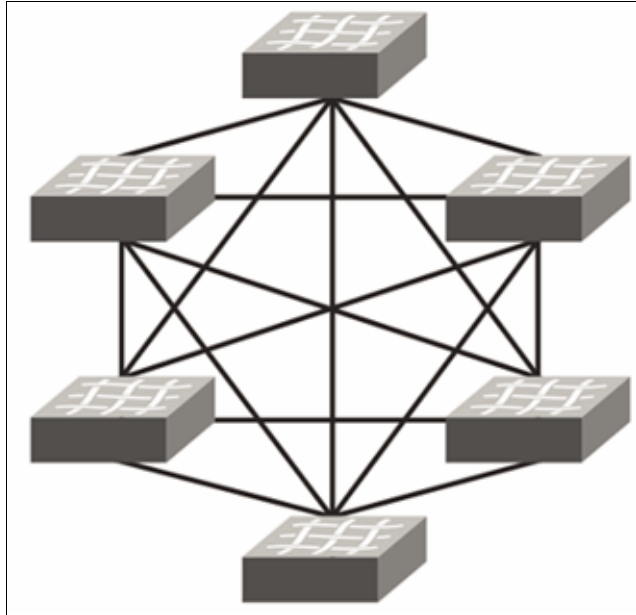


Figure 1-5 Full mesh topology

1.2.5 IBM Spectrum Virtualize/Storwize as a SAN bridge

IBM SAN Volume Controller nodes now have a maximum of sixteen ports. Besides the increased throughput capacity, this number of ports enables new possibilities and allows different kinds of topologies and/or migration scenarios.

One of these topologies is the use of a SAN Volume Controller/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 on page 9 has an example of a Spectrum Virtualize/Storwize as a SAN bridge:

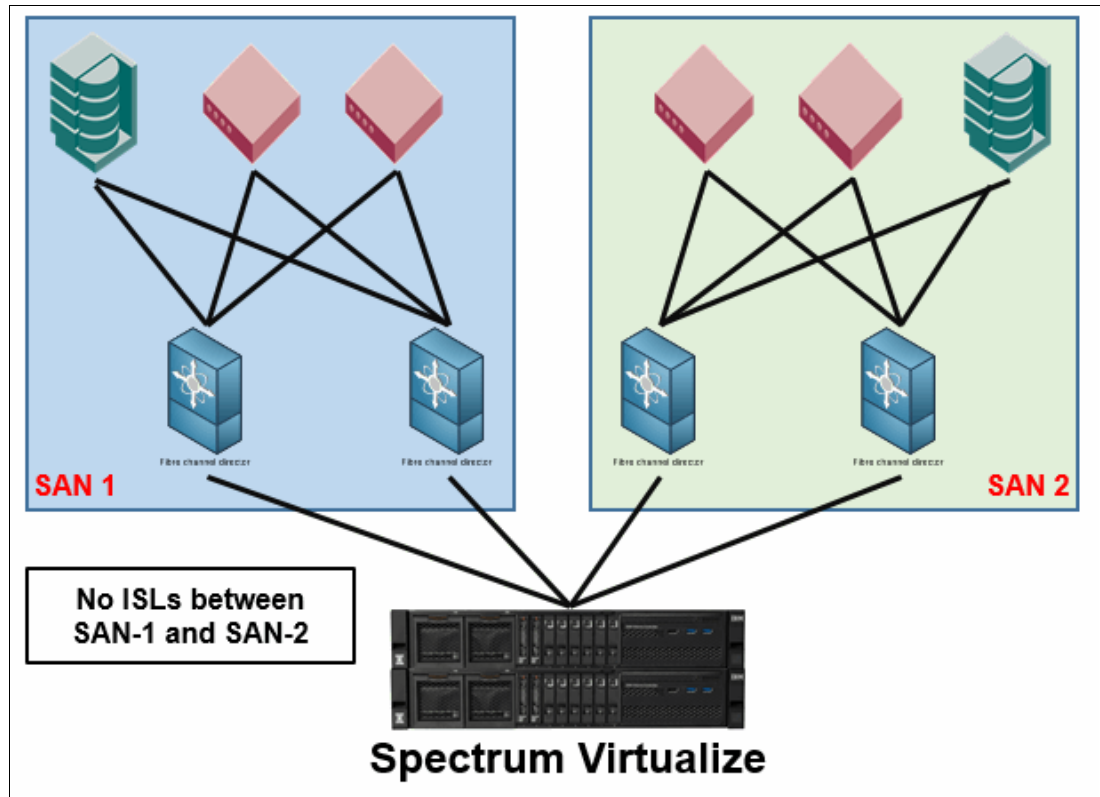


Figure 1-6 Spectrum Virtualize/Storage as SAN bridge

Notice in Figure 1-6 that both SANs (Blue and Green) are isolated and there is no communication via ISLs. When connected to both fabrics SAN Volume Controller/Storage is able to virtualize storages from either fabrics and provide disks from storage located on Green SAN, for example, to hosts on blue SAN.

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 and so ensure that the traffic is flowing in a certain way to avoid congestion and performance issues. The ways to balance the traffic consists of isolating traffic by using zoning, Virtual SANs (VSAN) and/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 on page 10.

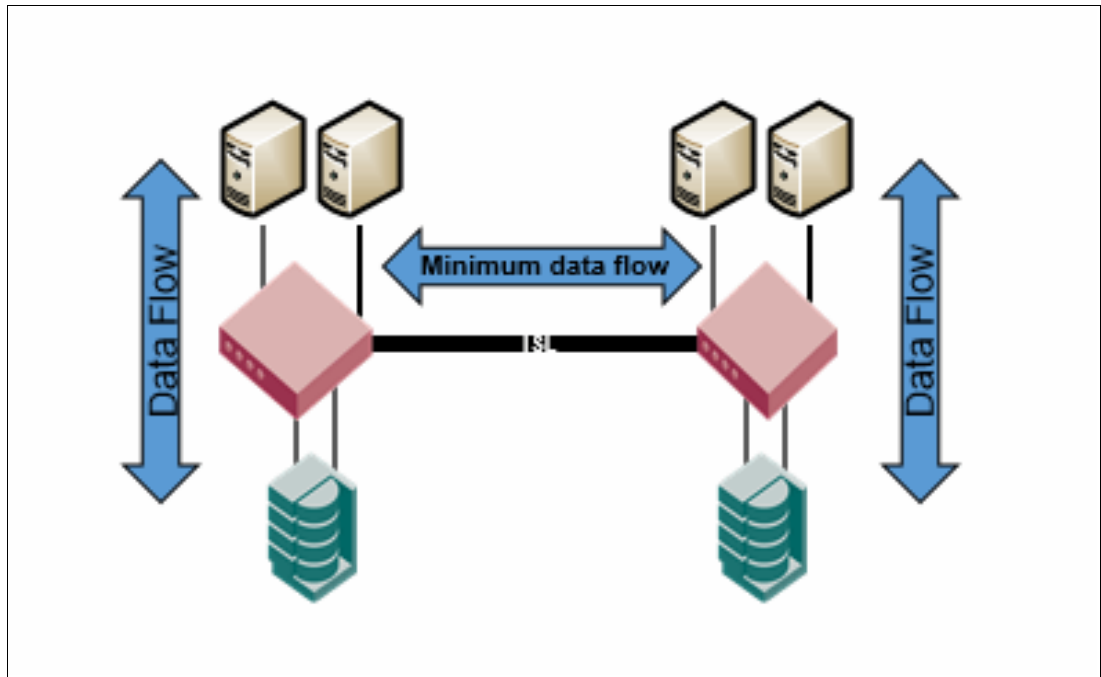


Figure 1-7 Storage and hosts attached to the same SAN switch

This solution can fit perfectly to small and medium SANs, however 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 on page 11 shows an example of traffic segregation on the SAN using edge-core-edge topology.

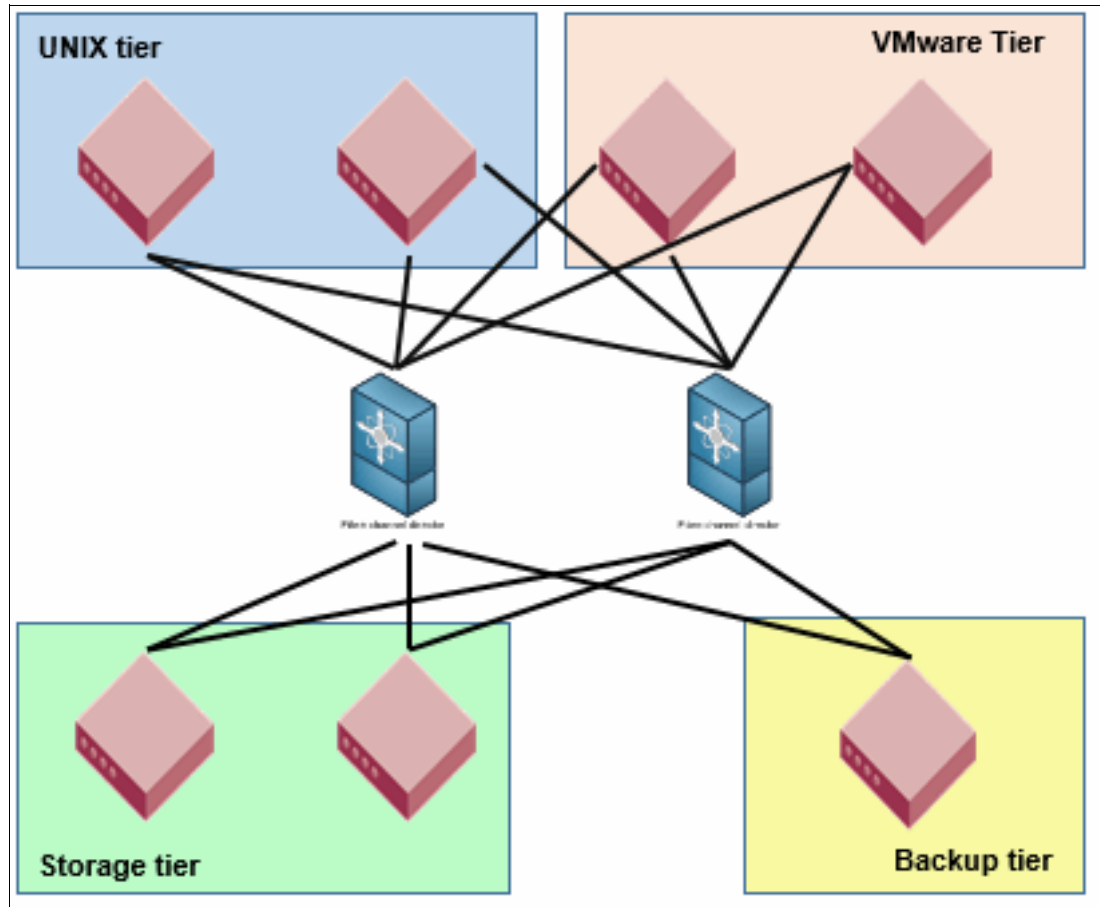


Figure 1-8 Edge-core-edge segregation

Even sharing the same core switches, it is possible to use Virtual Fabrics to isolate one tier from the other and, for example, avoid traffic congestion caused by slow drain devices connected to the backup tier switch.

Virtual Fabric Topology

Virtual Fabric (VF) is a hardware-level feature that permits SAN switches to share hardware resources by partitioning its hardware into different and isolated logical switches.

Hardware-level fabric isolation is accomplished through the concept of a logical switch, which provides the ability to partition physical switch ports into one or more “logical switches”. Logical switches are then connected to form logical fabrics. As the number of available ports on a switch continues to grow, partitioning switches gives storage administrators the ability to take advantage of high-port-count switches by dividing physical switches into different logical switches. Without VF, an FC switch is limited to 256 ports. A storage administrator can then connect logical switches through various types of ISLs to create one or more logical fabrics.

For additional information about virtual fabric, see Brocade Fabric OS Administrator's Guide version 7.4, which is available at:

<https://ibm.biz/BdsRRG>

1.3 SAN Volume controller ports

IBM SAN Volume controller hardware has significantly increased port connectivity options. Both 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. We will describe some best practices and use cases showing how to connect a SAN volume controller on the SAN to utilize this.

1.3.1 Slots and ports identification

The SAN volume controller can have up to up to four quad Fibre Channel (FC) HBA cards (16 FC ports) per node. The rear view of the 2145-SV1 node in Figure 1-9 shows the port location and Figure 1-10 on page 13 show the 2145-DH8 node port and slot locations.

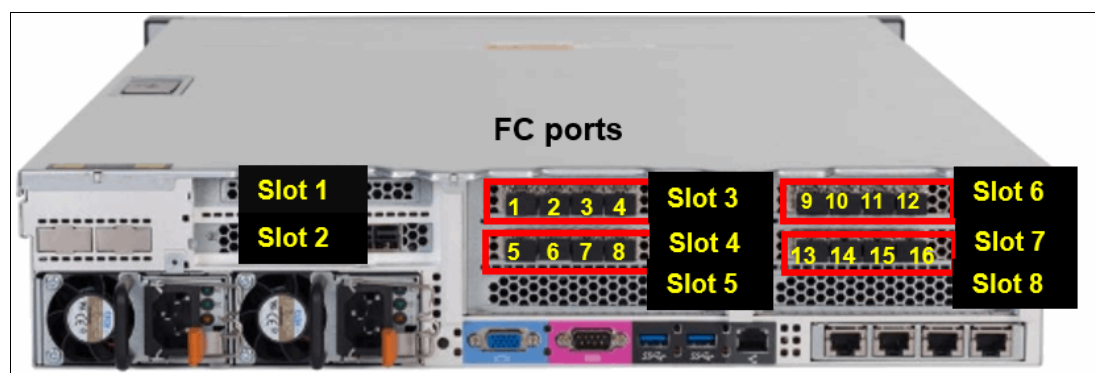


Figure 1-9 SAN Volume Controller 2145-SV1 rear port view

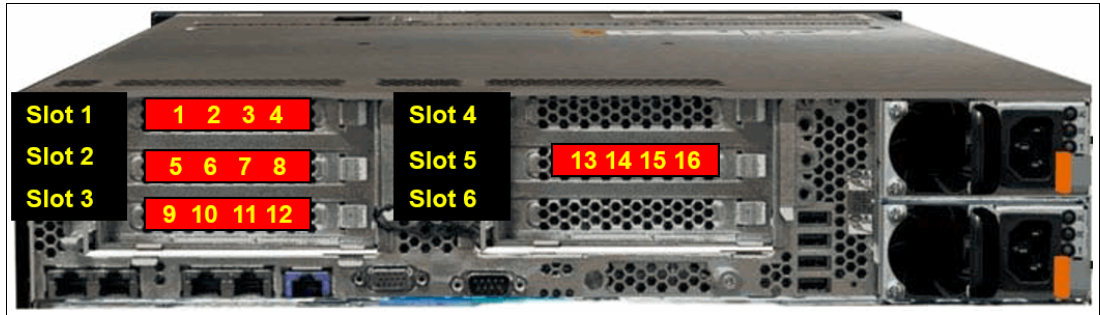


Figure 1-10 SAN Volume controller 2145-DH8 rear port view

To obtain maximum redundancy and resiliency it is recommended to spread the ports across different fabrics. As the port count vary, 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 the fabric naming convention vary, however it is common to find fabrics named, for example, “PROD_SAN_1” and “PROD_SAN_2” or “PROD_SAN_A” or “PROD_SAN_B”. Such a naming convention is used to simplify the SAN, once their denomination followed by “1” and “2” or “A” and “B” specifies that the devices connected to those fabrics contains the redundant paths of the same servers and SAN devices. So the suggestion to simplify the SAN connection identification and troubleshooting, is to 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.

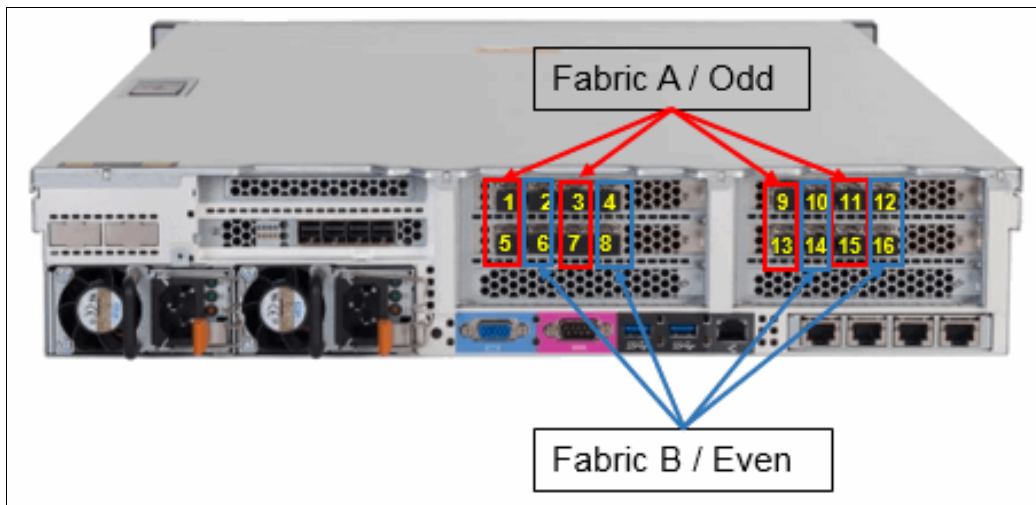


Figure 1-11 SAN Volume controller model 2145-SV1 Port distribution

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 best practice you must assign specific uses to specific SAN volume controller ports. By doing so it will optimize the port utilization by aligning the internal allocation of hardware CPU cores and software I/O threads to those ports. Apart from that, configuring the ports for specific uses will ensure the ability to replace nodes non-disruptively in the future.

Figure 1-12 on page 14 shows the specific port use guidelines for the 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	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	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
remotefcportmask			1100		01100000	01001000	001001000000
* 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 recommendation for the 2145-SV1 is similar, however the significant change is on the slot placement, and that there is no dual port HBA configuration option as shown in Figure 1-13 on page 14.

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, it is even possible to segregate the port assignment between hosts and storage, isolating their traffic. However, once host and storage ports have different traffic behavior, the recommendation is to not separate the host and storage ports. Keeping host and storage ports together extracts maximum port performance and utilization by benefiting from its full duplex bandwidth.

Important: Use port masking to assign specific uses to the SAN Volume Controller ports. For additional information refer to Chapter 6, “Hosts” on page 225.

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

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/Storwize requires three different zonings:

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

There are different guidelines to each zoning type. These guidelines will be detailed later on this chapter.

Note: Although internode/intra-cluster zone is not necessary for Storwize family systems, it is highly recommended 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.

Attention: Avoid the use of a zoning configuration that has a mix of port and worldwide name zoning.

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 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 302.

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 cause 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 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 1-15 on page 17 shows some alias examples.

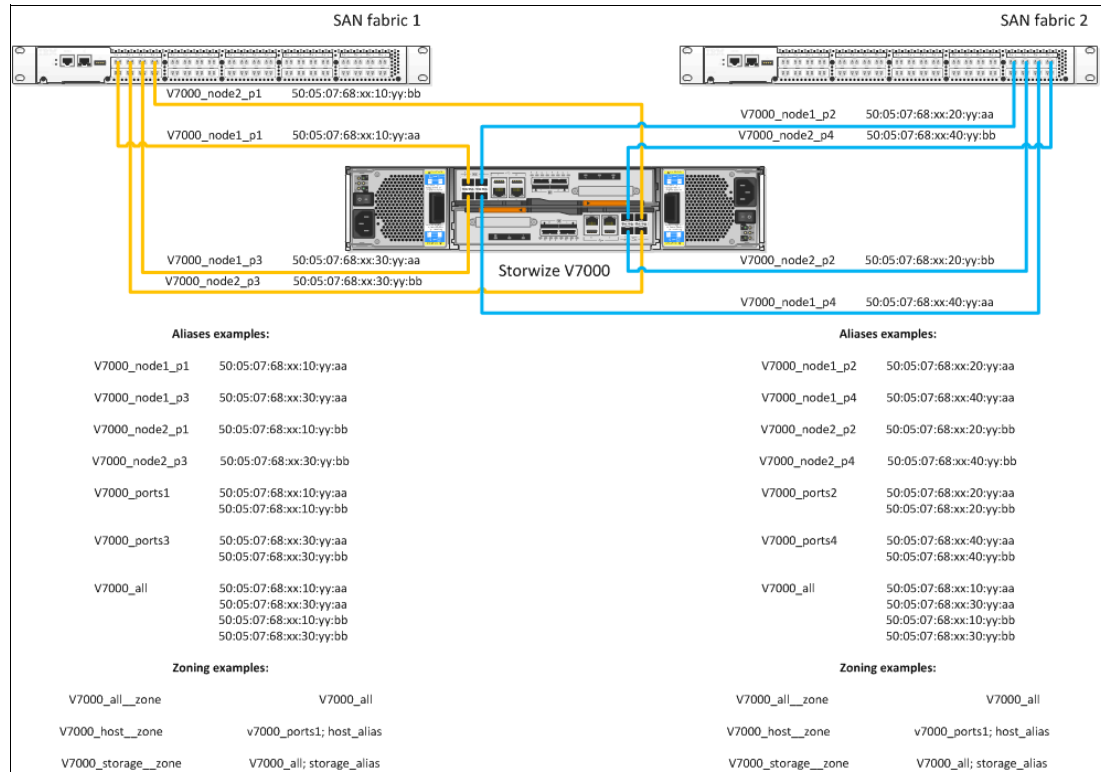


Figure 1-15 Different SAN Volume Controller/Storwize aliasing examples

1.4.3 SAN Volume Controller internode communications zone

The ports dedicated to internode communication are used for mirroring write cache and metadata exchange between nodes and they are critical to the stable operation of the cluster.

To establish efficient, redundant and resilient internode communication, the internode zone must contain at least two ports from each node. For SVC nodes with 8 ports or more, it is recommended to isolate the intracluster traffic by dedicating node ports specifically to internode communication. The ports to be used for internode communication varies according to the machine type-model number and port count. Refer to Figure 1-12 on page 14 (2145-DH8) and Figure 1-13 on page 14 (2145-SV1) for port assignment recommendations.

Only sixteen 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 port, member of the same zone, will be used for internode 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 into a port on node Y. This must not exceed 16. For additional port masking information refer to Chapter 6, “Hosts” on page 225.

1.4.4 SAN Volume Controller/Storwize storage zones

The zoning between SAN Volume Controller/Storwize and other storage is necessary to permit the virtualization of any storage space under the SAN Volume Controller/Storwize. We refer to this storage as backend storage.

A zone for each backend storage to each SAN Volume controller/Storwize node/canister must be created in both fabrics as shown in Figure 1-16 on page 18. By doing so it will reduce the overheads associated with a large number of logins. The ports from the storage subsystem must be split evenly across the dual fabrics.

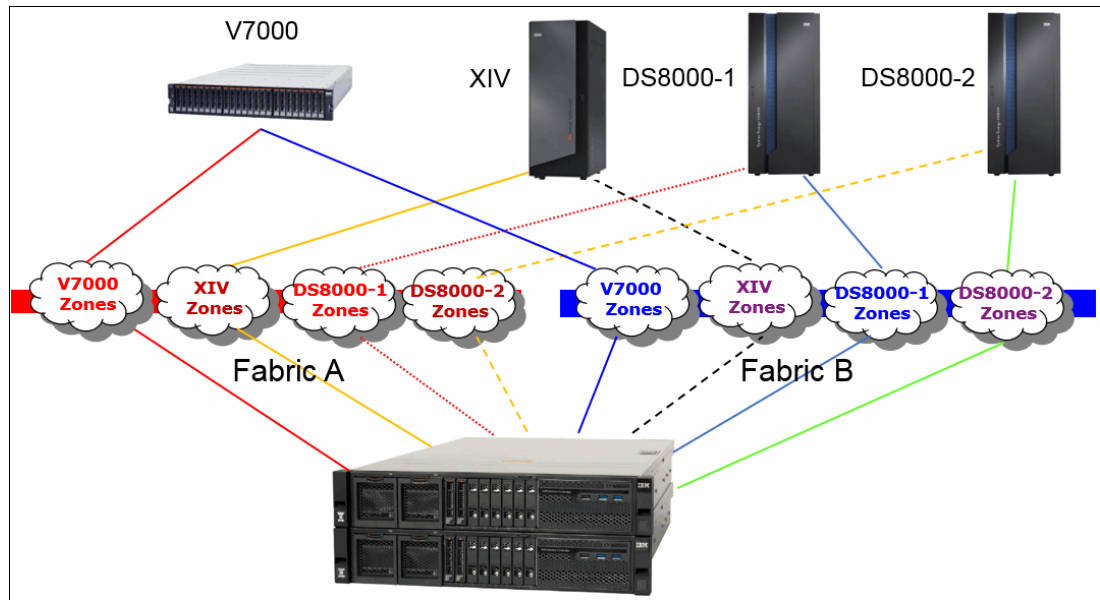


Figure 1-16 Back-end storage zoning

All nodes/canisters in a SAN Volume Controller/Storwize system must be zoned to the same ports on each backend storage system.

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, and the system will log errors that request a repair action. This situation can occur if inappropriate zoning is applied to the fabric.

Figure 1-17 on page 19 shows a zoning example (using generic aliases) between a two node SVC and a Storwize V5000. Notice that both SAN volume controller nodes has access to the same set of Storwize V5000 ports:

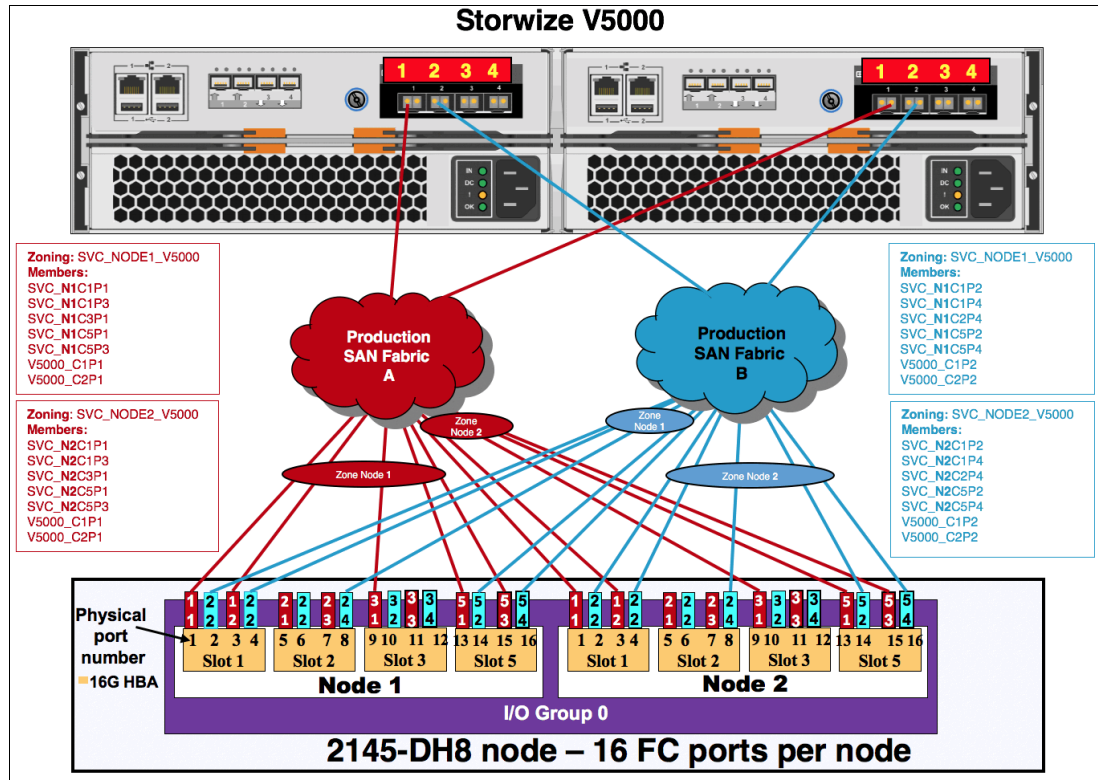


Figure 1-17 Storwize V5000 to SAN Volume controller zoning

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

Backend storage port count

The current firmware available (v7.8 at the time of writing), sets the limitation of 1024 WWNNs per SAN Volume Controller/Storwize cluster and up to 1024 WWPNNs. The rule is that each port represents a WWPN count on the SVC cluster, however the WWNN count will differ according to the storage.

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

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

The best practice recommendation is to assign up to sixteen (16) ports from each backend storage to the SAN Volume Controller/Storwize cluster. The reason of this limitation is that with v7.8 the maximum number of ports recognized by the SAN Volume controller/Storwize per each WWNN is sixteen. The more ports assigned, the more throughput will be obtained.

In a situation where your backend has hosts direct attached, do not mix the host ports with the SAN Volume Controller/Storwize ports. The storage backend ports must be dedicated to

the SAN Volume Controller/Storwize so, sharing storage ports are only reasonable during migration and for a limited time, however if you intend to have some hosts permanently directly attached to the backend storage you must segregate the SAN Volume Controller ports from the host ports.

XIV storage subsystem

XIV® storage is modular storage and is available as fully or partially populated configurations. XIV hardware can be presented from 6 and up to 15 modules configuration. Each additional module added to the configuration increase the XIV capacity, CPU, memory and connectivity.

From a connectivity standpoint there are four Fibre Channel ports 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 accordingly 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 instance, if a 9-module machine 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 choose 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, so 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 additional details refer to *IBM XIV Storage System Architecture and Implementation*, SG24-7659.

Figure 1-18 on page 21 shows how to connect an XIV frame to a SAN Volume Controller storage controller.

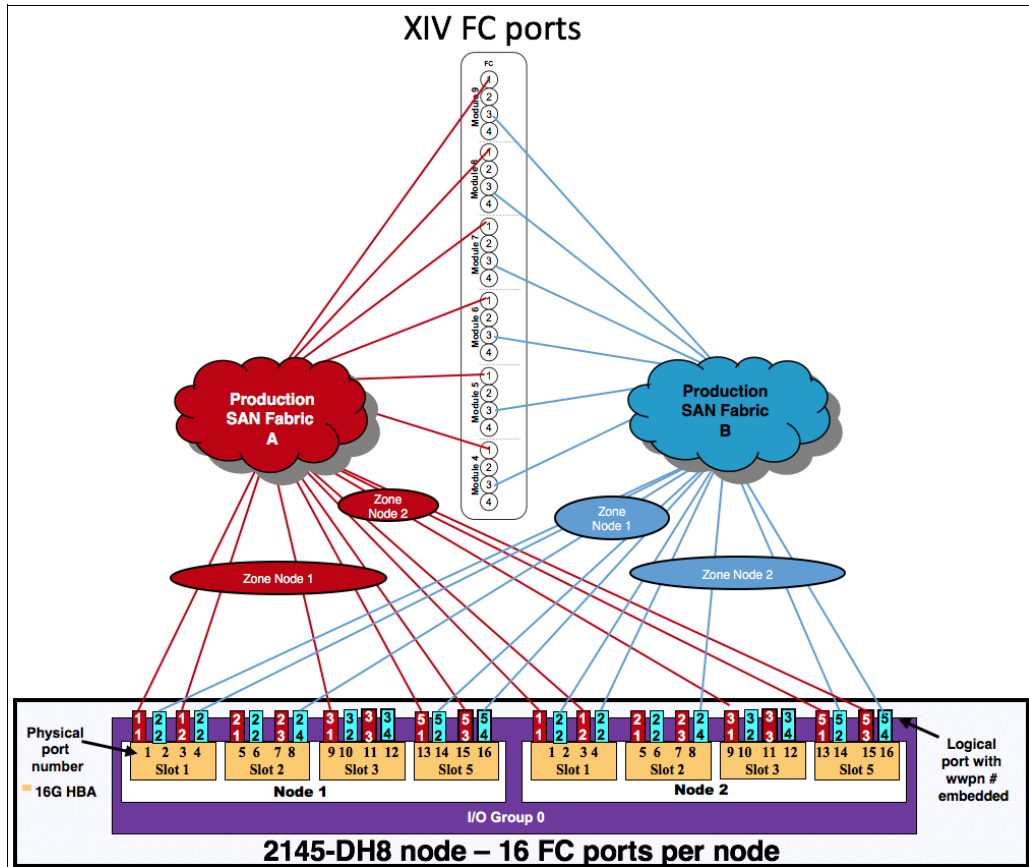


Figure 1-18 Connecting a XIV as a backend controller

The best practices 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 destined to connect host/backend storage. All nodes in an SVC/Storwize V7000 cluster must be able to see the same set of XIV host ports.

Notice on Figure 1-18 that there are a single zone to each XIV to SAN Volume controller node. So for this example there are 4 different zones:

1. Fabric A, XIV → SVC Node 1: All XIV fabric A ports to SVC node 1.
2. Fabric A, XIV → SVC Node 2: All XIV fabric A ports to SVC node 2.
3. Fabric B, XIV → SVC Node 1: All XIV fabric B ports to SVC node 1.
4. Fabric B, XIV → SVC Node 1: All XIV fabric B ports to SVC node 2.

Additional best practices and XIV consideration are detailed in Chapter 2, “Back-end storage” on page 37.

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 another Storwize, you must change the *layer* of the Storwize to be used as virtualizer. By default, SAN Volume Controller includes the layer of *replication* and Storwize 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 differences between the storage layer and the replication layer are:

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 means 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.

Additional information regarding storage layer can be found at:

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

To zone the Storwize as a backend 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 1-19 shows a zone between a 16 port Storwize V7000 to a SAN Volume Controller:

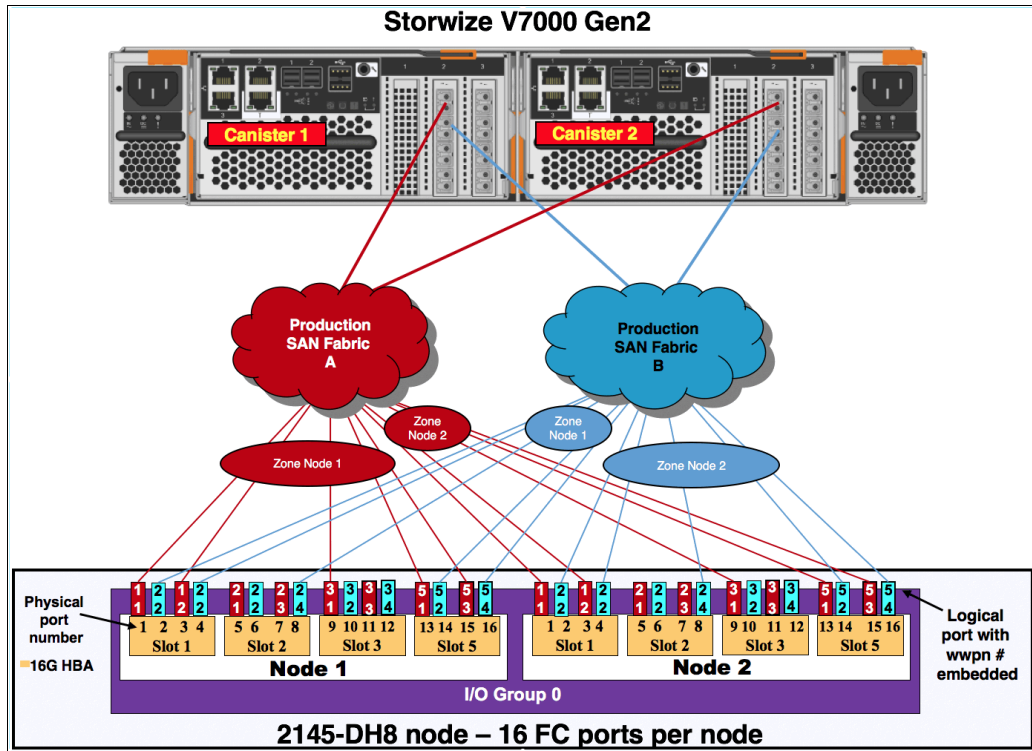


Figure 1-19 Storwize V7000 as a backend controller zone.

Notice that the ports from Storwize V7000 in Figure 1-19 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 as the mean time between failures (MTBF) of the adapters is not significantly less than that of the non-redundant canister components.

Connect to the SAN Volume controller as many ports as necessary to service your workload. For additional backend port limitations and recommendations refer to “Backend storage port count” on page 19.

FlashSystem 900

IBM FlashSystem® 900 is an all-flash storage array that provides extreme performance and is capable of sustaining highly demanding throughput and low latency across its FC interfaces. It includes up to 16 ports of 8 Gbps or 8 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 we recommend that you work closely with your IBM Sales, pre-sales and or IT architect to properly size the solution by defining the proper number of SAN Volume Controller I/O groups/cluster and FC ports that will be 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 13. 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:

- ▶ 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).
- ▶ Create one zone for each SAN Volume Controller/Storwize node with all FlashSystem 900 ports on each fabric.

Figure 1-20 show a 16 port FlashSystem 900 zoning to a SAN Volume Controller:

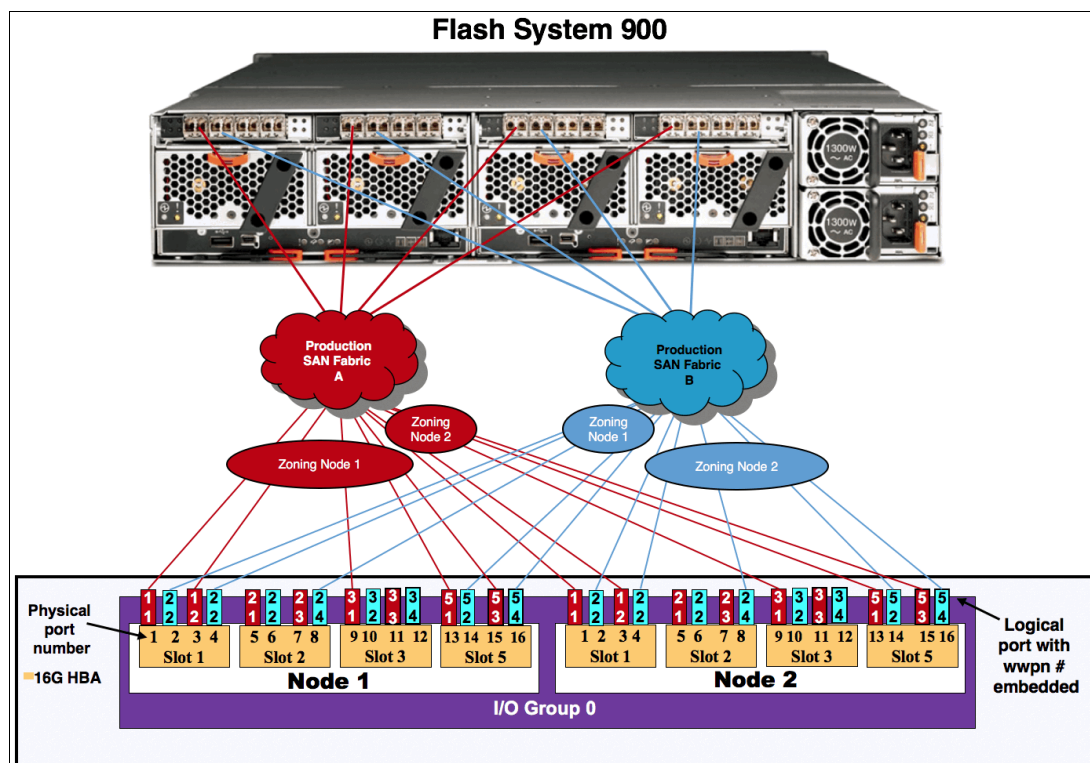


Figure 1-20 FlashSystem 900 to SAN Volume Controller zone

Notice that once the FlashSystem 900 is zoned to two SAN Volume Controller nodes there are a total of 4 zones: 1 zone per node and 2 zones per fabric.

You can decide to share or not the SAN Volume Controller/Storwize ports with other backend 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 37.

IBM DS88xx

The IBM DS8000 family is a high-performance, high capacity, highly secure, and resilient series of disk storage systems. The DS88xx 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 very 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 1-3.

Table 1-3 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 regarding DS888x hardware, port and connectivity refer to *IBM DS8880 Architecture and Implementation (Release 8.1)*, SG24-8323.

Despite the wide DS888x port availability, to attach a DS8880 series to a SAN Volume Controller you must use from 2 up 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 at the SAN Volume controller support website:

<http://www.ibm.com/storage/support/2145>

Figure 1-21 on page 26 shows the connectivity between a SAN Volume Controller and a DS8886.

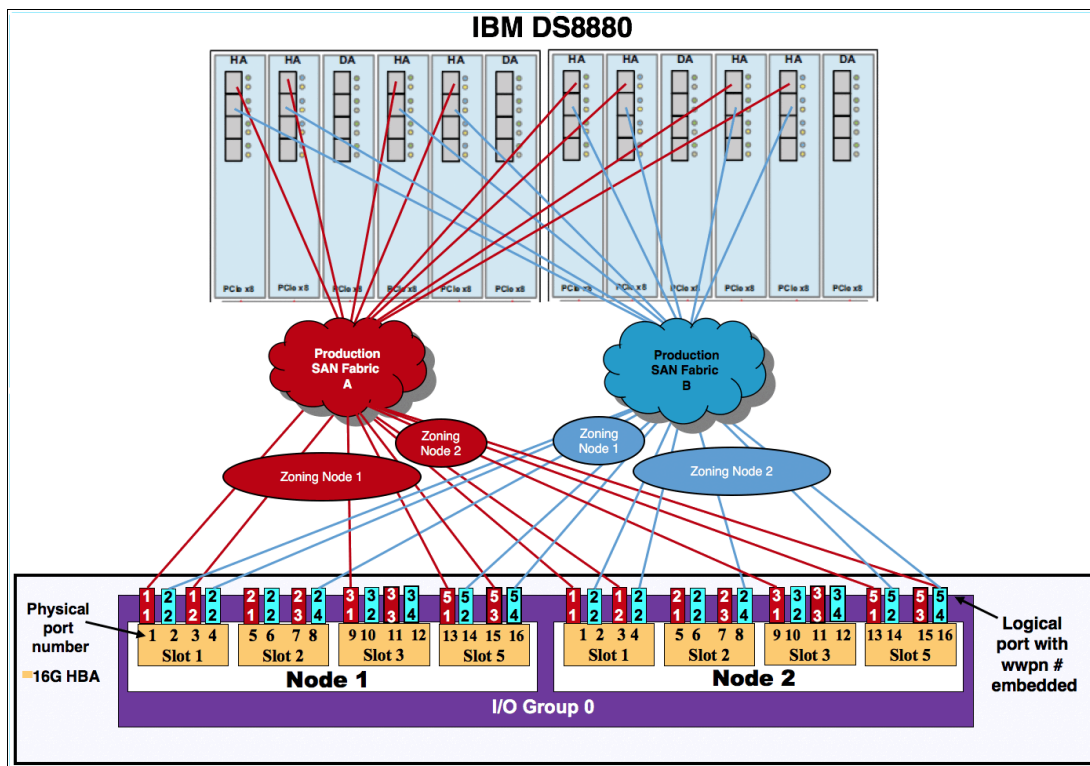


Figure 1-21 DS8886 to SAN volume controller connectivity

Notice that in Figure 1-21 there are 16 ports zoned to the SAN Volume Controller and the ports are spread across the different HAs available on the storage.

To maximize the performance that you can achieve 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 13.

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 will be set to degraded on the SAN Volume Controller.

After the zoning steps you must configure the *host connections*, via DS888x CLI (DSCLI) or GUI, to all SVC nodes WWPNs so create a single *Volume Group* adding all SVC cluster ports within this *Volume Group*. For additional information regarding *Volume Group*, *Host Connection* and DS8000 administration refer to *IBM DS8880 Architecture and Implementation (Release 8.1)*, SG24-8323.

The specific best practices to present DS8880 LUNs as backend storage to the SAN Volume Controller are detailed in Chapter 2, “Back-end storage” on page 37.

1.4.5 SAN Volume Controller/Storage host zones

The best practice to connect a host into a SAN volume Controller/Storage consist of creating a single zone to each host port. This zone must contain the host port and *one* port from each SAN Volume Controller/Storage 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 1-22 on page 27.

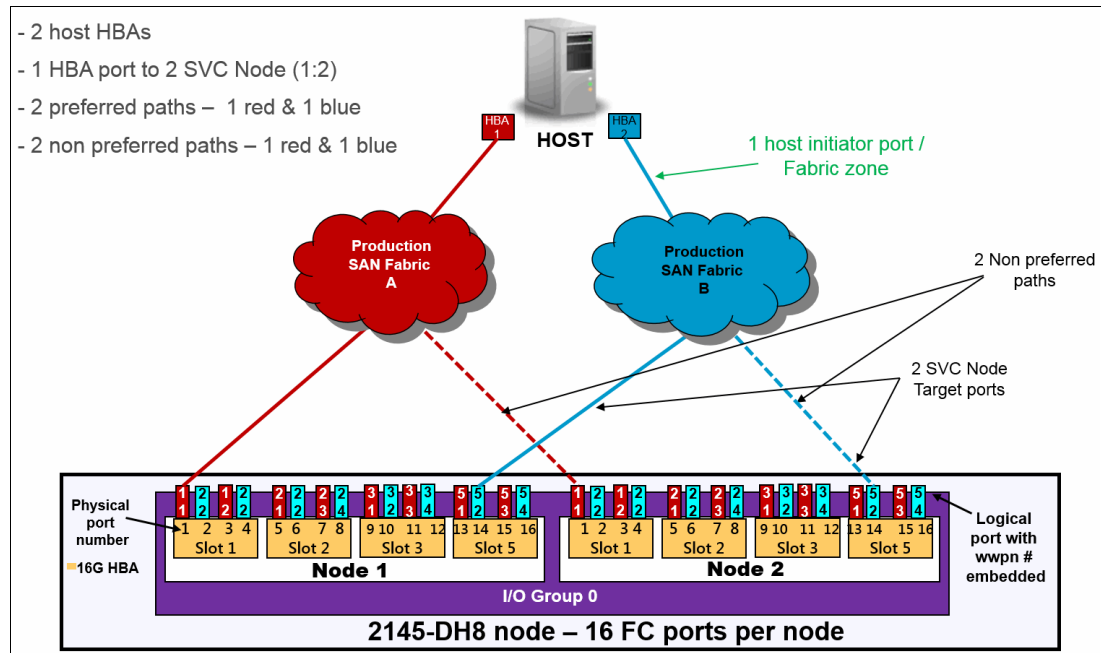


Figure 1-22 Typical host to SAN Volume Controller zoning

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/Storage, are optimized to work with.

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, possibly reduces performance. Also, it does not significantly improve reliability nor availability. On the other hand, less 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/Storage 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/Storage FC port.

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 master 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 load balance by I/O group instead so that the volume is assigned to nodes automatically.

- 2 host dual HBAs (4 ports)

- Still 1 HBA port to 2 SVC Node (1:2)

- Create 2 SVC Hosts to the same server

Ex. ITSO_HOST_1 (Red and Orange ports)

ITSO_HOST_2 (Blue and Green ports)

ITSO_HOST

Production SAN Fabric A

Production SAN Fabric B

Node 1

Node 2

I/O Group 0

2145-DH8 node – 16 FC ports per node

Physical port number

16G HBA

Logical port with wwpn # embedded

Figure 1-23 Four port host zoning

Because the optimal number of volume paths is four, you must create two or more hosts on SAN Volume Controller/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/Storwize host configuration. Instead of creating one host id with four WWPNs, you define two hosts with two WWPNs.

For ESX Clusters you must create separate zones for each host node in the ESX Cluster as shown in Figure 1-24 on page 29.

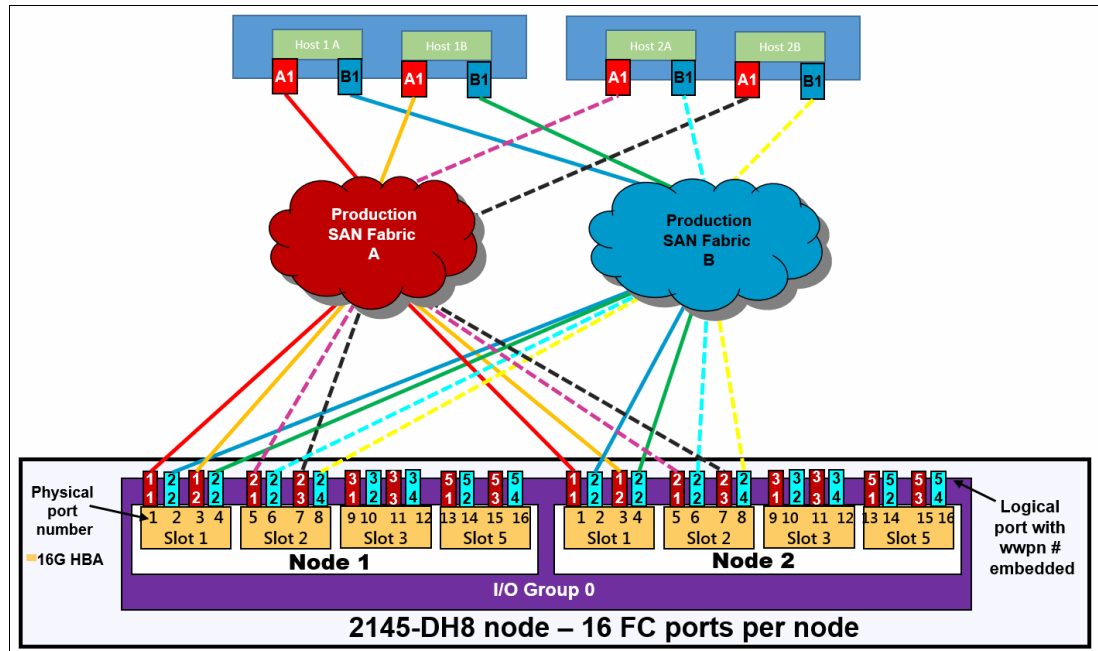


Figure 1-24 ESX Cluster zoning

Ensure that you apply the following best 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 IO Groups.
- ▶ Map LUNs/volume evenly across zoned ports, alternating the preferred node paths evenly for optimal I/O spread & 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 1-25 on page 30:

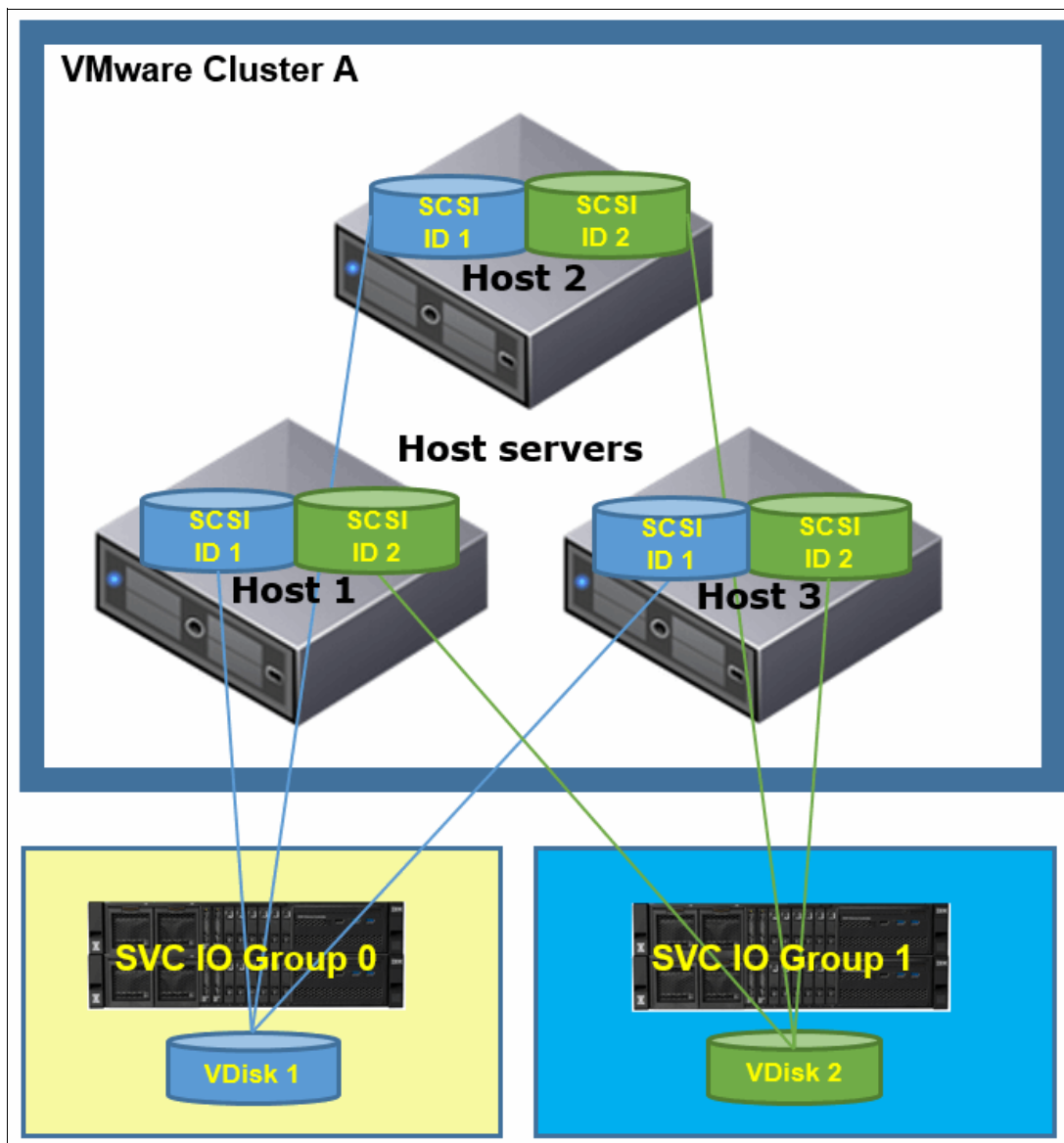


Figure 1-25 LUN/volume mapping to clustered ESX hosts

AIX Vios - LPM zoning

When zoning AIX® VIOs to Spectrum Virtualize you must plan carefully. Because of its complexity it is common to create more than 4 paths to each Volume/MDisk, or perhaps not provide for proper redundancy. The following best practices will help you to have a non-degraded path error on Spectrum Virtualize/Storwize and keep 4 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 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, it is recommended to make no more than a ratio of 1 physical adapter to 8 Virtual ports. It will avoid I/O bandwidth oversubscription to the physical adapters.

- ▶ Create a pseudo host in Spectrum Virtualize/Storage host definitions containing only two virtual WWPNs, one from each fabric as shown in Figure 1-26.
- ▶ Map the LUNs/volumes to the pseudo LPARs (both the active and inactive) in a Round Robin fashion.

Figure 1-26 shows a correct SAN connection and zoning for LPARs:

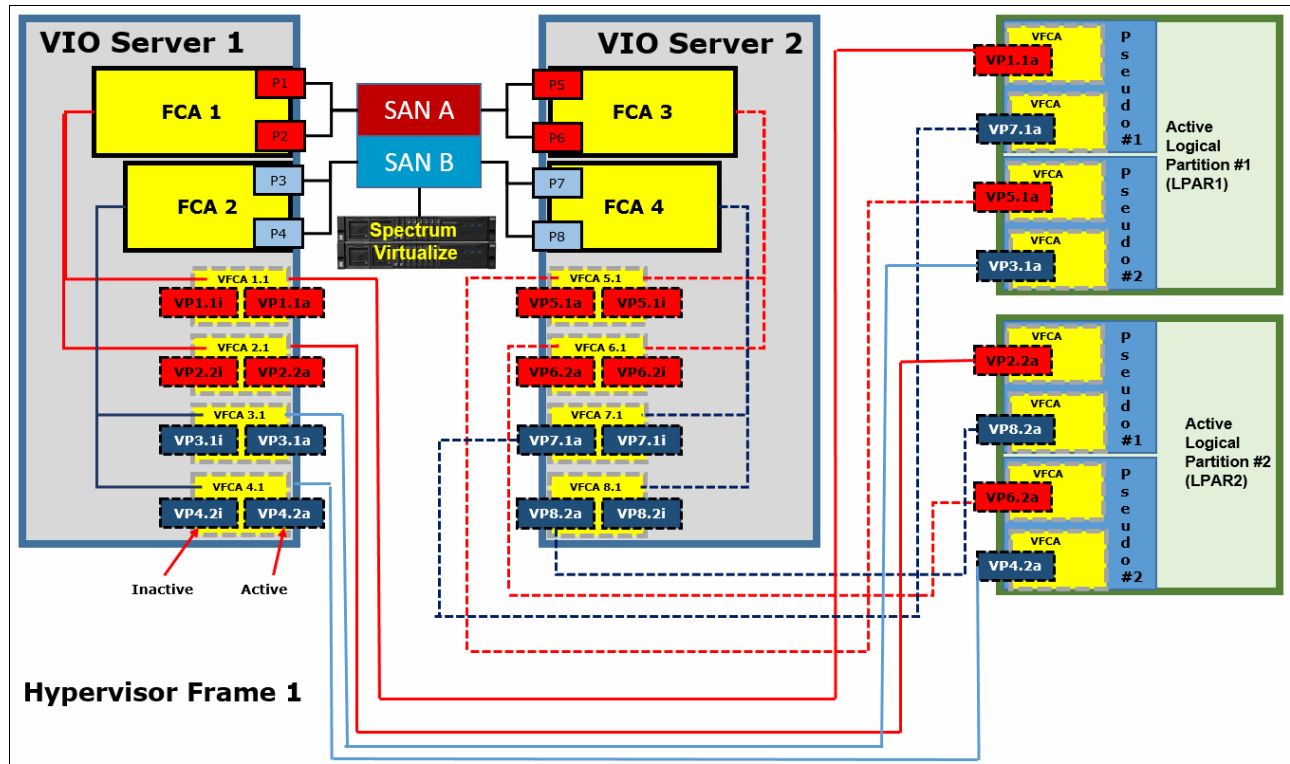


Figure 1-26 LPARs SAN connections

During Live Partition Migration (LPM) both inactive and active ports will be active. Upon the LPM completion the previously active ports will show as inactive and the previous inactive ports will show as active. Figure 1-27 on page 32 shows a Live partition migration from the hypervisor frame to another frame.

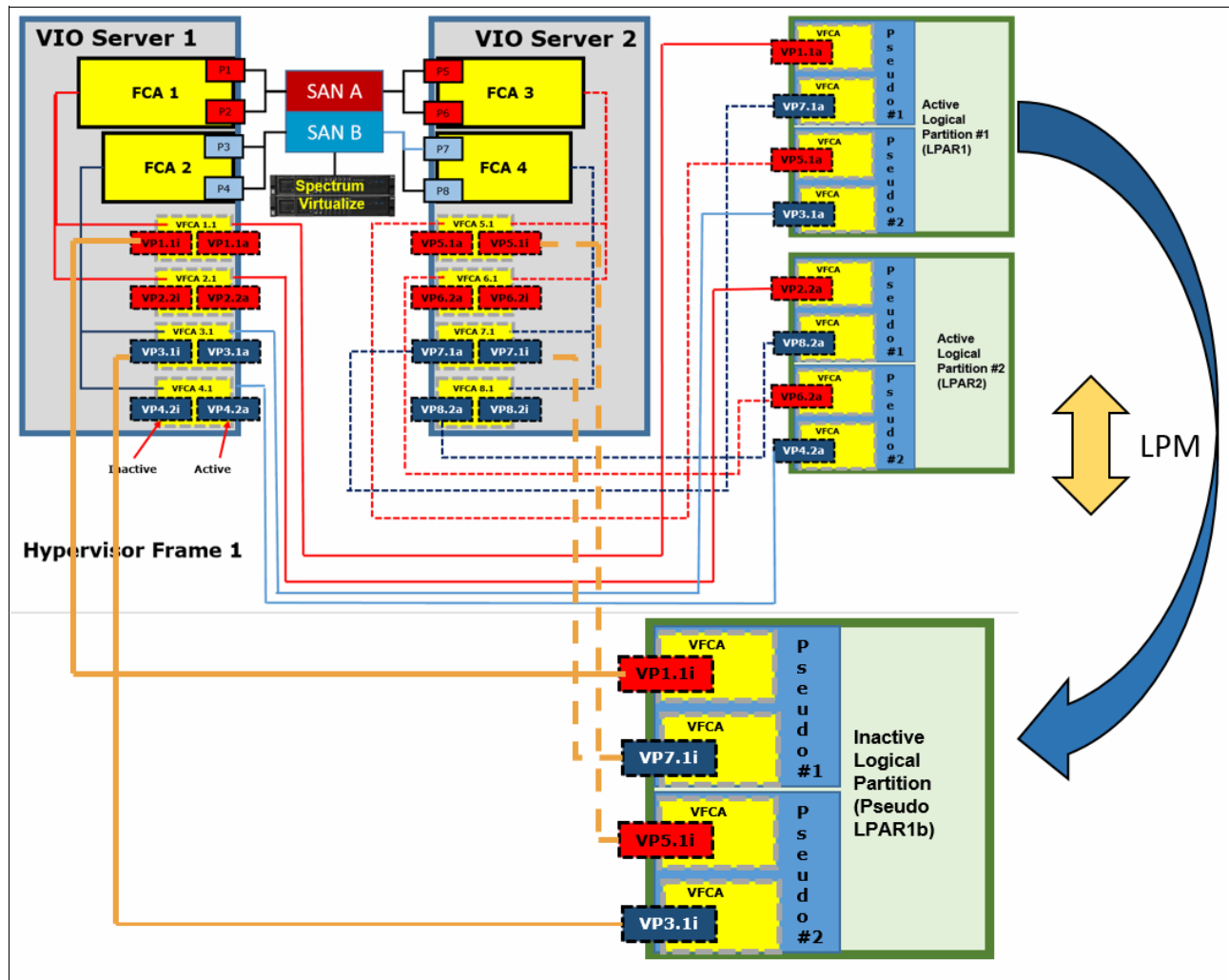


Figure 1-27 Live partition migration

Note: During LPM the number of paths doubles from 4 to 8. Starting with 8 paths per LUN/volume will render an unsupported 16 paths during LPM and that could lead to IO interruption.

1.4.6 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 backend storage zones must also be separate, even if the two clustered systems share a storage subsystem. You also must zone separate IO groups if you want to connect them in one clustered system. Up to four IO groups can be connected to form one clustered system.

1.4.7 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; therefore, it is unlikely that Cisco SFPs work in a Brocade 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 via the TCP/IP protocol. In most cases, the FCIP is implemented in DRC scenarios with some kind of

data replication between primary and secondary site. The 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 via FCIP link. To use FCIP, you need some kind of tunneling device on both sides of the TCP/IP link, such as FCIP blade in the SAN256B-6/SAN512B-6 directors or SAN06B-R router or SAN42B-R extension switch. 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:
 - **gmlinktolerance**
 - **gmmaxhostdelay**
 - **gminterdelaysimulation**
 - **gmintradelaysimulation**
- ▶ 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 Global Mirror with Change Volumes (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.
- ▶ Use VE and VEX ports on FCIP routers 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
- ▶ *Brocade Fabric OS Administrator's Guide version 7.4*

1.5.4 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 TCP/IP network and not on the FC network. For more information about native IP replication, see Chapter 5, “Copy Services” on page 131.

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 available at:

<http://www.redbooks.ibm.com/abstracts/redp5103.html>

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 Brocade switches are not recommended, except during fabric migrations and 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:

- ▶ 2.1, “Round Robin Path Selection” on page 38
- ▶ 2.2, “Considerations for DS88xx series” on page 39
- ▶ 2.4, “Considerations for IBM Storwize V7000/V5000/V3700” on page 47
- ▶ 2.5, “Considerations for IBM FlashSystem 900” on page 51
- ▶ 2.6, “Considerations for third-party storage with EMC VMAX and Hitachi Data Systems” on page 56

2.1 Round Robin Path Selection

Before V6.3 of SAN Volume Controller/Storwize code, all I/O to a particular MDisk was issued through only one backend storage controller FC port. Even if there were 12 (XIV) or 16 (DS88xx) FC ports that were zoned to SAN Volume Controller/Storwize, one MDisk was using only one port. If there was a port failure, another port on backend storage controller was chosen.

This changed in SAN Volume Controller/Storwize code v6.3. Since v6.3, each MDisk used one path per target port per SAN Volume Controller/Storwize node. This means that, in cases of storage systems without a preferred controller such as XIV or DS88xx, each MDisk uses all 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 IO 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 MDisk is sent via other ports. Moreover, because IOs to MDisk are sent via all backend storage FC ports the port failure can be detected more quickly.

Best 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.

At the time of writing, the Round Robin Path Selection is supported on the following storage systems:

- ▶ IBM Storwize V3700, V5000, and V7000
- ▶ IBM FlashSystem 720, 820, 840 and 900
- ▶ IBM DS8100, DS8300, DS8700, DS8800, and DS8870
- ▶ IBM XIV, A9000 and A9000R
- ▶ EMC Symmetrix (including DMX and VMAX), XtremIO
- ▶ Fujitsu Eternus
- ▶ Oracle Flash Storage Systems
- ▶ PureStorage
- ▶ Violin Memory 3100, 3200, and 6000

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 this website:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1003658>

2.2 Considerations for DS88xx series

Although all recommendations in this chapter are true for SAN Volume Controller and Storwize family storage systems, the DS88xx series might be virtualized behind one of Storwize storage systems for a short time, or even longer. Sometimes, it might be virtualized by Storwize only for data migration purposes.

To optimize the DS8880 resource utilization, use the following guidelines:

- ▶ Distribute capacity and workload across device adapter pairs.
- ▶ Balance the ranks and extent pools between the two DS88xx internal servers to support the corresponding workloads on them.
- ▶ Spread the logical volume workload across the DS88xx 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).

2.2.1 Balancing workload across DS88xx series controllers

When you configure storage on the DS88xx 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, a considerable performance degradation can result from uneven device adapter loading.

The DS88xx series controllers assigns 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 on page 40 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.

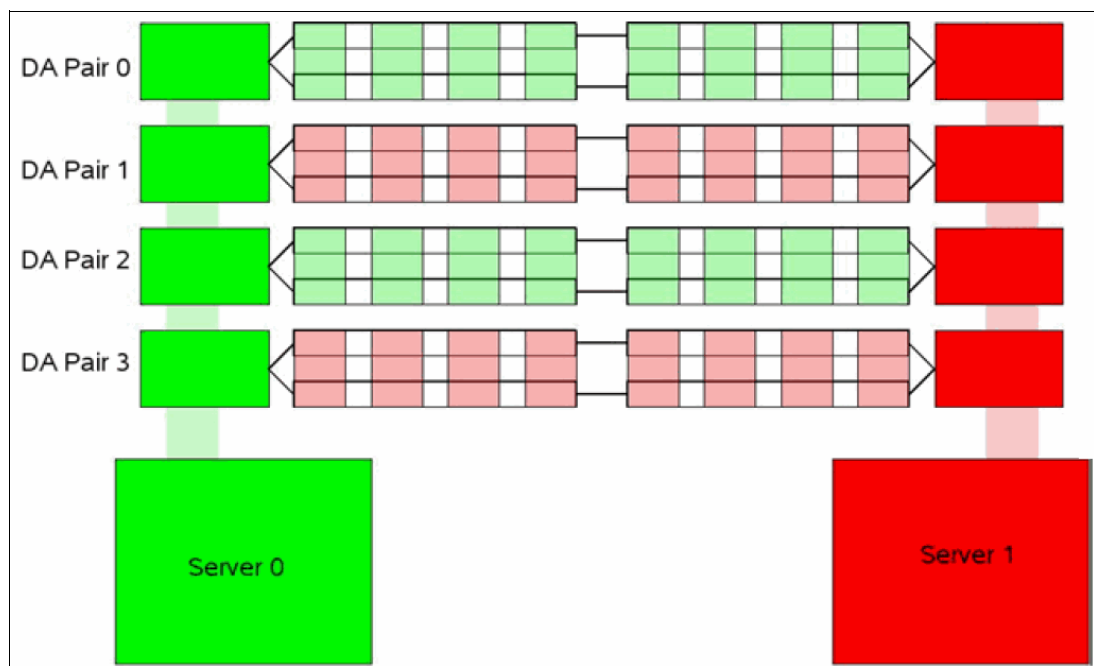


Figure 2-1 DA pair reduced bandwidth configuration

Example 2-1 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 DS88xx 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 2-1 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 (6+P+S) S1 R0 0 146.0 ENT
A1 Assign Normal 5 (6+P+S) S9 R1 1 146.0 ENT
A2 Assign Normal 5 (6+P+S) S17 R2 2 146.0 ENT
A3 Assign Normal 5 (6+P+S) S25 R3 3 146.0 ENT
A4 Assign Normal 5 (6+P+S) S2 R4 0 146.0 ENT
A5 Assign Normal 5 (6+P+S) S10 R5 1 146.0 ENT
A6 Assign Normal 5 (6+P+S) S18 R6 2 146.0 ENT
A7 Assign Normal 5 (6+P+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 on page 41 shows a configuration that balances the workload across all four DA pairs.

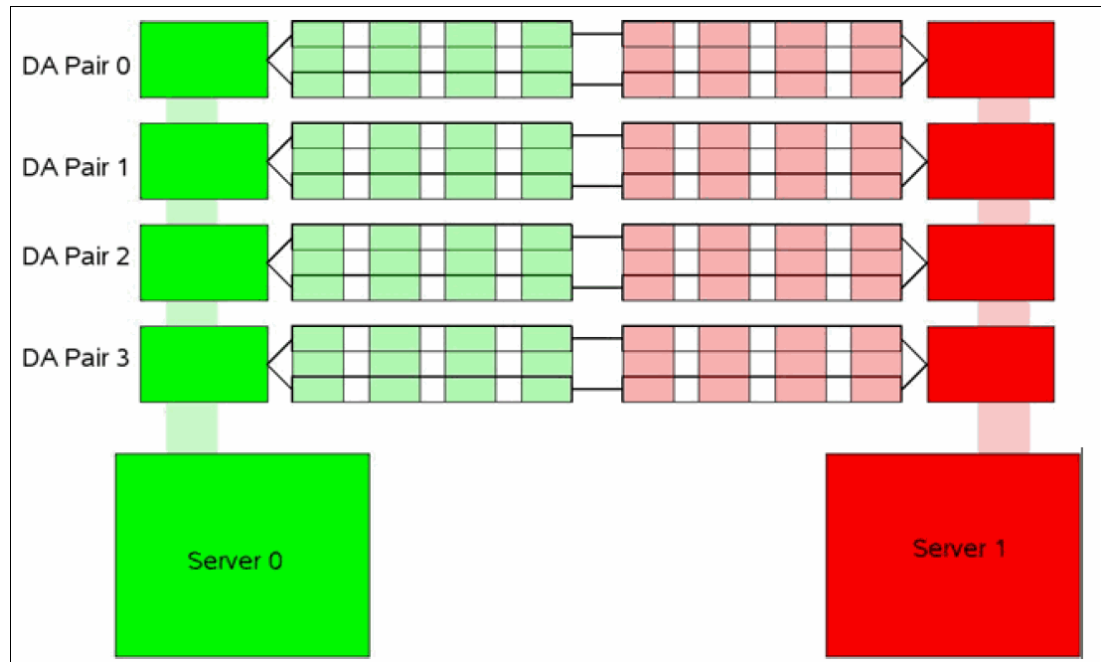


Figure 2-2 DA pair correct configuration

Figure 2-3 shows how 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 extpool10 fb 6348 6348 - MA1
R1 1 Normal Normal A1 5 P1 extpool11 fb 6348 6348 - MA2
R2 0 Normal Normal A2 5 P2 extpool12 fb 6348 6348 - MA3
R3 1 Normal Normal A3 5 P3 extpool13 fb 6348 6348 - MA4
R4 1 Normal Normal A4 5 P5 extpool15 fb 6348 6348 - MA5
R5 0 Normal Normal A5 5 P4 extpool14 fb 6348 6348 - MA6
R6 1 Normal Normal A6 5 P7 extpool17 fb 6348 6348 - MA7
R7 0 Normal Normal A7 5 P6 extpool16 fb 6348 6348 - MA8

```

Figure 2-3 Lsarray and lsrank output

2.2.2 DS88xx series ranks to extent pools mapping

When you configure the DS88xx series storage controllers, you can choose from the following approaches for rank to extent pools mapping:

- ▶ Use one rank per extent pool
- ▶ Use multiple ranks per extent pool by using DS88xx series extent pool striping

The old approach is to map one rank to one extent pool. It ensures that all volume allocation, from the selected extent pool, comes from the same rank. Although this approach provides good control for volume creation, the rank load balance and the extents distribution across the ranks is manual, requiring the storage administrator to create volumes on the different extent pools (that in this case is the same as ranks and disk arrays), and also distribute the volume across the storage controller (since each extent pool has its own controller affinity).

The storage pool striping feature, on the other hand, permits you to stripe a single volume across all ranks (disk arrays) in an extent pool. The function is often referred as *extent rotation*. Therefore, if an extent pool includes more than one rank, a volume can be allocated by using free space from several ranks. Also, storage pool striping can be enabled only at volume creation and reallocation is possible only when using easy tier.

To use the storage pool striping feature, your DS88xx series layout must be well-planned from the initial configuration to guarantee that all resources in the DS88xx are in use and its workload is balanced across DAs and controllers. Otherwise, storage pool striping can cause severe performance problems in a situation where, for example, you configure a heavily loaded extent pool with multiple ranks from the same DA pair.

The use of extent pool striping can boost performance per MDisk and this is the recommended method for extent pool configuration.

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

Additional recommendations regarding Ranks and Extent pools can be found in Chapter 3, “Storage pools and managed disks” on page 57.

In cases where you have different disk tiers (like flash, enterprise SAS and nearline SAS) on the DS88xx storage, you must create at least two extent pools for each tier in order 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 DS88xx pools. However, once virtualized, tier management has more advantages when handled by the SAN Volume Controller.

For additional information and recommendations regarding Easy Tier, refer to Chapter 7, “IBM Easy Tier function” on page 257.

2.2.3 Determining the number of controller ports for DS88xx series

Configure a minimum of eight controller ports to the SAN Volume Controller per controller, regardless of the number of nodes 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 cluster. Currently, 16 ports per storage subsystem are the maximum that is supported from the SAN Volume Controller side.

It is recommended to use ports from different host adapters and if possible from different I/O enclosures. This configuration is also important because during a DS88xx LIC update, a host adapter port might need to be taken offline. This configuration allows the SAN Volume Controller 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 will be virtualized on SAN Volume Controller:

- ▶ Use 8 ports to support up to 40 Ranks
- ▶ Use 16 ports (the maximum supported by SVC) for 40+ ranks.

For additional information of SAN best practices and connectivity refer to Chapter 1, “SAN topology” on page 1.

2.2.4 LUN masking

For a storage controller, all SAN Volume Controller nodes 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 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 DS88xx series controllers perform LUN masking that is based on the volume group. Example 2-2 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 2-2 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 2-3 shows output for the `lshostconnect` command from the DS88xx 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 2-3 Output for the lshostconnect command

```

dscli> lshostconnect -volgrp v0
Date/Time: Oct 22, 2016 10:45:23 AM BRT IBM DSCLI Version: 7.8.1.62 DS: IBM.2107-75FPX81
Name          ID    WWPN          HostType Profile          portgrp volgrpID ESSIOport
=====
ITSO_SVC_N1C1P4 0001 500507680C145232 SVC      San Volume Controller      1 V0      all
ITSO_SVC_N1C2P3 0002 500507680C235232 SVC      San Volume Controller      1 V0      all
ITSO_SVC_N2C1P4 0003 500507680C145231 SVC      San Volume Controller      1 V0      all
ITSO_SVC_N2C2P3 0004 500507680C235231 SVC      San Volume Controller      1 V0      all
dscli>

```

From Example 2-3 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; that is, direct-attached hosts.

Next, you see how the SAN Volume Controller 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 2-4 shows the general details of the output storage controller by using the SAN Volume Controller command-line interface (CLI).

Example 2-4 Output of the lscontroller command

```

IBM_2145:ITSO_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:ITSO_SVC:admin>

```

DS88xx LUN/MDisk Size considerations

As mentioned previously the maximum number of DS88xx ports to be presented to the SAN Volume Controller is 16. Each port represents a path to the SAN Volume Controller. So when sizing the number of LUN/MDisks to be presented to the SAN Volume Controller, the recommendation is to present, at least, from 2 to 4 volumes per path. So utilizing the maximum of 16 paths you should create 32, 48 or 64 DS88xx volumes, and for this configuration it will maintain a good queue depth.

For additional MDisk information refer to Chapter 4, “Volumes” on page 95.

2.3 Considerations for IBM XIV Storage System

When you configure the controller for the IBM XIV Storage System, you must remember the considerations that are described in this section.

2.3.1 Connectivity considerations

The XIV supports iSCSI and FC protocols; however, when you connect to SAN Volume Controller/Storwize, only FC ports can be used.

The best practices related to cabling and zoning are detailed in Chapter 1, “SAN topology” 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 WWNN, which identifies the entire XIV system.

Creating a host object for SAN Volume Controller/Storwize for an IBM XIV type 281x

A single host instance can be created for use in defining and then implementing the SAN Volume Controller/Storwize. By doing so, the host configuration is easier to configure. However, the ideal host definition for use with SAN Volume Controller/Storwize is to consider each node of the SAN Volume Controller/Storwize (a minimum of two) as an instance of a cluster.

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 Managed disks considerations

As modular storage, XIV storage can be presented from 6 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 on page 46 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 <u>active</u> 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 55 TB 84 TB 112 TB 169 TB	44 TB 88 TB 132 TB 177 TB 267 TB	51 TB 102 TB 154 TB 207 TB 311 TB	56 TB 111 TB 168 TB 225 TB 338 TB	63 TB 125 TB 190 TB 254 TB 382 TB	67 TB 134 TB 203 TB 272 TB 409 TB	75 TB 149 TB 225 TB 301 TB 453 TB	81 TB 161 TB 243 TB 325 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

Due to the different possibilities of the XIV capacity, the size of the volumes to be presented to the SAN Volume Controller/Storwize cluster will vary. The best practices related to XIV MDisks are described in Chapter 4, “Volumes” on page 95.

2.3.4 Additional considerations

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

MDisk mapping:

When mapping a volume to the SAN Volume Controller/Storwize you must use the same LUN ID to all SVC/Storwize nodes. 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 SVC/Storwize MDisk. 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.

2.4 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 into single larger environments for scalability reasons.
- ▶ Where SAN Volume Controller is already virtualizing other storage systems and more capacity is provided by 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 node). Therefore, adding the SAN Volume Controller on top can provide more caching capability, which is beneficial for cache-friendly workloads.
- ▶ Storwize V7000 with SSDs can be added to the SAN Volume Controller setup to provide Easy Tier capabilities at capacities that are larger than is possible with internal SAN Volume Controller SSDs. This setup is common with back-end storage that does not provide SSD disk capacity, or when too many internal resources are used for them.

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

Note: The 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.

2.4.1 Cabling and zoning

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 backend Storwize system is probably not used for this function as 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 best practices, refer to Chapter 1, “SAN topology” on page 1.

2.4.2 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 will be the most suited and appropriate for your environment. The recommended RAID type will vary according to the capacity, performance and protection level required. For instance, RAID5 can perform faster writes than RAID6, however RAID6 (double parity), provides protection in cases 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 will be, and so, the disk array will remain unprotected for a longer time (RAID 5 tolerates one disk failure at a time). Figure 2-5 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 2-5 Traditional RAID rebuild time.

Distributed RAID (DRAID) was introduced in V7.6 and is an improved RAID solution. It distributes and reserves disk spare space across all disks in the array. Since there are no idle disks to be used as spare, 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 array/MDisk: 128 disks.

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

- ▶ Up to 36 disk drives: 1 spare
- ▶ 37 to 72 disk drives: 2 spares
- ▶ 73 to 100 disk drives: 3 spares
- ▶ 101 to 128 disk drives: 4 spares

DRAID consumes more resources from the storage controller than traditional RAID, so keep in mind that the number of arrays/MDisks per I/O group will decrease and so, depending on the storage disk count/capacity, creating small DRAID arrays is not recommended, as you might run out of resources to create new arrays/MDisks.

Figure 2-6 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 2-6 DRAID and TRAIID arrays per system limitation

Note: The limitations in Figure 2-6 are related to the V7.8 firmware version. More information and limitations for other versions can be found at:

SAN Volume Controller:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1003658>

Storwize V7000:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1003741>

Storwize V5000:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004971>

Storwize V3700:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1004388>

When creating a new array, the default array width, on the storage GUI, is 9 disks for DRAID 5 and 12 disks for DRAID 6. However the recommendations for the array width may vary according to the number of disks and tiers the storage has.

For example, when using DRAID with NL-SAS disks, prefer a minimum of 24 disks per array/MDisk (1 enclosure), however an optimal disk utilization on DRAID is to keep the array width around 48 to 60 disks per array/MDisks. When using SSDs disks, you can consider reducing the number of disks on the array in accordance with the disk count in the storage.

When creating an MDisk/array the disk drives that should be part of the MDisk/array is not usually a concern for RAID5 or RAID6, however, the best practice is to build MDisks/Arrays with disk drives within enclosures in the same SAS chain.

The use of the GUI to configure internal storage is helpful when you configure Storwize as a general-purpose storage system for different hosts. However, when you want to use it only as backend storage for SAN Volume Controller or another front-end Storwize, it stops being general-purpose storage and starts to be SAN Volume Controller-specific storage. Because the GUI storage configuration wizard does not know that this storage is mapped to another

SAN Volume Controller/Storwize, it can create unbalanced arrays (when optimize for capacity is chosen) or it can leave some drives unconfigured (if optimize for performance is chosen). Therefore, if you know exactly what type of storage pools you want to have on SAN Volume Controller or front-end Storwize, it is better to use the CLI to configure backend Storwize internal drives.

Important: To ensure optimal performance, all MDisks in a storage pool tier must have the same technology and performance characteristics.

Managed Disks (MDisks)

After deciding which RAID protection for your environment and creating the MDisks/Arrays on the Storwize storage, you should create and present volumes to the SAN Volume Controller/Storwize to be virtualized as external/backend storage. The recommendation is to create volumes in multiples of 4 per MDisk array. For SAS and NL-SAS disks the goal is to stripe the volume on 8 disks; for example in a 64 disk drive array you should create 8 volumes (64/8) to be presented to SAN Volume Controller/Storwize.

Storwize systems can have a mixed disk drive type, such as solid-state drives (SSDs), serial-attached SCSI (SAS), and nearline SAS (Multi-tier storages). Therefore, pay attention when you map the Storwize volume to the SAN Volume Controller storage pools (as MDisks). Assign the same disk drive type (array) to the SAN Volume Controller storage pool with the same characteristics.

In cases of multi-tier 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 backend storage. However, the use of Easy Tier on both storage at the same time is not recommended. Since SAN Volume Controller/Storwize Easy Tier does not monitor Easy Tier on both backend/external storage, SAN Volume Controller/Storwize and the backend storage will independently rebalance the hot areas according to their own heat map, causing a rebalance over a rebalance. Such a situation can cancel the performance benefits of tier reallocation.

Specific recommendations for Easy Tier on external storages can be found in Chapter 7, “IBM Easy Tier function” on page 257.

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.

2.5 Considerations for IBM FlashSystem 900

The main advantage 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.

When you configure the controller for IBM FlashSystem storage systems, you must remember the considerations that are described in this section.

2.5.1 Physical FC port connection and zoning

The physical FC port connections and zoning are described in details on Chapter 1, “SAN topology” on page 1.

2.5.2 Storage capacity and Arrays

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 daughter board are either half-populated with 1.06 TiB (1.2 TB) or fully populated with 2.62 TiB (2.9 TB). The optional daughter 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 and 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.

2.5.3 Extent size

The extent size is not a real performance factor; rather, it is a management factor. If you have some storage pools already, you are advised to create a FlashSystem storage pool with the same extent size as the extent size of existing storage pools. The use of the same extent size to all pools on your SAN Volume Controller/Storwize allows you to use transparent volume migration between pools. If you do not have any other storage pools, you can leave the default extent size, which in V7.x versions of SAN Volume Controller/ Storwize code equals 1024 MB.

2.5.4 Storage pools

If you use FlashSystem 900 as the primary data storage, add all of the MDisks from the controller to a single managed disk group (also known as a storage pool in the SAN Volume Controller GUI). However, if more than one FlashSystem 900 is presented to SAN Volume Controller/Storwize, a preferred practice is to create a single storage pool per controller.

If you use FlashSystem 900 with the SAN Volume Controller/Storwize Easy Tier function you must create multiple volumes for each hybrid storage pool. Create four or more volumes for each hybrid pool on the SAN Volume Controller/Storwize.

For SAN Volume Controller/Storwize to properly recognize FlashSystem MDisks as flash disks, remember to change the MDisk tier on: *Pools* → *External Storage* and click with the right button on the MDisk as shown on Figure 2-7 on page 53.

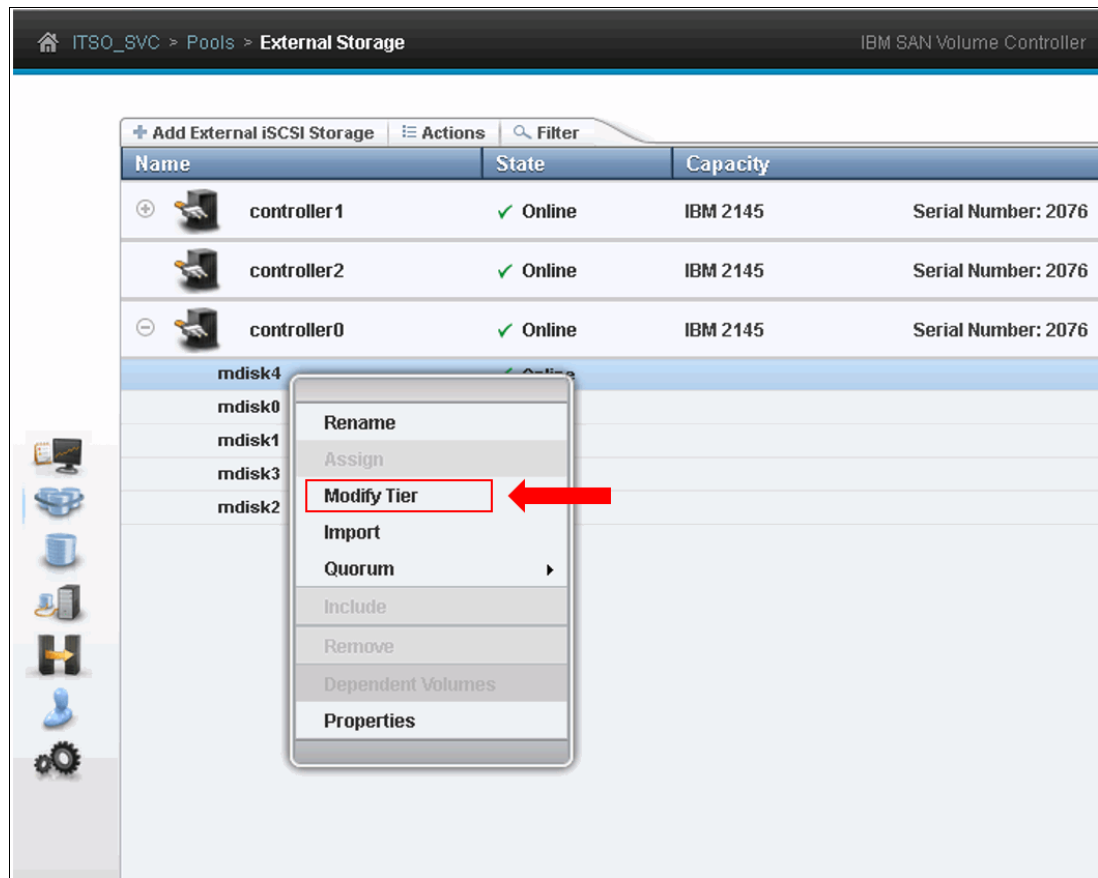


Figure 2-7 Modify MDisk Tier on SAN Volume Controller GUI.

Click with the left button, on the drop down menu option *Modify Tier*. A new window pop-up will appear on the GUI to select the tier for the selected MDisk. For FlashSystem MDisk you must select Tier 0 as shown in Figure 2-8 on page 54.

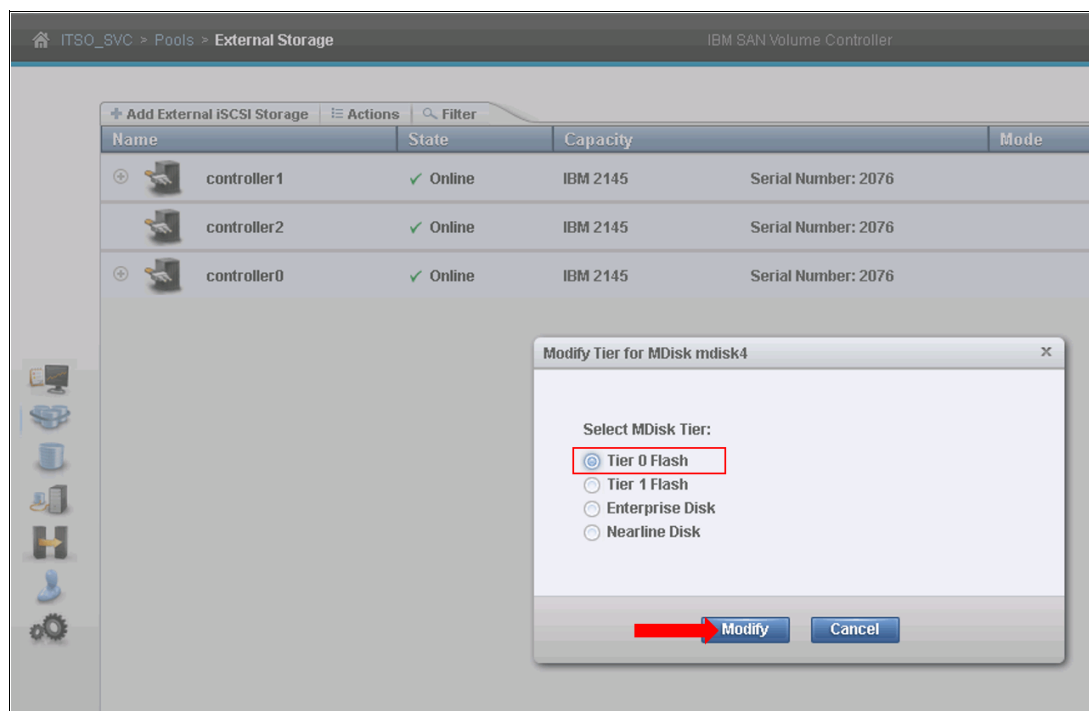


Figure 2-8 Modify MDisk Tier on SAN Volume Controller interface.

Note: For detailed Flash System 900 information, refer to *Implementing IBM FlashSystem 900*, SG24-8271

2.5.5 Volumes

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 MDisk 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 you might consider disabling total cache (both Read and Write) in cases where you are experiencing FlashSystem volume latency issues.

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

- ▶ If Flash Systems volumes are Compressed.
- ▶ If Flash Systems volumes are in a Metro/Global Mirror relationship.
- ▶ If Flash Systems volumes are in a FlashCopy relationship (copy on write).
- ▶ If Flash Systems volumes are in an Easy Tier pool.

Note: Latency can be reduced and maximum IOPs optimized by turning off cache for external FlashSystems volumes. The reason for this is that in the event of a node failure writes destage faster to flash volumes than SVC node to node partner caching. Reads will be redriven through other paths whether it is a cache hit or miss.

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 MDisk but the other copy comes from the spinning-type of MDisk, it is possible to 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 primary copy will prioritize the read from the Flash System volumes as shown in Figure 2-9.

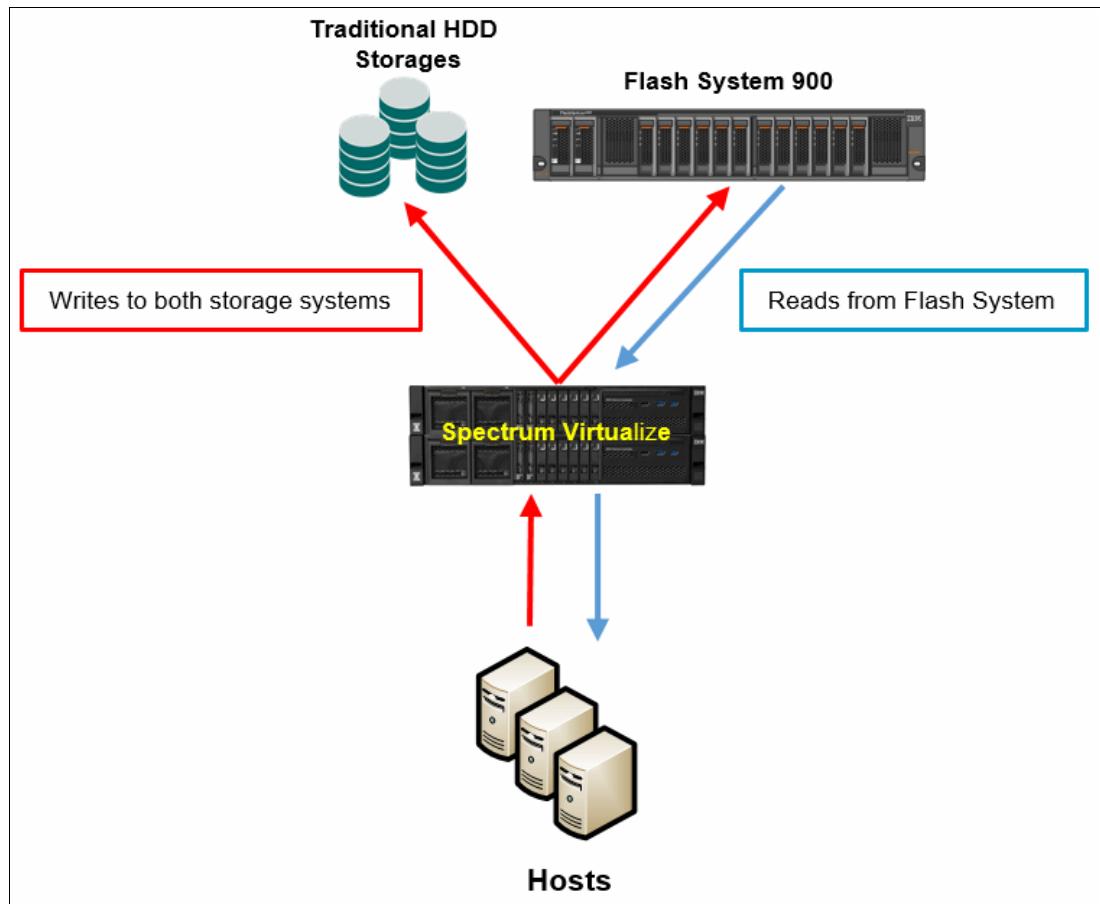


Figure 2-9 Flash System volume as primary copy - read priority.

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 `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 `chvdisk -mirrorwritepriority latency|redundancy volume_name_or_volume_id`.

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

For more information regarding flash system, refer to *Implementing IBM FlashSystem 900*, SG24-8271 available at:

<http://www.redbooks.ibm.com/redpieces/abstracts/sg248271.html>

2.6 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 of storage controllers, when presented to the SAN Volume Controller/Storwize are recognized as a single World Wide Node Name (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 will appear as a different external storage controller.

SAN Volume Controller/Storwize V7.8 supports a maximum of 16 WWNN per storage system so, for this reason, it is recommended to connect up to 16 storage ports to a SAN Volume Controller/Storwize cluster, what will result 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 (System Storage SAN Volume Controller) or 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/Storwize
- ▶ Volume consideration
- ▶ Tiered storage
- ▶ 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 for Storwize systems
- ▶ Considerations when using Encryption for Storwize systems
- ▶ MDisk group (Storage pool) considerations when using Real-time Compression

3.1 Availability considerations for storage pools

Although Spectrum Virtualize provides many advantages through consolidation of storage, you must understand the availability implications that storage subsystem failures can have on availability domains within the Spectrum Virtualize cluster. The Spectrum Virtualize 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 MDisk(s) 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 equate to a gain in application performance.

Remember that Spectrum Virtualize 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 an MDisk 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. For more information, see 3.3.4, “Considerations for the IBM XIV Storage System” on page 66.

If the solution you are going to implement has to provide Business Continuity (BC) and High Availability (HA) capabilities, all the best practices explained later on in this chapter are still valid, but it is strongly suggested to take review specific IBM Redbooks for Spectrum Virtualize BC and HA solutions. Refer to:

- ▶ *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:

- ▶ To minimize potential impact in case of a complete back-end storage subsystem failure, each storage subsystem is suggested to be used with only a single Spectrum Virtualize cluster. This is almost impossible in a data center because a solution like that will increase the TCO, and it is common to share a back-end storage subsystem with different Spectrum Virtualize cluster(s). In this case, specific attention needs to be taken into account about the back-end storage subsystem maximum performance that will be represented by the aggregation of all the Spectrum Virtualize cluster(s) accessing it.
- ▶ It is suggested that each storage pool must contain only MDisk(s) 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 257.
- ▶ It is suggested that each storage pool contains only MDisk(s) from a single Storage Tier (SSD or Flash, Enterprise, 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 257.

- It is suggested to not address more than 250TB per each Storage Pool to minimize the impact in case of a single MDisk failure.

3.2 Selecting storage subsystems

When you are selecting storage subsystems, the decision comes down to the ability of the storage subsystem to be more reliable, resilient, and meet application requirements. When the Spectrum Virtualize does not provide any data redundancy, the availability characteristics of the storage subsystems' controllers have the most impact on the overall availability of the data that is virtualized by the Spectrum Virtualize. This is also true for Storwize family systems unless you use Storwize internal drives. 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 type that is selected.

Performance is also a determining factor, where adding Spectrum Virtualize 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 DS88XX series is a scale-up architecture that delivers “best of breed” 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 Spectrum Virtualize cluster. All storage subsystems possess an inherent failure rate, and therefore, the failure rate of a storage pool becomes the failure rate of the storage subsystem times the number of units.

Other factors can lead you to select one storage subsystem over another. For example, you might use available resources or a requirement for more features and functions, such as the IBM System z® attach capability.

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 an 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.

A starting best practices are:

- Create a Storage Pool for each Storage Subsystem.
- Without any specific workload profile, use a 250TB addressing space per each IOgrp, that scales out to a 1PB Spectrum Virtualize cluster with four IOgrp (8 nodes).
- Create a dedicated storage pool if there is a specific performance application request, such as using FlashSystem dedicated for a particular application. One example could be a

SAP HANA application that will mean you create a dedicated pool for this application. In this scenario keep in mind that if you want to segregate a specific workload at a Spectrum Virtualize level it makes sense only if you will 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/MDisk supplied by the internal storage directly connected to one IOgrp SAS chain only. This will avoid needless IOgrp to IOgrp communication traversing the SAN and consuming Fibre Channel bandwidth.
- ▶ in a Spectrum Virtualize environment, when using 12F or 24F expansion enclosure, create storage pools with IOgrp and Expansion Enclosure affinity. That means you have to use only arrays/MDisk supplied by the internal storage directly connected to one IOgrp SAS chain only. This will avoid needless IOgrp to IOgrp communication traversing the SAN and consuming Fibre Channel bandwidth.
- ▶ Try to limit the number of storage pool to five or less for better cache utilization when using 2145-CG8 nodes and up to seven when using 2145-DH8 or 2145-SV1 nodes. With the introduction of Spectrum Virtualize or Storwize 7.3 version, IBM introduced a new two level cache architecture. With the new cache architecture, potential performance problem caused by an excessive cache partitioning has been mitigated.
- ▶ For storage pool extent size, as a guide, for most clusters, a 1 - 2 PB capacity is sufficient. In general, use 256 MB to address a 1PB space and for larger clusters use 512 MB as the standard extent size to address up to a 2PB space. Alternatively, when you are working with the XIV system or DS88XX family system, use an extent size of 1 GB. Keep in mind that the smaller the extent size the better is the usage of T0 disk like SSD or Flash because Easy Tier will have a better chance to move around a larger number of extents and in most circumstances that will give you better performance.
- ▶ Keep the same extent size for all pools. Volumes can not 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 sub-divide a storage pool (or managed disk group) but maintain a larger number of MDisk in that pool. Child pools are logically similar to storage pools, but allow you to specify one or more sub divided child pools. Quotas and warnings can be set independently per child pool.

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.

3.3.1 Selecting the number of arrays per storage pool

The capability to stripe across disk arrays is the most important performance advantage of the Spectrum Virtualize. 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 required in terms of performance has to 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 will increase the failure domain as described in 3.1, “Availability considerations for storage pools” on page 58. On the other hand, reducing the arrays number using bigger disk (like 2TB to 8TB NL_SAS disk or 1.2TB to 1.8TB disk) could affect performance because the bigger the disk the bigger the I/O density. It is important to find the trade off 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, but 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, meaning that when workload is spread evenly across all storage pools, no difference in performance occurs.

RAID 5 compared to RAID 10

In general, RAID 10 arrays are capable of higher throughput for random write workloads than RAID 5 because RAID 10 requires only two I/Os per logical write compared to four I/Os per logical write for RAID 5. For random reads and sequential workloads, often no benefit is gained. With certain workloads, such as sequential writes, RAID 5 often shows a performance advantage.

Selecting RAID 10 for its performance advantage comes at a cost in usable capacity, and in most cases, RAID 5 is the best overall choice but, as always, it depends. Today, storage subsystems typically have a larger cache size than in the past, and with Spectrum Virtualize in front of a storage subsystem, the benefits of RAID10, in terms of performance, means there is negligible performance benefit between RAID5 and RAID10.

If you are considering RAID 10, use Disk Magic to determine the difference in I/O service times between RAID 5 and RAID 10. If the service times are similar, the lower-cost solution might make the most sense. If RAID 10 shows a service time advantage over RAID 5, the importance of that advantage must be weighed against any additional cost.

RAID 10 offers, in addition to better performance, better resilience as well. If a RAID 5 solution will supply the required performance, consider Spectrum Virtualize Enhanced Stretched Cluster (ESC) solution or Storwize V7000 and SVC ESC HyperSwap® (HS) Solution that will add more benefits and resilience than a simple RAID 10 back-end storage subsystem.

RAID 6

RAID 6 consists of block-level striping with double distributed parity. Double parity provides fault tolerance up to two failed drives. This makes larger RAID groups more practical, especially for high-availability systems, as large-capacity drives take longer to restore. RAID 6 requires a minimum of four disks. As with RAID 5, a single drive failure results in reduced performance of the entire array until the failed drive has been replaced. With a RAID 6 array, using drives from multiple sources and manufacturers, it is possible to mitigate most of the problems associated with RAID 5. The larger the drive capacities and the larger the array size, the more important it becomes to choose RAID 6 instead of RAID 5.

Figure 3-1 on page 62 shows the estimated traditional RAID rebuild time.

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

Figure 3-1 Estimated traditional RAID rebuild time

Distributed RAID (DRAID)

Distributed RAID was launched in 7.6.0 and allows 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 with each stripe having a data stripe on seven of the drives and a parity strip on the eighth. In distributed RAID5 you specify the stripe width and the number of drives separately so you can still have seven data stripes protected by a parity stripe but those eight drives are selected from 64. Additionally we have added distributed sparing, this is the concept that instead of having a spare set on the side that isn't being used, each drive in the array gives up some of its capacity to make a spare.

Rebuild Performance

The main reason for distributed RAID is to improve rebuild performance. When a drive fails, the data from that drive has to be rebuilt from the surviving drives and written to a spare. By having a larger set of drives in the array those rebuild reads are coming from more drives and distributed sparing means 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 4TB nearline drive. The rebuild time could be up to 24 hours. With RAID5 a second drive failure during a rebuild means the array goes offline and if the drive can't be resuscitated the entire array needs to be restored from backup, RAID6 copes with a second concurrent failure but a third is going to be terminal. Some products offer RAID7 to cope with three failures, but it is all just delaying the inevitable as the drives get bigger and rebuilds take longer.

Number of Drives per DRAID Array

One of the key decisions anyone has to make is how many drives to put into an array. As you increase the number of drives the rebuild time shortens but it isn't linear, and it doesn't go on forever, as you likely start to hit other limits in the system. Our testing shows that about 64 is the sweet spot for spinning disks, the GUI will recommend between 40 and 80 assuming that you have at least 40 of the drive class you want to use. Typically we suggest going with what the GUI recommends.

One of our recommendations is that you have heterogeneous pools, having different drive classes within the same tier is not a great idea. The current GUI implementation can be a little over zealous, and so can start disallowing some things that are fairly reasonable, such as someone trying to add a second array with drives that are slightly larger than the ones that make up the current array in the pool. For now you will have to use the CLI to work around this behavior.

There are times when you will want to override the GUI recommendations, the main reason is likely to boil down to the current limitation that arrays can't be expanded by adding new drives. So you might want to think about what you will do when the pool runs out of space, if it's a case of buying a new expansion with 24 drives then you might want to use 24 drives per array, or you might want to look at buying two expansions and adding 48 drives when it gets close to capacity.

Array Performance

Typically we follow the 'rule of four' as a commandment that cannot be disobeyed. The rule of four stems back to the fact that Storwize V7000 Gen1 has four cores for processing IO, each volume gets assigned to a core and each array gets assigned to a core, so a single volume using a single array could be only using one core for the majority of the IO processing. If you are in that situation and you have a performance critical system, and you have a workload that may be limited by the processing power, then you might want to use four arrays. However, with distributed RAID came improvements that allow some of the IO processing for an array to be done outside of the assigned core so that the problem is reduced, and there are also performance advantages of having a single distributed raid array. If your system happens to fall nicely into having four arrays then that's great, but going out of your way to contort it to fit the rule is just going to give you more overheads.

The reason why one large array can improve performance is due to the Spectrum Virtualize code allocating extents from a pool to a volume. By default those extents are now 1GB, that means if reading or writing to a 1GB extent you are only using the drives in one MDisk. For a random workload you would expect multiple extents to be active across all of the MDisks so all the drives are in use, but for a sequential workload you tend to hit one extent rather hard then move onto the next. The cache does improve things but it's still very easy to get a situation where some drives are sat idle. A distributed array uses every drive in the array for every extent so getting all drives active is much easier.

Thanks to the performance improvements introduced in version 7.7.1 that have removed restrictions of SSDs usage on DRAID, even with 6+ SSD there are benefits. The main advantage (in addition to the rebuild time of course) is the active use of the spare disk performance. Since the spare disks are within the DRAID, their performance contributes to the current use. This is especially important when using SSDs. With 9 SSD, you have 7 + P + S within the matrix, although the ability to spare is not used, its performance makes a contribution.

For HDD arrays of any DRAID type the default recommendation is RAID6 and there is various information out there that shows DRAID-6 performs better than RAID5 in traditional form.

There are always discussions and questions regarding Storwize DRAID performance, and the current design that still assigns a single MDisk to a core/thread internally. While this may seem less than ideal today, lets do some maths and work out if its really an issue for HDD.

Taking the worst case, i.e. highest performing drives, 15K, when we run flat out (hitting drive limits) for miss workloads today, we only see about 80% overall CPU utilisation on the system as measured in our IBM Lab.

So that is 504 drives, giving 80% utilisation.

We have 14 active cores across two nodes doing the work (the other two cores are dedicated to cache mirroring over PCIe).

Therefore if we push to 100% utilisation, then that's 45 drives per core. $(504 / 14) / 0.8$.

But we have two nodes servicing the array so that is 90 drives.

Therefore, with a 15K drive, you can do up to 90 drives in a single DRAID before CPU becomes a question - and that's best case, running the array flat out (around 400 iops per drive - which needs to be short stroked to achieve!).

If we take that scenario to NL-SAS, where most DRAID will be used, that means the upper limit of 120 drives per DRAID will be hit before the core becomes a bottleneck.

Rebuild Areas

One final topic is about the number of rebuild areas, and a rebuild area is the equivalent capacity to a single drive. So the more rebuild areas 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. Once a drive has been replaced, the data then gets copied back from all the spare spaces to the replaced drive, so this is another case of writing to a single drive which can take a few days on the really large Nearlines. So the copy back time needs to be taken into account, replacing a drive doesn't immediately give you back the redundancy. Our advice would be to go with the default suggested by the GUI, then add an extra one if the data is critical (but don't use the fact you are using extra rebuild areas to drop down from RAID6 to RAID5) or if you want to have some leeway to batch up replacing failed drives.

3.3.2 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, which defeats the subsystem's ability to perform "full stride writes" for RAID 5 arrays. Also, I/O queues for multiple LUNs directed at the same array can overdrive the array.

Higher-end storage controllers, such as the DS8000 series, make this situation much less of an issue by using large cache sizes. In addition, on higher 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 the version 7.x, the maximum supported MDisk size equals 1 PB, so the maximum LUN size on the storage controller side no longer is 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 LUNs 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 on each pool LUNs with the same performance characteristics. Varying the performance level will mean you are not consistent with the type of data 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 Spectrum Virtualize does not provide any RAID capabilities 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 95.

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 best practice to optimize the performance, availability and cost of the infrastructure.

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

- ▶ It is suggested that each storage pool contains a number of MDisks at least $\geq$ the number of SVC nodes CPU:
 - 6 MDisk or 6x for 2145-CG8 node
 - 8 MDisk or 8x for 2145-DH8 and 2145-SV1 node
- ▶ Max size of MDisks $\leq$ 2TB or $>$ 2TB where the device type allows for better queue depth. Remember that the fewer the backend MDisks the better the queue depth. All Storwize and Spectrum Virtualize based products have a per node queue depth of 10,000, so a single control enclosure can have 20,000. There is no per LUN queue depth limit, so all 10,000 could be on one LUN.
- ▶ However, in performance testing about 1,500 per node is all that is needed with a decent back-end solution and low latency to maximize out the system.
- ▶ Spread all MDisks equally across all Spectrum Virtualize node ports for better load balancing.

3.3.3 Considerations for Storwize family systems

For the Storwize family, you can have the following possible scenarios:

- ▶ Storwize as back-end storage system to Spectrum Virtualize or another Storwize
- ▶ Storwize as front-end storage systems to hosts

In a case where Storwize is a back-end controller for Spectrum Virtualize or another Storwize, see Chapter 2, “Back-end storage” on page 37.

If you have Storwize as a front-end storage system to your hosts and you want to use Storwize internal drives, you must consider the raid type and width and stripe size. When you are configuring internal drives, each raid becomes an MDisk with the type of *array*. You can use the GUI for internal storage configuration, which uses the default settings that are considered as the best for general purpose.

However, if you know the I/O characteristic of your applications, you can use the CLI to tune it. For example, if you have an application that uses GPFS™ file system, you might want to create some arrays with the stripe size of 1 MB because GPFS always uses 1 MB I/O to disks. Therefore, it is beneficial to create, for example, RAID5 arrays 4+1 with stripe size of 256 KB, RAID5 with 8+1 with stripe size of 128 KB, RAID10 4+4 with stripe size of 256 KB, or RAID10 8+8 with stripe size of 128 KB, and so on.

Best practice: For general-purpose storage pools with various I/O applications, use the storage configuration wizard in the GUI. For specific applications with known I/O patterns, use CLI to create arrays that suits your needs.

3.3.4 Considerations for the IBM XIV Storage System

The XIV system currently supports the following configurations:

- ▶ 27 - 79 TB of usable capacity when you use 1 TB drives
- ▶ 55 - 161 TB when you use 2 TB disks
- ▶ 84.1 - 243 TB when you are using 3 TB disks
- ▶ Up to 325 TB usable capacity that uses 4 TB disks

The minimum volume size is 17 GB. Although you can create smaller LUNs, define LUNs on 17 GB boundaries to maximize the physical space available.

Support for MDisks larger than 2 TB: MDisks larger than 2 TB on the XIV system are supported.

Spectrum Virtualize has a maximum of 511 LUNs that can be presented from the XIV system. Spectrum Virtualize does not currently support dynamically expanding the size of the MDisk.

Because the XIV configuration grows 6 - 15 modules, usage of the Spectrum Virtualize rebalancing script to restripe volume extents to include new MDisks is no longer required and Easy Tier, Intra-Tier capability, will take care of this. With Easy Tier activated on this pool when a new MDisk will be added, the space allocated on each MDisk will be rebalanced based on the workload profile of each MDisk in order to get the best performance.

The size of the volume created on the XIV acting as back-end storage system for a Spectrum Virtualize can vary, depending on some parameters.

The XIV can handle a queue depth of 1400 per Fibre Channel host port and a queue depth of 256 per mapped volume per host port:target port:volume tuple. However, the Spectrum Virtualize or Storwize V7000 sets the following internal limits:

- ▶ 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 the 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 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 Spectrum Virtualize or Storwize V7000 cluster (2, 4, 6, or 8)

M Number of volumes presented by the XIV to the Spectrum Virtualize or Storwize V7000 cluster (detected as MDisks)

C 1000 (the maximum SCSI queue depth that an Spectrum Virtualize or Storwize V7000 uses for each XIV host port) Spectrum Virtualize

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

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

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

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

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

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

Starting with version 6.4, Spectrum Virtualize and Storwize V7000 clusters support MDisks greater than 2 TB from the XIV system. If you use earlier versions of the Spectrum Virtualize code, smaller volume sizes for 2 TB, 3 TB, and 4 TB drives are necessary.

This leads to the suggested volume sizes and quantities for Spectrum Virtualize or a Storwize V7000 system on the XIV with different drive capacities, as shown in Figure 3-2 on page 67.

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 quantity	Ratio of volumes to XIV host ports
6	4	1600	3201	4801	6401	17	4.3
9	8	1600	3201	4801	6401	27	3.4
10	8	1600	3201	4801	6401	31	3.9
11	10	1600	3201	4801	6401	34	3.4
12	10	1600	3201	4801	6401	39	3.9
13	12	1600	3201	4801	6401	41	3.4
14	12	1600	3201	4801	6401	46	3.8
15	12	1600	3201	4801	6401	50	4.2

Figure 3-2 Suggested volume size and quantity

The best use of the Spectrum Virtualize virtualization solution with XIV systems can be achieved by running LUN allocation with the following basic parameters:

- ▶ Allocate all LUNs (MDisks) to one storage pool. If multiple XIV systems are being managed by Spectrum Virtualize, each physical XIV system should have a separate storage pool. This design provides a good queue depth on the Spectrum Virtualize to drive XIV adequately.
- ▶ Use 1 GB or larger extent sizes because this large extent size ensures that data is striped across all XIV system drives.

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

- ▶ *XIV Gen3 with Spectrum Virtualize and Storwize V7000*, REDP-5063, which is available at this website:
<https://ibm.biz/Bdr7UQ>
- ▶ “Can you use Spectrum Virtualize with XIV as storage?”, which is available at this website:
<https://ibm.biz/Bdr7U3>

3.3.5 IBM FlashSystem A9000/A9000R considerations

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

- ▶ Minimum 7.4 version 7.4.0.10
- ▶ Minimum 7.5 version 7.5.0.8
- ▶ Minimum 7.6 version 7.6.1.4
- ▶ Version 7.7 and higher.

You can also check the Spectrum Virtualize Support matrices for any further updates.

What volume size and how many volumes?

The best case is that we have run the data reduction estimation tool (available in Fix Central) and know the total data reduction ratio we expect to get. The data reduction estimation tool can be downloaded at:

<https://ibm.biz/Bdr7Uk>

The next best case is that we know the compression ratio of the data, and the worst case is that we don't have any of that information.

So, when confronted with this worst case scenario, we will use a data reduction ratio of 2.5:1.

But in any case, calculate the effective capacity of the FlashSystem A9000/A9000R by multiplying the usable capacity times the data reduction ratio, and then divide by the number of paths X 2, to get the volume size.

One point to remember, the SVC and A9000/A9000R work very well together and data reduction information from the underlying storage has to be collected directly from FlashSystem A9000/A9000R.

It is extremely important not to run out of hard capacity on the underlying system, as that will take the storage pool offline. Close monitoring of the FlashSystem A9000/A9000R is very important. Take care to enable all the different levels of alerts for the pool thresholds as shown in Figure 3-3, and turn off alerts for snaps as the snap space needs to be zeroed out anyway.

Pool Properties Units: GB

Pool Thresholds (%)

☐ Derived From System
 ☒ Custom Threshold

Volume Used	Snapshot Used
Informational ☒ 65	Informational ☐ N/A
Warning ☒ 70	Warning ☐ N/A
Minor ☒ 80	Minor ☐ N/A
Major ☒ 90	Major ☐ N/A
Critical ☒ 95	Critical ☐ N/A

Figure 3-3 Monitoring alert setup on A9000 and A900R

If you start to run out of space, you can use the migration characteristics of Spectrum Virtualize to move data to another storage system.

Let's look at a couple of quick 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 we will be virtualizing and know we have a data reduction ratio of 4.2:1. 4.2×57 gives you 239.4 TB. Divide this by 12 (6 paths $\times$ 2), and you will get 19.9TB per volume.

- For our next example, we have a 5 grid element FlashSystem A9000R, using 29TB Flash enclosures, for a total usable capacity of 145TB.

We are using 10 paths and have not run any of the estimation tools on the data, but we know the host is not compressing the data. 2.5×145 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 Spectrum Virtualize.

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 1PB 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 a lot of volumes to manage.

Sizing a back end storage subsystem with compression and deduplication capability in a Spectrum Virtualize environment

Usage of the correct compression and or deduplication ratio is key to get a stable environment. If you are not sure about the real compression and or deduplication ratio contact IBM support to get more information as to how to approach this.

Usage of the wrong compression and/or deduplication ratio can cause difficulties for several reasons:

- ▶ The amount of data in a storage system is not static, it varies. Data deletion and replacement is normal behavior. Those operations can create many gaps within the data structure. Re-usage of those gaps can be difficult for the storage system. The capacity for garbage collection is the amount of obsolete data on the physical disk which will be given back as usable data by the garbage collection process. Usually this process runs after data changes and in an environment with a high data change rate this may cause a delay in releasing physical space.

- For example:

Assumption 1: No capacity calculated for garbage collection:

Physical capacity: 20TB

Calculated capacity: $20\text{TB} \times 5.2 = 104\text{TB}$

Assumption 2: 5% capacity required for garbage collection:

Calculated capacity: $104\text{TB} \times 95\% = 98,8\text{TB}$

If we had assumed a compression and/or deduplication ratio of ~5:1, that is a little optimistic in this case and because of the nature of data, we have seen different ratios, such as 3:1.

There are multiple consequences:

- ▶ Using the wrong ratio for capacity assignment to Spectrum Virtualize will cause an out of space situation if the Spectrum Virtualize Managed Disk doesn't provide enough capacity and Spectrum Virtualize will disable the whole pool. All Volumes related to the pool will go offline, so multiple Spectrum Virtualize pools may be affected. As a workaround additional spare capacity should be planned and allocated (filled with dummy data with a small deduplication ratio) to avoid the deadlock situation. The spare capacity volume can be removed on the storage subsystem in the deadlock situation and the storage subsystem is operational again.

- For example:

Assumption 1: Sizing with ~5:1 rate

Assumption 2: Real rate is 3:1

Physical Capacity: 20TB

Calculated capacity: $20\text{TB} \times 5 = 104\text{TB}$

Volume assigned from compressed and/or deduplicated storage subsystem to SVC:
104TB

Real usable capacity: $20\text{TB} \times 3 = 60\text{TB}$

The physical capacity is fully allocated by 60 TB host data, instead of the calculated 104 TB data. If this happens across the whole storage subsystem, the storage subsystem cannot provide any capacity anymore, and all volumes used as Spectrum Virtualize Managed Disks will go offline, and all related pools will go offline.

Spectrum Virtualize data removal or volume deletion are not reflected on the storage subsystem, because the data is Spectrum Virtualize 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 zeroes and is still in use by the Spectrum Virtualize cluster. So 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 zeroes.

Final thoughts

Definition of the compression and/or deduplication ratio is the key to get a stable system. The storage subsystem volume size, provided to Spectrum Virtualize, depends on the rate. So the assumption must be made before final configuration.

Some additional space is required for garbage collection. One additional volume should be prepared to get the storage subsystem back to operation if there is no capacity available any longer in the system. Some spare capacity should be included in the calculation for several reasons:

- ▶ Reduction of data space saving ratio because of workload changes (for example, you start with documents, but later store videos and music)
- ▶ Keep some free space for flexibility (as already known from other storage systems)
- ▶ So the capacity sizing depends on:
 - Real compression and/or deduplication rate (IBM clients can use the Comprestimator or Data Reduction Estimation Tool available from IBM Support Fix Central):
<https://www.ibm.com/support/fixcentral/>
 - Capacity for garbage collection
 - Capacity for a single, removable volume in case of an out of space scenario
 - Buffer to keep some spare capacity

This kind of calculation is always error prone. There are two possible scenarios:

1. The calculated space saving rate is too high (i.e. 5:1):
 - a. You must expect an outage of the whole system at a certain utilization level.
2. The calculated space saving rate is too low (i.e. 1:1):
 - a. The system is running stable, not all capacity is used and you can add the free capacity later step by step to the Spectrum Virtualize and make use of it.

In any case we recommend monitoring the available physical capacity in the storage subsystem.

- ▶ A final example with a compression and or deduplication ratio of 3:1 (real values may differ):
 - Logical capacity required: 100TB
 - Physical capacity required: $100\text{TB} / 3 = 33,3 \text{ TB}$
 - Additional capacity for garbage collection i.e. 5% = 1.6 TB
 - Additional capacity for out of space scenario i.e. 1TB -> 1.0TB

- Buffer to keep some spare capacity, i.e. 20% = 6.6 TB
- **Total physical capacity:** 42.5TB

To provide a logical capacity of 100 TB at a compression and/or deduplication ratio of 3:1, a capacity of 42.5TB is required, which is equivalent to a total space saving ratio of approximately 2.35:1.

If you are not sure about the sizing, it is suggested to involve IBM Support for your country.

3.3.6 Considerations for the DS88XX family

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.

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

For Spectrum Virtualize attachments, some clients formatted the DS8000 arrays in 1:1 assignments between arrays and extent pools, which disabled any DS8000 storage pool striping or auto-rebalancing activity. Then, they located one or two volumes (MDisks) in each extent pool exclusively on one rank only, and put all of those volumes into one Spectrum Virtualize Storage Pool. Spectrum Virtualize controlled striping across all these volumes and balanced the load across the RAID ranks by that method. No more than two MDisks (DS8000 volumes) per rank are needed with this approach. So, the rank size determines the MDisk 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 MDisks 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 SAN Volume Controller striping occurs because each MDisk in a storage pool is the same size then.

The Spectrum Virtualize extent size is the stripe size that is used to stripe across all these single-rank MDisks.

This approach delivered good performance and has its justifications. However, it also has a few drawbacks.

There can be natural skew, for example, a small file of a few hundred KiB that is heavily accessed. Even with a smaller Spectrum Virtualize 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 MDisks on one rank, and not as many Spectrum Virtualize storage pools, Spectrum Virtualize might start striping across many entities that are effectively in the same rank, depending on the storage pool layout and 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.

You have two options:

- Go for huge multitier hybrid pools, having just one pair of DS8000 pools where the DS8000 internal Easy Tier logic is also doing the cross-tier internal optimization.
- Create in the DS8000 storage system as many extent pool pairs as you have tiers in the DS8000 storage system, report each such DS8000 pool separately to the Spectrum Virtualize, and the Spectrum Virtualize-internal Easy Tier logic makes the cross-tier optimization.

In the latter case, the DS8000 internal Easy Tier logic can still do intra-tier auto-rebalancing.

You need only one MDisk volume size with this multi-rank approach because plenty of space is available in each large DS8000 extent pool.

Often, clients choose 2 TiB (2048 GiB) MDisks for this approach. Create many 2-TiB volumes in each extent pool until the DS8000 extent pool is full, and provide these MDisks to the Spectrum Virtualize to build the storage pools.

Two extent pools are needed at least so that each DS8000 processor complex (even/odd) is loaded about equal, or it might take a larger number of pools.

If you use DS8000 Easy Tier, even only for intra-tier auto-rebalancing, do not use 100% of your extent pools. You must leave some small space of a few extents per rank free for Easy Tier so that it can work.

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

3.4 Quorum disk considerations

When back-end storage is initially added to a Spectrum Virtualize 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 Spectrum Virtualize 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 another external storage system.

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 Spectrum Virtualize cluster quorum MDisk 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 MDisk 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 one-half portion 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 Spectrum Virtualize 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.

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 this website:

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

Attention: Running a Spectrum Virtualize 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 (which 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 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 MDisk are acting as quorum in your environment, we suggest you create three 1GB LUN, one on each back-end storage subsystem if possible, and put these MDisk in their own dedicated storage pool. For Storwize V7000, quorum disk are HDD drives by default so the SVC best practice does not apply. However, if your Storwize V7000 is virtualizing external storage then the same best practice for the SVC does apply.

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

When implementing a Spectrum Virtualize ESC or HS solution, some clients like to have a preferred winning site in case of a split brain scenario and using quorum disk this is possible. Clients need to implement the active quorum disk in the site that is considered as preferred this storage subsystem need to be defined as to be in Site 3 anyway, connecting to the remote site using SAN extension connectivity, the same connectivity used to extend the

Public and Private SAN. In this way the winning site will be always the one with two quorum implemented.

IP Quorum

With respect to quorum disk, IP quorum is a feature 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 MDisk or drives for quorum resolution. As stated earlier, normally three MDisk 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. In the case of a quorum loss, that is 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 that is, Fibre Channel (FC) or serial-attached SCSI (SAS) attached storage. But in case of an Enhanced Stretched Cluster or HyperSwap environment, the need for accessibility to the quorum device in the case of a site failure necessitates the presence of a third independent domain for quorum resolution. In versions before 7.6.0, the requirement of a third site is necessary over FC connectivity to the third site with an associated storage controller at the third site. Maintaining this third site and storage controller over FC makes the system costly in case of 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, the alternative is to 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 host application over Ethernet connection can reduce the implementation and maintenance cost.

IP Quorum requirements

Connectivity from server(s) to the service IP addresses of all nodes.

- ▶ Only the first Ethernet port can be used. There is no VLAN support yet.
- ▶ Port 1260 (via SSL/TLS) for inbound connections from the app to the service IP address of each node.
- ▶ Maximum round-trip time = 80 ms, minimum bandwidth of 2 MB/s.
- ▶ 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 if 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 to recreate the Java quorum application package.

So, how many IP Quorum do we have to set? Spectrum Virtualize and Storwize V7000 can support up to five IP Quorum at the same time — but only one is active.

The active IP Quorum Application will be either the first one started and detected by the Spectrum Virtualize cluster, or it will be 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 may be able to achieve this.

It is suggested to 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 Quorum one per each site.

Note that even with three IP Quorum, one on each site, currently it is impossible to be certain who will be the winning site in case of a split brain scenario, because FC connectivity interruption between your Spectrum Virtualize Nodes 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 App and a quorum disk at site 1 and site 2, and you lose all IP quorum App 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 `lshw`, and normally the node that was used when the system was first set up) will be used to resolve the tie-break.

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

Note 2: This id can change as 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 refer to:

http://www.ibm.com/support/knowledgecenter/en/STPVGU_7.8.0/com.ibm.storage.svc.console.780.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 at:

<https://ibm.biz/BdrWGT>

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

The script will restart the IP Quorum Application automatically in case of a failure

3.5 Volume Consideration

Volumes in Spectrum Virtualize or Storwize can be created as *striped* or *sequential*.

The general rule now is always to create as striped, but as usual it depends. The main target is to take the 60 MDisk queue depth into account, which typically says for HDD MDisk, to aim for 8 spindles per MDisk, which when coupled with the new cache algorithms, works much better (number of drives/8 spindles).

So here, the rule would be that for the number of drives you have on the back-end in a pool, for example 64 drives, then $64/8 = 8$ volumes created on the back-end and presented to the SVC as 8 MDisk. This means when we get the 60 MDisk queue, we get roughly a queue depth of 8 per drive, which keeps a spinning disk well utilized. It also gives nice concurrency across the ports of the back-end controller.

3.6 Tiered storage

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

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

- They have the same hardware characteristics; for example, the same RAID type, RAID array size, disk type, and disk revolutions per minute (RPMs). This is mostly true about same disk with same RPMs. Today we can have *Enterprise* disks with different size but same RPMs, for example 10K RPM 600GB SAS disk as well as 900GB, 1.2TB and 1.8TB.

The suggestion here is to try to keep size and speed consistency in a *single-tiered storage pool*. If this is not possible because of storage pool space, upgrades at different time, and disks with the same size would no be longer available, it is suggested to use disks with sizes closer to the original one. For example, if the pool was configured with 10K RPMs 900GB SAS disk, a mix with 10K RPM 1.2TB SAS disk could be accepted. This would not have a serious side effect on the performance because Easy Tier has introduced Intra-Tier balance that will balance the workload on different MDisk (in this case with different drive size) basing on I/O density and response time.

- The MDisk that will be configured have to have the same size as long as possible. If this requirement is not feasible, Spectrum Virtualize Easy Tier with Intra-Tier balance will balance the workload on different MDisk (in this case with different Drive size) basing on I/O density and response time.

In a *multitiered storage pool*, you have a mix of MDisk 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 has changed as following:

- 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 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 version 7.8 the naming has changed as follows:

- tier3_nearline

Figure 3-4 shows changes to Tech Types introduced from V7.8.

Old Name	New Name
<code>sas_ssd</code>	<code>tier0_flash</code>
	<code>tier1_flash</code>
<code>sas_hdd</code>	<code>tier2_hdd</code>
<code>sas_nearline_hdd</code>	<code>tier3_nearline</code>

Figure 3-4 Tech Types

A multitiered storage pool, therefore, contains MDisk 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 Spectrum Virtualize Easy Tier function. For more information about IBM System Storage Easy Tier, see Chapter 7, “IBM Easy Tier function” on page 257.

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

Therefore, a storage tier attribute is assigned to each MDisk basing on the Storage Controller where it belong because MDisk with same HDD size and RPMs could be supplied by different Storage Controller like Storwize V7000 or DS88XX family with different hardware characteristics and, of course, with different performance.

Before Spectrum Virtualize version 7.8 there was three types of storage tier: **ssd**, **enterprise** and **nearline**.

Starting with 7.8 we have three tiers encompassing four Tech Types as shown in Figure 3-5.

Tier	Tech Types
<code>ET_Tier1</code>	<code>tier0_flash</code>
<code>ET_Tier2</code>	<code>tier1_flash, tier2_hdd</code>
<code>ET_Tier3</code>	<code>tier3_nearline</code>

Figure 3-5 Tier and Tech Types

For Storwize, 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 Spectrum Virtualize or Storwize, a specific Easy Tier **easytierload** profile is assigned and 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 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 MDisk that are presented from external storage systems accordingly with the owning storage controller using **chmdisk** command.

When multiple storage tiers pool 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 MDisk.

Effectively we can say we have four tiers within a 3-Tier Mapping; tier1_ssd and tier2_hdd will “share” with SSD getting the default data and a greater portion of I/O in balancing.

Spectrum Virtualize 7.8 brings XML assignment to Tier1_SSD so the balancing is effective.

Figure 3-6 on page 80 shows where data lands based on Tier and Tech Types combination.

User (VG) Tiers	EasyTier Tier (by configuration)													
	T0	T0+T1	T0+T1 +T2	T0+T1 +T2+T 3	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 NearLi ne)				3		3	2			3	3		3	3

Figure 3-6 Tier and Tech Types combination

3.7 Adding MDisk to existing storage pools

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

3.7.1 Checking access to new MDisk

Be careful when you add MDisk 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 Spectrum Virtualize 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 Spectrum Virtualize cluster suffers a software error during the MDisk test.

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

3.7.2 Persistent reserve

A common condition where MDisk can be configured by Spectrum Virtualize (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 LUNs. Then, 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 **lquerypr** (which is included with SDD and SDDPCM) or the Microsoft Windows SDD Persistent Reserve Tool.

3.7.3 Renaming MDisk

After you discover MDisk, rename them from their Spectrum Virtualize-assigned name. To help during problem isolation and avoid confusion that can lead to an administration error, use a naming convention for MDisk that associates the MDisk with the controller and array.

When multiple tiers of storage are on the same Spectrum Virtualize 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.

Best practice: Use a naming convention for MDisk that associates the MDisk with its corresponding controller and array within the controller; for example, DS8K_R5_12345.

3.8 Rebalancing extents across a storage pool

From 7.3 onwards, Spectrum Virtualize Easy Tier has introduced Intra-Tier balancing that will balance the extent and workload on different MDisk basing it on I/O density and response time.

Before this feature, adding MDisk to existing storage pools before 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 MDisk to storage pools, rebalancing extents across all available MDisk was by using the CLI or alternatively using a Perl script. This is now automatically taken care of.

3.9 Removing MDisk from existing storage pools

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

Sufficient space: The removal occurs only if sufficient space is available to migrate the volume data to other extents on other MDisk 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 destroyed. This means the array on which this MDisk was built is also destroyed and all drives that were included in this array convert 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, and so on.

3.9.1 Migrating extents from the MDisk to be deleted

If an MDisk contains volume extents, you must move these extents to the remaining MDisk in the storage pool. Example 3-2 shows how to list the volumes that have extents on a MDisk by using the CLI.

Example 3-2 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 Spectrum Virtualize to automatically move all used extents on the MDisk to the remaining MDisk in the storage pool.

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

3.9.2 Verifying the identity of an MDisk before removal

MDisk must appear to the Spectrum Virtualize cluster as unmanaged before their controller LUN mapping is removed. Unmapping LUNs from the Spectrum Virtualize 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 the Spectrum Virtualize 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 Spectrum Virtualize 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 3.9.3, “Correlating the back-end volume with the MDisk” on page 83.

3.9.3 Correlating the back-end volume with the MDisk

The correct correlation between the back-end volume (LUN) with the Spectrum Virtualize 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 Spectrum Virtualize cluster, the LUN ID must be combined with the worldwide node name (WWNN) attribute to uniquely identify LUNs within the Spectrum Virtualize 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; for example, from 5005076305ffc74c to 6005076305ffc74c.

When detected as Spectrum Virtualize 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 3-3.

Example 3-3 DS8000 UID example

[illegible]

In Example 3-3 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 3-7 shows the Volume Properties dialog box that opens.

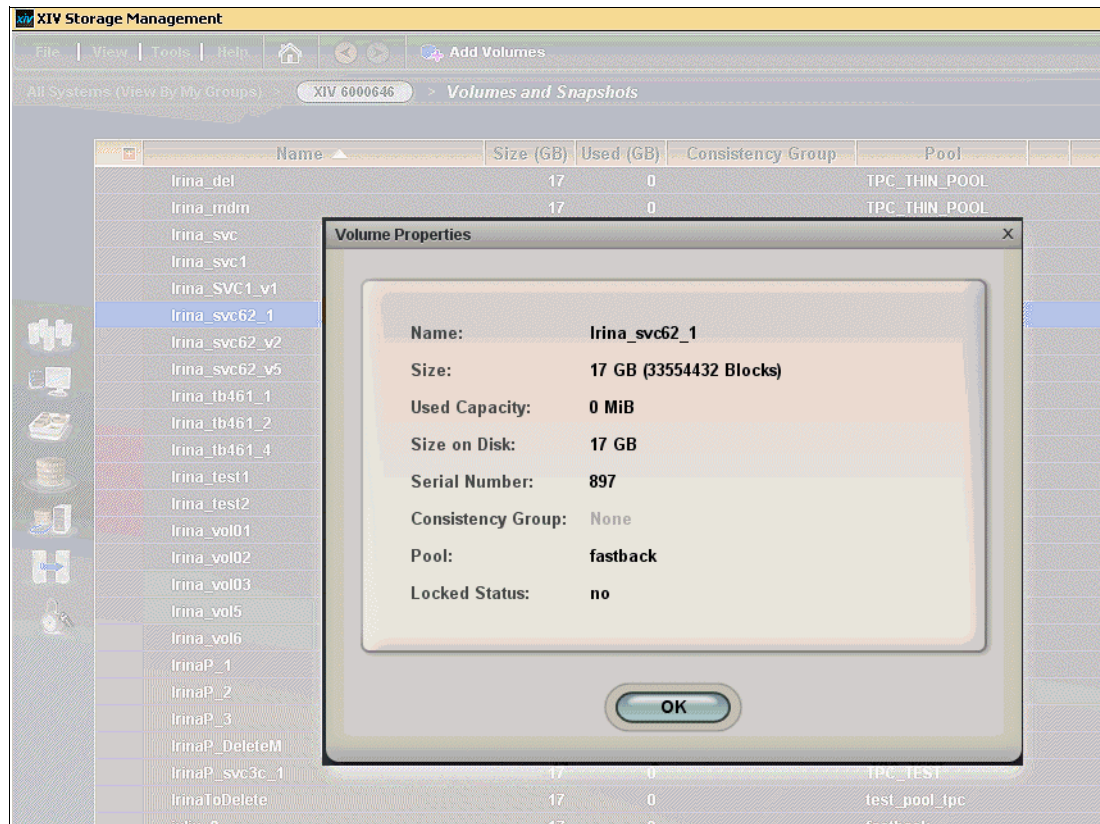


Figure 3-7 XIV Volume Properties dialog box

To identify your LUN, in the Volumes by Hosts view, expand your Spectrum Virtualize host group and then review the LUN column, as shown in Figure 3-8 on page 85.

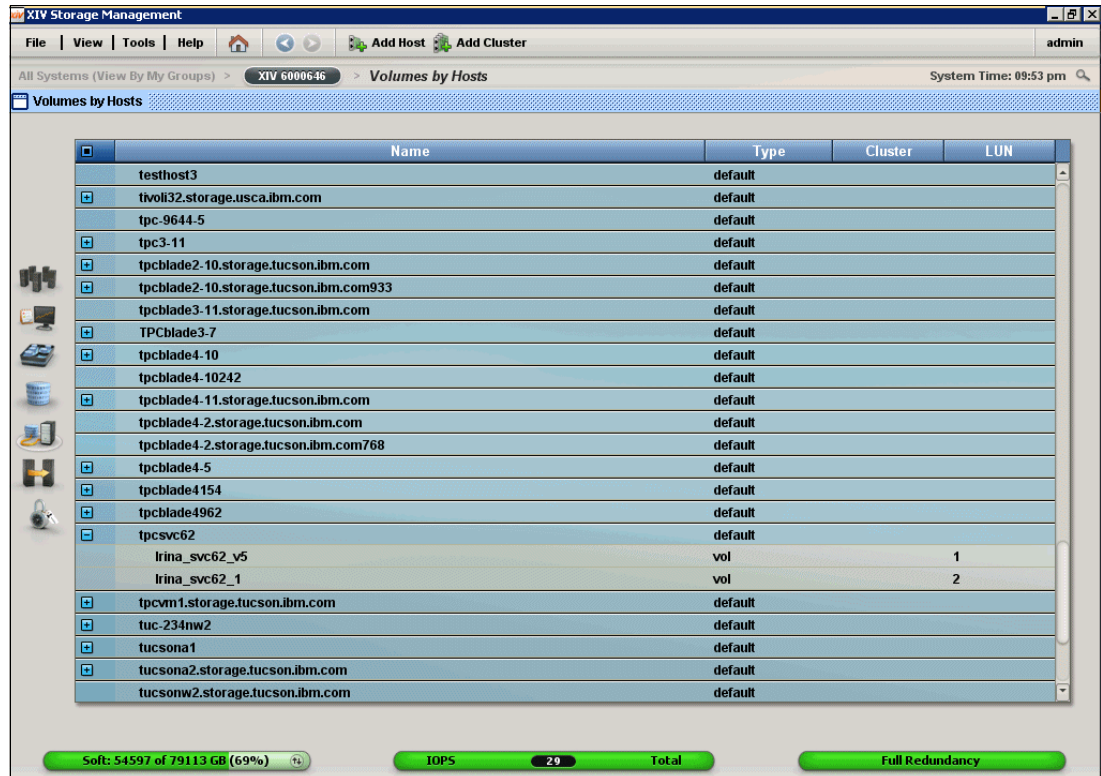


Figure 3-8 XIV Volumes by Hosts view

The MDisk UID field consists of part of the controller WWNN from bits 2 - 13. You might check those bits by using the **svcininfo lscontroller** command, as shown in Example 3-4.

Example 3-4 The lscontroller command

```
IBM_2145:tpcsvc62:admin>svcsinfo lscontroller 10
id 10
controller_name controller10
WWNN 5001738002860000
...
```

The correlation can now be performed by taking the first 16 bits from the MDisk UID field. Bits 1 - 13 refer to the controller WWNN, as shown in Example 3-4. Bits 14 - 16 are the XIV volume serial number (897) in hexadecimal format (resulting in 381 hex). The translation is 0017380002860381000000000000000000000000000000000000000000000000, where 0017380002860 is the controller WWNN (bits 2 - 13) and 381 is the XIV volume serial number that is converted in hex.

To correlate the Spectrum Virtualize ctrl_LUN_#, convert the XIV volume number in hexadecimal format and then check the last three bits from the Spectrum Virtualize ctrl_LUN_#. In this example, the number is 0000000000000002, as shown in Figure 3-9 on page 86.

[illegible]

Figure 3-9 MDisk details for XIV volume

Storwize volumes

The IBM Storwize solution is built upon the IBM Spectrum Virtualize technology base and uses similar terminology.

Complete the following steps to correlate the Storwize volumes with the MDisks:

1. From the Storwize side first, check the Volume UID field that was presented to the Spectrum Virtualize host, as shown in Figure 3-10 on page 87.

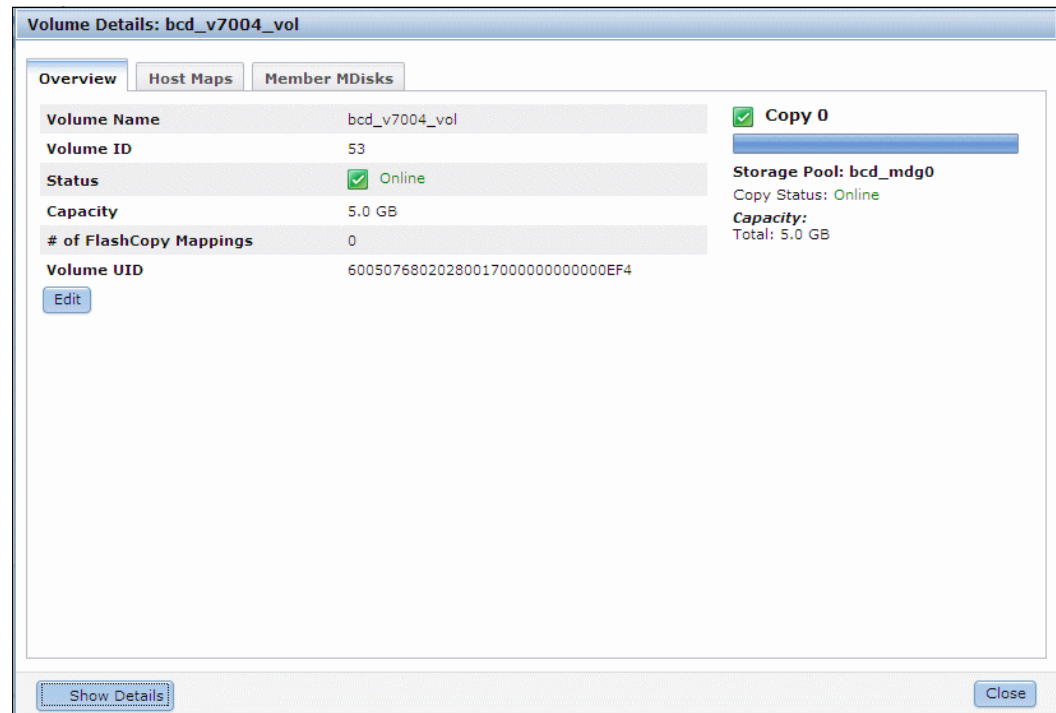


Figure 3-10 Storwize Volume details

- On the Host Maps tab (see Figure 3-11), check the SCSI ID number for the specific volume. This value is used to match the Spectrum Virtualize ctrl_LUN_# (in hexadecimal format).

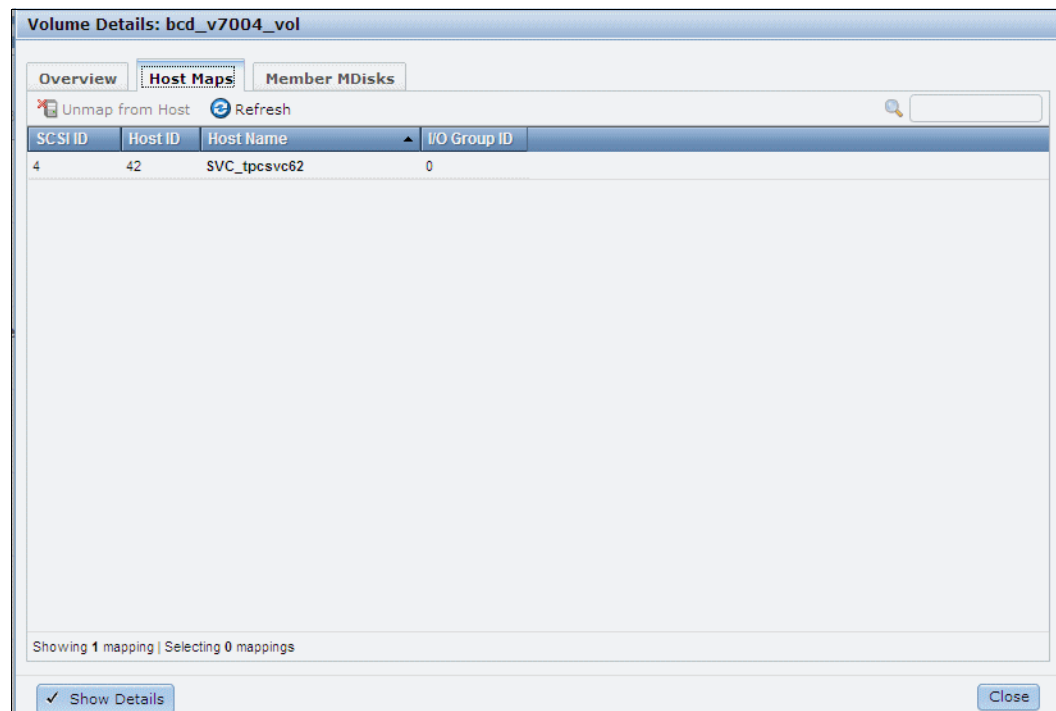


Figure 3-11 Storwize Volume Details for Host Maps

- On the Spectrum Virtualize side, review the MDisk details (see Figure 3-12) and compare the MDisk UID field with the Storwize Volume UID. The first 32 bits should be the same.

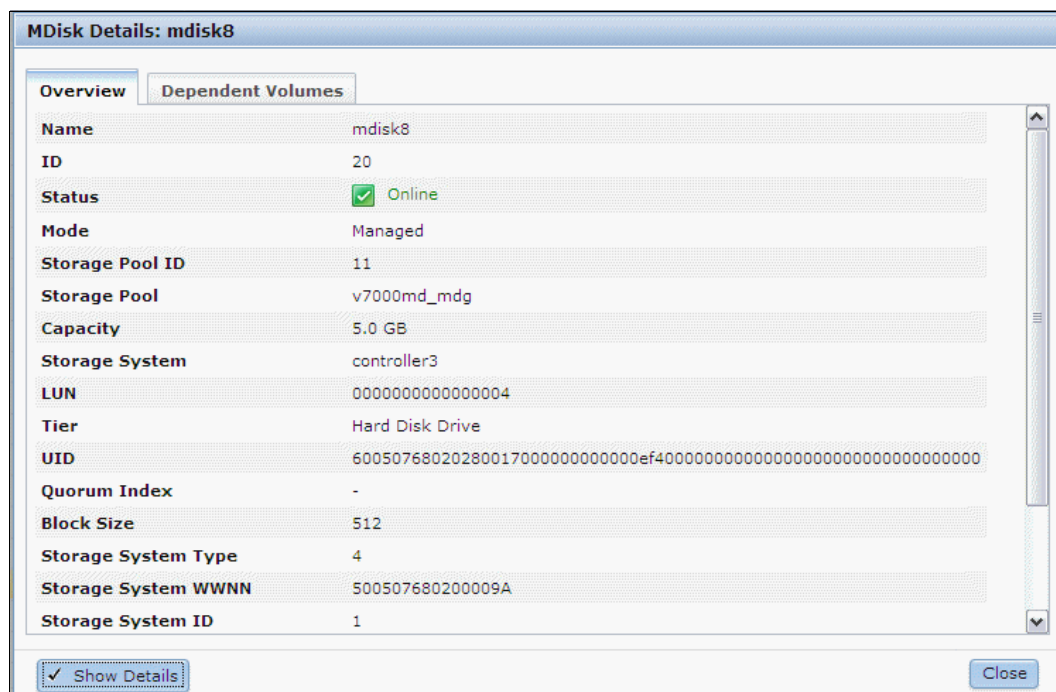


Figure 3-12 Spectrum Virtualize MDisk Details for Storwize volumes

- Double-check that the Spectrum Virtualize ctrl_LUN_# is the Storwize SCSI ID number in hexadecimal format. In this example, the number is 0000000000000004.

3.10 Remapping managed MDisks

Generally, you do not unmap managed MDisks from Spectrum Virtualize because this process causes the storage pool to go offline. However, if managed MDisks were unmapped from Spectrum Virtualize for a specific reason, the LUN must present the same attributes to Spectrum Virtualize before it is mapped back. Such attributes include UID, subsystem identifier (SSID), and LUN_ID.

If the LUN is mapped back with different attributes, Spectrum Virtualize recognizes this MDisk as a new MDisk and the associated storage pool does *not* come back online. Consider this situation for storage controllers that support LUN selection because selecting a different LUN ID changes the UID. If the LUN was mapped back with a different LUN ID, it must be mapped again by using the previous LUN ID.

Another instance where the UID can change on a LUN is when DS4000® support regenerates the metadata for the logical drive definitions as part of a recovery procedure. When logical drive definitions are regenerated, the LUN appears as a new LUN as it does when it is created. The only exception is that the user data is still present.

In this case, you can restore the UID on a LUN only to its previous value by using assistance from DS4000 support. The previous UID and the SSID are required. You can obtain both IDs from the controller profile.

3.11 Controlling extent allocation order for volume creation

With early version of Spectrum Virtualize, when you create a virtual disk, you might want to control the order in which extents are allocated across the MDisk 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 simply choosing the next disk in a round robin fashion. The pseudo random algorithm avoids the situation whereby the “striping effect” inherent in a round robin algorithm places the first extent for a large number of volumes on the same MDisk. Placing the first extent of a number of volumes on the same MDisk could lead to poor performance for workloads which place a large IO load on the first extent of each volume or which create multiple sequential streams. This effect had been seen to cause poor performance with early code, and today has been totally fixed.

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

3.12 Considerations when using Encryption

IBM 2145-DH8 and 2145-SV1 Spectrum Virtualize nodes and Storwize system provides optional encryption of data at rest, which protects against the potential exposure of sensitive user data and user metadata that is stored on discarded, lost, or stolen storage devices.

For more detailed information about Spectrum virtualize and Storwize Encryption, refer to *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V7.6*, SG24-7933

3.12.1 Encryption at rest with USB Key

As a general best practice here, when implementing the Encryption feature, consider a potential CPU usage increment of about 15% to 20% depending on the number of volumes encrypted.

Remember that Encryption is Storage Pool based, so to start using Encryption you have to:

- ▶ Create a new encrypted pool:
 - For existing data, move volumes from existing pools to new pools (there is no “convert in place” function to encrypt existing pools)
 - May require additional capacity as swing/working space

Keep in mind that:

- ▶ Encryption is enabled at the array level:
 - When enabled in a system, all new arrays/MDisks are created encrypted by default
 - Storage pools contain multiple arrays/MDisks
 - Volumes are usually striped across MDisks in a storage pool
 - With Easy Tier, volumes migrate among MDisk tiers in a pool

- To ensure all data is properly protected, encrypt all or none of the arrays/MDisks in a pool
- All volumes created in an encrypted pool are automatically encrypted

Attention: Pay special attention when encryption is enabled on some storage pools. You must have inserted a USB drive with the stored encryption keys, otherwise the data will not be readable after restart.

- ▶ Master keys are only persistently stored on USB devices
 - The master key is required on all nodes at all times, but mostly will be handled by the key manager
 - Master keys from the USB device are required at coldboot of the entire cluster
 - They are also required from the USB device for **T3** recovery with the IBM support team
- ▶ Storwize V7000 and Storwize V7000 Unified have built-in key management
- ▶ There are two types of keys
 - Master key (one per system)
 - Data encryption key (one per encrypted array)
- ▶ Master key is created when encryption enabled
 - Stored on USB devices
 - Stored as a simple file
 - May be copied or backed up
 - Required to use a system with encryption enabled
 - May be changed
 - Must be stored securely because enables access to encrypted data
- ▶ Master key is required for a system with encryption enabled
 - System will not operate without access to master key regardless of whether or not any arrays are encrypted
 - Protect the USB devices holding the master key and consider secure backup copies

Note: Only IBM USB devices are supported for encryption key use so be sure to order them in eConfig.

- ▶ Data encryption key is used to encrypt data and is created automatically when an encrypted array/MDisk is created
 - Stored in secure memory in SAS controller hardware
 - Stored encrypted with the master key
 - No way to view data encryption key
 - Cannot be changed
 - Discarded when an array is deleted (secure erase)

3.12.2 Encryption at rest with Key Server

IBM already supports Encryption managed by Key Server on other storage subsystems such as the DS88XX family and XIV.

Starting from V7.8 of Spectrum Virtualize and Storwize, Encryption with Key Server is supported as well.

Spectrum Virtualize and Storwize take advantage of key servers and the value that these bring to any solution for key generation and management. As well as supporting the use of USB devices for existing encryption solutions, Spectrum Virtualize and Storwize are capable of dealing with customers that are just happy to use USB devices to manage their keys, through to customers connected to multiple key servers with a variety of roles, regular re-keying, audits chains, and so on.

The key server is used to provide and manage keys that are used to unlock the hardware in a similar fashion to the USB Flash Drives. The server is accessed via TCP/IP and requires the cluster to be operational. In order to cater for deadlocks, T3 and other scenarios that need access to keys before the cluster is operational then an alternative would be required.

Figure 3-13 shows the basic difference between encryption with USB Key and encryption with Key server.

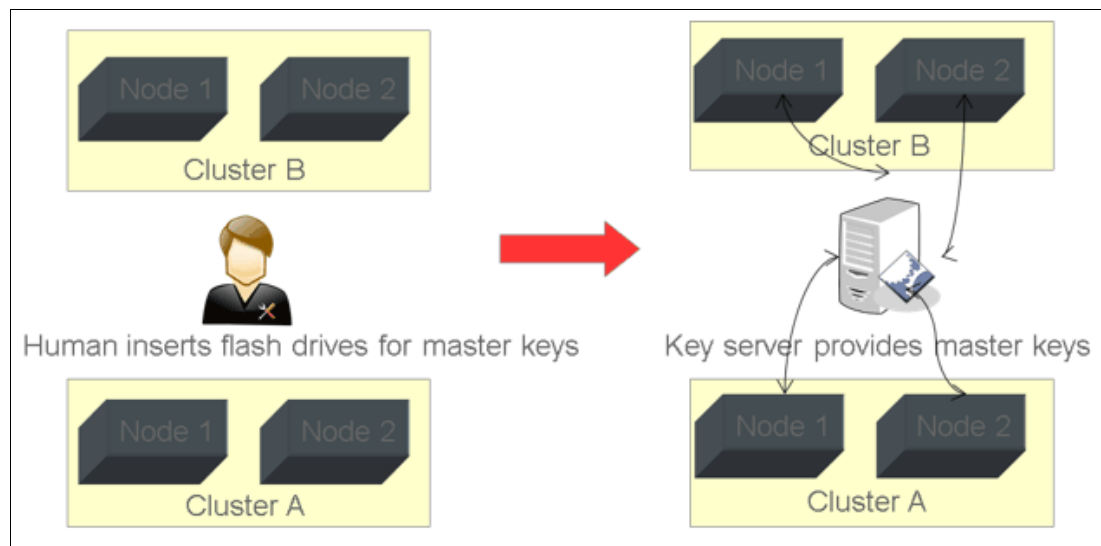


Figure 3-13 Encryption difference

Why use a key server

Key servers provide useful features that make them desirable to use. Here's some further detail about public key servers:

- ▶ Responsible for key generation
- ▶ Good use of entropy and random number generators
- ▶ Responsible for backups
- ▶ There is an open standard that key server implementations are following which will aid interoperability
- ▶ Audit detail
- ▶ Ability to administer access to data separately from storage device(s)

Managed vs unmanaged keys

Key servers support the concepts of managed vs unmanaged keys.

Managed keys are keys which the key server makes an effort to maintain, backup and make available to every key server end point that needs them. The keys are stored on the server and are fetched as and when they are needed.

Unmanaged keys are keys which the key server makes no effort to maintain. They do not get backed up or continuously made available to end points. Once the key server has generated unmanaged keys, the expectation is that they are then stored and maintained elsewhere.

Types of key server

The main area of difference between Key Server vendors is in the area of authentication. The operations of make/get/delete keys are otherwise being covered by the protocol. Due to its history, SKLM uses vendor-defined credentials. KeySecure uses all or none of the username/password credentials and client certificate credentials.

Security Key Lifecycle Manager

IBM Security Key Lifecycle Manager, formally known as Tivoli® Key Lifecycle Manager (TKLM). TKLM was based on the IPP protocol, but it was re-branded to SKLM and now also supports a protocol known as Key Management Interface Protocol (KMIP). However, since KMIP didn't exist when TKLM was originally made, SKLM's models pre-date KMIP and make it a slightly different beast to generic KMIP.

The SKLM server code can be downloaded and installed via Extreme Leverage, search for 'IBM Security Key Lifecycle Manager V2.6.0 eAssembly Multiplatform, Multilingual'.

- ▶ An SKLM server requires an IP address, IP port and device group.
- ▶ SKLM can create and manage keys in different ways. For Spectrum Virtualize and Storwize, IBM used the 'Spectrum Scale' device family which creates managed keys for the client and provides access to them via the client certificate. It then requests that the user makes a backup of the keys and then restore them onto redundant SKLM servers.
- ▶ SKLM also acts as a key foundry allowing ISKLM to be used to generate unmanaged keys.

At the time of writing Spectrum virtualize and Storwize have some restriction in Encryption implementation with Key Server:

- ▶ Only SKLM key servers are supported
- ▶ Only one key server end point (only master, no clones) is supported
- ▶ Key server or USB encryption, not both

In time these restrictions will be removed and enhancements will be added.

For the most up to date information about Encryption with Key Server, check the Spectrum Virtualize Knowledge Center at:

<https://ibm.biz/BdsSSq>

and the SKLM Knowledge Center at:

<https://ibm.biz/BdsvEP>

3.12.3 Encryption on external storage

With Spectrum Virtualize and Storwize, when you are virtualizing external storage subsystems, it might be the case that some of those storage subsystems have in place their own encryption.

Encrypting MDisk already encrypted does not make any sense and it will waste CPU resources in your Spectrum Virtualize or Storwize system.

If you are in this position, it is suggested to configure the external MDisk supplied by an external storage subsystem, with the parameter **-encrypt yes** in the **chmdisk** command as shown in Example 3-5. This is really important if you use Spectrum Virtualize in front of an encrypted applying encryption to your Storwize V7000 system.

Example 3-5 chmdisk -encrypt example

```
IBM_2145:SVC_ESC:superuser>chmdisk -encrypt yes my_encrypted_mdisk_0
```

In Example 3-5 **-encrypt yes** | **no** specifies whether the MDisk is encrypted using its own encryption resources.

Important:

- ▶ If you apply encryption to your system, you must identify the encrypted MDisk before applying encryption.
- ▶ If you specify **chmdisk -encrypt**, the setting is permanent in SAN Volume Controller.
- ▶ Do not use the **-encrypt** parameter if one of the MDisk groups has an encryption key, parent pool, and child pools.
- ▶ Use **chmdisk** for existing self-encrypting MDisk before starting any migration.
- ▶ If an MDisk is self-encrypting, i.e. when the MDisk is supplied by external Storage System and is already encrypted.

Users that move to an encrypted solution and then migrate their data are at risk of leaving behind an unencrypted version of the data on their MDisk that will eventually get overwritten when that extent/block is re-used.

In some scenarios, customers might not care, however, if they do, then they need to be aware that they should remove an MDisk that may contain an unencrypted copy of their data in a no longer used extent, format the MDisk to erase such data, and then add the MDisk back to the system.

Note: Just migrating to an encrypted pool doesn't delete the old data unless actions are taken to do so.



Volumes

This chapter explains how to create, manage, and migrate volumes (formerly known as vdisks) across I/O groups. It also explains how to use IBM FlashCopy.

This chapter includes the following sections:

- ▶ Overview of volumes
- ▶ Creating volumes
- ▶ Creating volumes
- ▶ Volume migration
- ▶ IPreferred paths to a volume
- ▶ Non-Disruptive volume move (NDVM)
- ▶ VMware Virtual Volume (VVols)
- ▶ Cache mode and cache-disabled volumes
- ▶ Volume Governance
- ▶ Using Spectrum Virtualize or Storwize with FlashSystem

4.1 Overview of volumes

Three types of volumes are possible: striped, sequential, and image. These types are determined by how the extents are allocated from the storage pool.

A *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.

An *image-mode volume* is a one-to-one mapped extent mode volume.

4.1.1 Striping compared to sequential type

With a few exceptions, you must always configure volumes by using striping. One exception is for 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 to configuration by using volume striping is an environment with a high dependency on many flash copies. In this case, FlashCopy loads the volumes evenly, and the sequential I/O, which is generated by the flash copies, has a higher throughput potential than what is possible with striping. This situation is rare considering the unlikely need to optimize for FlashCopy as opposed to an online workload.

The general rule now is to always go with striped. The main target is to take the 60 MDisk queue depth into account, which typically says for HDD MDisks, to aim for 8 spindles per MDisk. Previous cache algorithms meant that for sequential workloads, and to avoid stripe on stripe issues we suggested keeping 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 is was added with version 7.3).

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 back-end and presented to Spectrum Virtualize or Storwize as 8 MDisks. This means when we get the 60 MDisk queue, we get roughly a queue depth of 8 per drive, which keeps a spinning disk well utilized. It also gives better concurrency across the ports of the back-end controller.

There could be other scenarios, for example when Spectrum Virtualize or mostly Storwize are acting as back-end storage for IBM TS7650G ProtecTIER® Gateway. In this scenario it is still suggested to use sequential volumes mostly when using disk drives with very big sizes such as 2TB or 3TB for user data repository. The reason is that those big disk drives will end up to have very big array/MDisk/LUN. If ProtecTIER will handle this big LUN by itself it will be able to optimize its File System structure and workload without over committing or congesting the single array, rather than striping the LUNs over an entire multi array Storage Pool.

4.1.2 Thin-provisioned volumes

Volumes can be configured as *Thin-provisioned* or *fully allocated*. Thin-provisioned volumes are created with real and virtual capacities. 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 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, 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 thus has a zero contingency capacity, goes offline when the real capacity is used and must expand.

Warning threshold: Enable the warning threshold (by using e-mail or an SNMP trap) when you are working with Thin-provisioned volumes, on the volume, and on the storage pool side, especially when you do not use the autoexpand mode. Otherwise, the thin volume goes offline if it runs out of space.

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.

4.1.3 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 is defined when the volume is created and can be 8KB, 32 KB, 64 KB, 128 KB, or 256 KB (default). **Note:** 16KB is not supported.

Smaller granularities can save more space, but they have larger directories. When you use Thin provisioning with FlashCopy, specify the same grain size for the Thin-provisioned volume and FlashCopy.

The 8KB grain is an optimization for particular clients using flash storage that want to achieve the maximum flash storage space usage optimization for better protection of their investment on valuable disk space.

4.1.4 Compressed volumes

A compressed volume is, first of all, a Thin-provisioned volume. The compression technology is implemented into the 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 *Real-time Compression in SAN Volume Controller and Storwize V7000*, REDP-4859, which is available at:

<http://www.redbooks.ibm.com/abstracts/redp4859.html?Open>

IBM recommends that when using Real-time Compression (RtC), always use 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 in order to perform a space and performance assessment first.

Use the RtC estimator tool that is available on your Spectrum Virtualize and Storwize Command Line Interface starting with version 7.6 and with the GUI starting from version 7.7 to identify the best volume candidates to be compressed.

When using the CLI use the commands, as shown in Example 4-1, to run volume analysis on a single volume.

Example 4-1 analyzevdisk example

```
IBM_2145:SVC_ESC:superuser>svctask analyzevdisk -h
.
analyzevdisk
.
Syntax
.
>>- analyzevdisk -- --+-----+-- --+ vdisk_id ---+-----><
                        '- -cancel-'      '- vdisk_name -'
.
For more details type 'help analyzevdisk'.
.
IBM_2145:SVC_ESC:superuser>svctask analyzevdisk fcs2
IBM_2145:SVC_ESC:superuser>
```

When using the CLI use the commands, as shown in Example 4-2, to run Volume analysis for an entire subsystem.

Example 4-2 analyzevdiskbysystem example

```
IBM_2145:SVC_ESC:superuser>svctask analyzevdiskbysystem -h
```

```

.
analyzevdiskbysystem
.
Syntax
.
>>- analyzevdiskbysystem -- --+-----+-- -----><
                                '- -cancel-'
.
For more details type 'help analyzevdiskbysystem'.
.
IBM_2145:SVC_ESC:superuser>svctask analyzevdiskbysystem
IBM_2145:SVC_ESC:superuser>

```

Note: `analyzevdisk` and `analyzevdiskbysystem` will return to the prompt.

To see the result of the analysis and its progress run the CLI command as shown in Example 4-3.

Example 4-3 lsvdiskanalysis example

```

IBM_2145:SVC_ESC:superuser>svcinfolsvdiskanalysis
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 fcs0 sparse 161011155433 5.00GB 0.00MB 0.00MB 0
0.00MB 0.00MB 0
0.00MB 0 0
.
lines omitted for brevity
.
8 tgtrm sparse 161011155438 5.00GB 0.00MB 0.00MB 0
0.00MB 0.00MB 0
0.00MB 0 0
IBM_2145:SVC_ESC:superuser>svcinfolsvdiskanalysisprogress
vdisk_count pending_analysis estimated_completion_time
9 0

```

When using the GUI go to the menu as shown in Figure 4-1 on page 100 to run volume analysis by single volume or by multiple volumes selecting all the volumes you need to be analyzed.

From the same menu shown in Figure 4-1 on page 100, you can download the Saving report in .csv format.

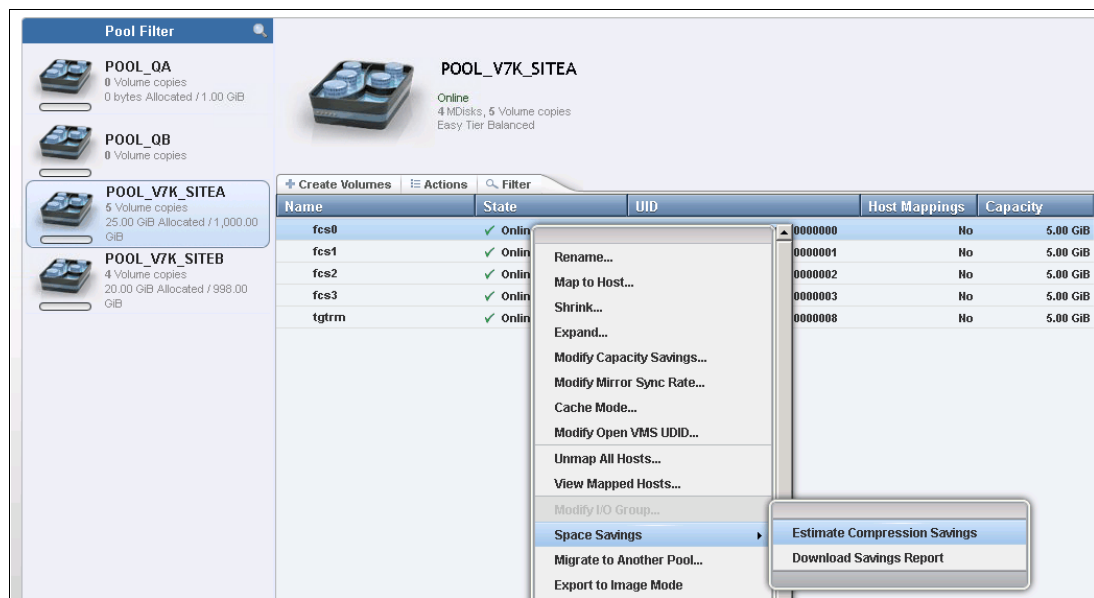


Figure 4-1 Use of Estimate Compression Saving with GUI

In case you are planning to virtualize volumes that currently are connected to your hosts directly from any storage subsystems, and you want to know what the space saving you will achieve is using RtC on those volumes, run the Comprestimator Utility available at:

<https://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, and the above link provides all the instructions needed.

The best practices here are:

- ▶ 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 trade off between space saving and hardware resource consumption to compress your data may not make sense.
- ▶ After you have compressed your selected Volumes, take a look at what volumes have the most space saving benefits from Thin Provisioning rather than RtC. Consider not compressing those ones, but simply move them to Thin Provisioning only. It will require some effort, but will save hardware resources that will be available to give better performance to those Volumes that will achieve most 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 on page 101. Customize the Volume view to get all the metrics you might need to help make your decision.

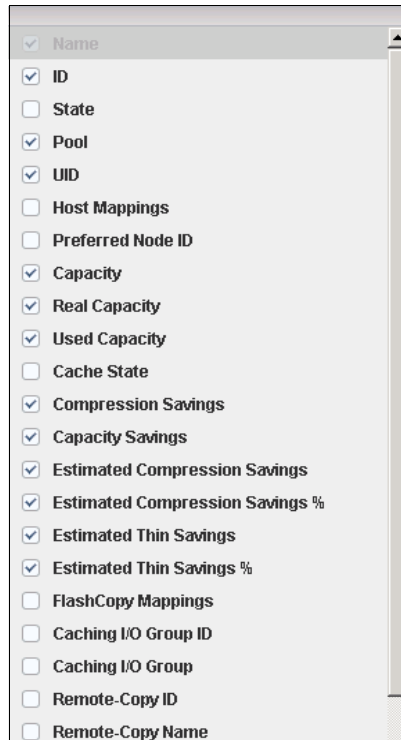


Figure 4-2 Customized view

4.1.5 Thin-provisioned volume

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 allocated the space.

A volume that is Thin-provisioned by SVC or Storwize is a volume where the large chunk of binary zeros are not stored in the storage pool. So, if you've never written to the volume yet, then you don't need to use valuable resources storing data that doesn't 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 on page 102 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 filesystems

There are a number of different properties of Thin-provisioned volumes. It will be useful to understand them for the rest of the explanations later:

- ▶ The size of the volume presented to the host. This doesn't really have a name, but we will refer to this as '*volume capacity*'.
- ▶ The amount of user data that has been written to the storage pool. This is called the '*used capacity*'.
- ▶ The capacity which 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.
- ▶ There is also a warning threshold.
- ▶ For a Compressed Volume only (because Compressed Columns are based on Thin-provisioned Volumes), there is the amount of uncompressed user data which has been written into the volume. This is called the '*uncompressed used capacity*'.
 - It is used to calculate the compression ratio, where:

$$((\text{uncompressed used capacity} - \text{used capacity}) / \text{uncompressed used capacity} * 100 = \text{compression ratio})$$

Since there are at least two ways of calculating compression ratios, It useful to remember that bigger is better - 90% compression ratio is better than 50% compression ratio.

As we stated, Thin provisioning means “don't store the zeros”, so what does overallocation mean? Simply put, a storage pool is only overallocated once 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, then if 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 isn't 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 monitor it you must.

Even if you are creating a new storage pool, it is likely that you won't start over provisioning for a few weeks after you start writing to that pool. You don't 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. The capacity planner needs to monitor the amount of capacity being used versus the capacity of the storage pool, and make sure you purchase more capacity before you actually 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 growth. Thus you need to approach capacity planning carefully.

To avoid unpleasant situations where some volumes may go offline due to lack of space, the storage administrator needs to monitor the real capacity rather than the volume capacity. And that's 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 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.

Spectrum Virtualize or Storwize will 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 will set this to 80% of the capacity of the storage pool by default, although you can change it.

It is really suggested to have event notifications turned on so that someone gets an e-mail or pop up on your monitoring system when this error gets added to the event log. Note that this event will not call home to IBM - you need to respond to this yourself.

What to do if you run out of space

There are a lot of options here. You can use just one of these options, or a combination of as many as you like.

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 then if the system does not have any more capacity to store the host writes, then the volume will go offline. But it's not only that one volume that goes offline, all the volumes in the storage pool are now at risk of going offline.

So what mechanisms and processes should you put in place for when this happens for real?

- ▶ Automatic out of space protection provided by the product
- ▶ Buy more storage

If the storage pool has run out of space, each volume now has its own emergency capacity. That emergency capacity is normally pretty sizable (2% is the default). The emergency capacity which is dedicated to a volume could allow that volume to stay online for anywhere between minutes to days depending upon the change rate of that volume. This means that when you run out of space, you do have some time to repair things before everything starts going offline.

So you could implement a policy of 10% emergency capacity per volume if you wanted to be safer. Also remember that you don't need to have the same contingency capacity for every volume.

Note: This automatic protection will probably solve the vast majority of immediate problems, but remember that once you are woken up to be told that you've run out of space, this automatic protection won't buy you an infinite amount of time. You need to know what you are going to do next.

- Have unallocated storage sitting as standby.

You can always have one or more managed disks, or a collection of drives sitting on the floor. These spare drives or managed disks can be added to whichever storage pool runs out of space within only a few minutes. This gives you some breathing room whilst you take some of the other actions. The more managed disks or drives that you have available the more breathing room you have.

- Move or delete volumes

Once you have run out of space, you always have the option to migrate volumes to other pools to free up space. This is perfectly reasonable, but remember that data migration on Spectrum Virtualize and Storwize is designed to go slowly to avoid causing performance problems, so it may be that you can't complete this migration before your favorite application goes offline.

Of course a very rapid but dire solution is to delete one or more volumes to make space. This is not really recommended, but it's something on your list. Especially if you are sharing the storage pool with both production and development. You (with management approval of course) may choose to sacrifice less important volumes to preserve the critical volumes.

- Policy based solutions

Of course no policy is going to solve the problem if you actually run out of space, but you could use policies to reduce the likelihood of that ever happening to the point where you feel comfortable doing less of the other options.

What type of policies could you use for Thin provisioning?

Note: In the policies below we used 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 this is all about business risk and it is something different for any 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 100TB of capacity in the storage pool, then the sum of the volume capacities in the same pool must not exceed 200TB.
- Always start the process of buying more capacity when the storage pool reaches 60% full.
- If you keep your FlashCopy backups and your production data in the same pool, you could choose to not overallocate the production data, so that if you run out of space you could simply start deleting the backups.
- Child Pools

Version 7.4 introduced a feature called child pools which allows you to make a storage pool which 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 would prevent any problems with a server in child pool A affecting a server in child pool B. If Child Pool A ran out of space, and the parent pool still had space then you can easily grow the child pool.
- You could use child pools to create a child pool which 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 a set of emergency capacity that can be given back to the parent pool (after you've worked out which server was eating up all the space and stopped whatever it was doing).

For more information about Thin Provisioning usage and best practices read:

<https://ibm.biz/BdschV>

4.1.6 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 4-1 shows the maximum Thin-provisioned volume virtual capacities for an extent size.

Table 4-1 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	524,288	520,000
8192	1,048,576	1,040,000

Table 4-2 show the maximum Thin-provisioned volume virtual capacities for a grain size.

Table 4-2 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

4.2 Creating volumes

To create volumes, follow the procedure that is described in *Implementing the IBM System Storage SAN Volume Controller with IBM Spectrum Virtualize V7.6*, SG24-7933.

When you are creating volumes, adhere to the following guidelines:

- ▶ 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 afterward.
- ▶ 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. For information about moving a volume across I/O groups, see , “The functionality used to change preferred node across I/O groups is named Non-Disruptive Volume Move (NDVM)” on page 107.

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.

- ▶ 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 4.2.1, “Changing the preferred node within an I/O group or cross I/O group” on page 107.
- ▶ In Stretched Cluster environments, it is recommended to configure the *preferred node* that is based on site awareness.
- ▶ The maximum number of volumes per I/O group, at the time of writing, is 2048 and 8192 per cluster for versions up to 7.7.x. These limits can change from time to time with newer versions. You must always confirm the limits related to your specific version as shown in the following links for version 7.7.x.

<http://www.ibm.com/support/docview.wss?uid=ssg1S1005826>

and for version 7.8.x.

<http://www.ibm.com/support/docview.wss?uid=ssg1S1009558>

- ▶ The smaller the extent size is that you select, the finer the granularity is of the volume of space that is occupied on the underlying storage controller. 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.

The counter view to this view is that, the smaller the extent size is, the smaller the total storage volume is that the Spectrum Virtualize or Storwize can address. The extent size does not affect performance. For most clients, extent sizes of 128 MB or 256 MB give a reasonable balance between volume granularity and cluster capacity. Extent size is set during the Storage Pool creation.

Important: You can migrate volumes using the `migratevolume` command only between storage pools that have the same extent size, except for mirrored volumes that can allocate space on different storage pools with different extent size.

As described in 4.1, “Overview of volumes” on page 96, 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 striping mode.

4.2.1 Changing the preferred node within an I/O group or cross I/O group

Currently, a nondisruptive method is available to change the preferred node within an I/O group and across I/O groups. The correct way is to migrate the volume to a recovery group and migrate back with the preferred node.

Changing the preferred node within an I/O group is nondisruptive but can lead to some delay in performance and for some specific operating systems or application could affect some specific time out.

Changing the preferred node within an I/O group can always be done using the CLI and with the GUI only if you have at least two I/O groups.

To change the preferred node across I/O groups there are some limitations, mostly in a Host Cluster environment. Read carefully the *Supported Hardware List, Device Driver, Firmware and Recommended Software Levels for Spectrum Virtualize and Storwize* for your specific version available at:

<http://www.ibm.com/support/docview.wss?uid=ssg1S1009558>

and read also the *Configuration Limits and Restrictions for IBM System Storage SAN Volume Controller* for your specific version available at:

<http://www.ibm.com/support/docview.wss?uid=ssg1S1009560>

The functionality used to change preferred node across I/O groups is named Non-Disruptive Volume Move (NDVM)

Attention: These migration tasks can be nondisruptive if performed correctly and hosts that are mapped to the volume support nondisruptive volume move (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. 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 can be done when 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.

Ensure that you create paths to I/O groups on the host system. After the system successfully added the new I/O group to the volume's access set and you 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 as per the instruction shown in the IBM Knowledge Center for Spectrum Virtualize that is available at:

<https://ibm.biz/BdsvXi>

4.3 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 4-3 on page 108.

Table 4-3 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 by using the volume. Depending on the workload of 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.

4.3.1 Image-type to striped-type migration

When you are migrating existing storage into the 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`.

In general, *image-type volume* is migrated to a *striped-type volume*, to make it more easily managed together with all other striped volumes in the cluster, even if an *image-type-volume* can apply to all the functionality as any other regular volume but can not be expanded. Example 4-4 on page 109 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 4-4 migratevdisk command

```
IBM_2145:svccg8:admin>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 the use of 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 V7.6*, SG24-7933.

You can monitor the migration process by using the `svcinfolsmigrate` command, as shown in Example 4-5.

Example 4-5 Monitoring the migration process

```
IBM_2145:svccg8:admin>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:svccg8:admin>
```

4.3.2 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.

Image-type disks are used to migrate existing data to a Spectrum Virtualize or Storwize and to migrate data out of virtualization. Image-type volumes cannot be expanded.

Often 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 reside 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.

Remember: This command cannot be used to if the source volume copy is in a child pool or if the MDisk group that is specified is a child pool.

This command does not work if the volume is fast formatting.

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.

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 owned by a filesystem 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 MDisk 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.

Spectrum Virtualize and Storwize **migratetoimage** command is very powerful and makes your life easier when you might want to use your system as *data mover*. To better understand all requirements and specification for that command refer to IBM Knowledge Center available at:

<https://ibm.biz/BdsvX2>

4.3.3 Migrating with volume mirroring

Volume mirroring offers the facility 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.

The best practice here is to try to not overload the systems with a too high *syncrate* and do 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 Figure 4-4 for the supported *-syncrate* values and their corresponding rates. Use this parameter to alter the rate at which the fully allocated volume or mirrored volume format before synchronization

User-specified <i>syncrate</i> 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

Figure 4-4 *syncrate* values

For further information refer to IBM Knowledge Center at:

<https://ibm.biz/Bdsvxb>

4.4 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. With the addition of Virtual Volumes, we enable a fundamentally more efficient operational model for storage in virtualized environments, centering it around the virtual machine (VM) rather than the physical infrastructure.

Before the arrival of Virtual Volumes, a virtual machine disk (VMDK) would be presented to a virtual machine (VM) in the form of a file. This file represents a disk to the VM, which is then accessed by the guest operating system in the same way as a physical disk is accessed on a physical server. This VMDK is stored on a VMware file system (VMFS) formatted Datastore.

The VMFS Datastore is hosted by a single volume on a storage system such as IBM Spectrum Virtualize or Storwize. A single VMFS Datastore, sometimes referred to as the VMFS blender, could have hundreds or even thousands of VMDKs.

Virtual Volumes provides 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 utilized at the VVol level, when performing snapshot and clone operations.

For detailed information about VVols prerequisites, implementation and configuration in Spectrum Virtualize or Storwize environment, refer to: *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.

4.5 IPreferred paths to a volume

For I/O purposes, Spectrum Virtualize and Storwize nodes within the cluster are grouped into pairs, which are called *I/O groups* (sometimes cache I/O group). 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.

Spectrum Virtualize and Storwize implements 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 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, Spectrum Virtualize and Storwize assigns 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*. Because it is expected that hosts access volumes through the preferred nodes, those nodes can become overloaded. When a node becomes overloaded, volumes can be moved to other I/O groups because the ownership of a volume cannot be changed after the volume is created. For more information, see 4.2.1, “Changing the preferred node within an I/O group or cross I/O group” on page 107.

Multipathing Software, SDDPCM or SDDDSM (SDD for brevity) is aware of the preferred paths that Spectrum Virtualize or Storwize sets per volume. SDD uses a load balancing and optimizing algorithm when failing over paths. That is, it tries the next known preferred path. If this effort fails and all preferred paths were tried, it load balances on the nonpreferred paths until it finds an available path. If all paths are unavailable, the volume goes offline. Therefore, it can take time to perform path failover when multiple paths go offline. SDD also performs load balancing across the preferred paths where appropriate.

Sometimes when debugging performance problems it can be useful to take a look to a specific metric in Spectrum Control, named *Non-Preferred Node Usage Percentage*. 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 further information about this metric and more, refer to IBM Spectrum Control™ Knowledge Center available at:

<https://ibm.biz/BdsBsZ>

4.5.1 Governing of volumes

I/O governing effectively throttles the number of I/O operations per second (IOPS) or MBps that can be achieved to and from a specific volume. You might want to use I/O governing if you have a volume that has an access pattern that adversely affects the performance of other volumes on the same set of MDisk. An example is a volume that uses most of the available bandwidth.

If this application is highly important, you might want to migrate the volume to another set of MDisk. However, in some cases, it is an issue with the I/O profile of the application rather than a measure of its use or importance.

Base the choice between I/O and MB as the I/O governing throttle on the disk access profile 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 much throttling. It is better to use an IOPS throttle.

Conversely, 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, setting an I/O governing throttle that is based on IOPS does not achieve much throttling. For a streaming video application, it is better to use an MBps throttle.

Quality of Services Enhancement

As we already stated, I/O throttling is a mechanism to limit the volume of I/O processed by a storage system. Throttling primarily limits IOPS or Bandwidth available to a volume or a host and I/O rate is limited by queuing I/O requests if preset limits are exceeded.

On previous version we could limit IOPS per volume or MB/s per volume, and we could not limit both at the same time.

Starting with Spectrum Virtualize and Storwize V7.7 we have two kinds of throttling.

The new I/O throttling works at a finer granularity interval, short bursts are allowed to avoid delays in workload I/O bursts, and fairness across throttled and incoming I/Os.

Figure 4-5 on page 114 shows how throttle algorithm interacts with I/Os.

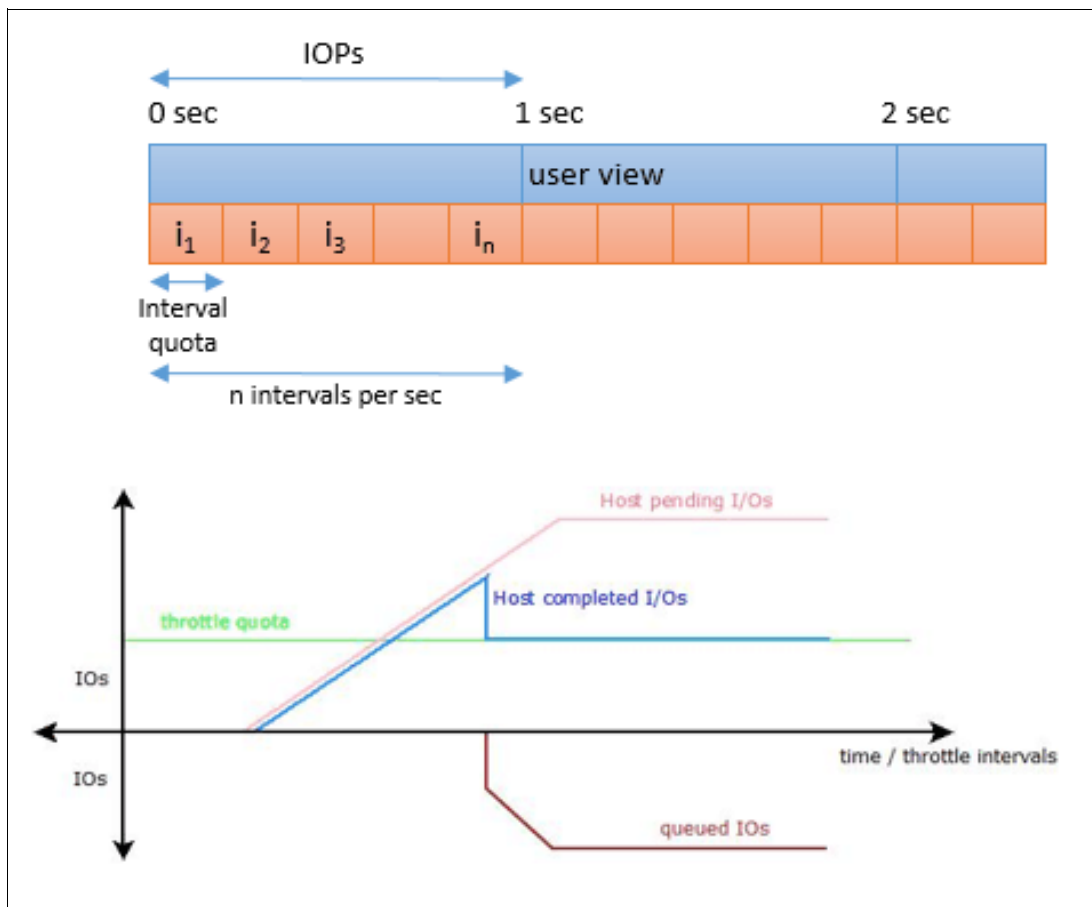


Figure 4-5 Throttle activity

Volume throttles

With Volume throttles, a per volume throttle can be configured, and IOPs and/or bandwidth limits can be set. V7.7 has also introduced a node level throttle enforcement and Standard SCSI read and write operations are throttled.

Throttling at a volume level can be set using the **chvdisk** and **-rate throttle_rate -unitmb** parameter.

It specifies the I/O governing rate for the volume, which caps the amount of I/O that is accepted. The default throttle_rate units are I/Os. By default the throttle_rate is disabled. To change the throttle_rate units to megabytes per second (MBps), specify the **-unitmb** parameter. The governing rate for a volume can be specified by I/Os or by MBps, but not both. However, you can set the rate to I/Os for some volumes and to MBps for others. When the Input/Output Operations Per Second (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.

Note: To disable the throttling on a specific volume, set the throttle_rate value to zero.

To set Volume throttle use the **chvdisk** command and a throttle object will be created. Then you can list your created throttle objects using the **lsthrottle** command and change their parameters with the **chthrottle** command. We show some command examples in Example 4-6.

Example 4-6 Throttle command example

```
IBM_2145:SVC_ESC:superuser>chvdisk -rate 100 -unitmb fcs0
```

```
IBM_2145:SVC_ESC:superuser>lsthrottle
```

throttle_id	throttle_name	object_id	object_name	throttle_type	IOPs_limit	bandwidth_limit_MB
0	throttle0	0	fcs0	vdisk	100	

```
IBM_2145:SVC_ESC:superuser>chthrottle -iops 1000 fcs0
```

```
IBM_2145:SVC_ESC:superuser>lsthrottle
```

throttle_id	throttle_name	object_id	object_name	throttle_type	IOPs_limit	bandwidth_limit_MB
0	throttle0	0	fcs0	vdisk	1000	100

```
IBM_2145:SVC_ESC:superuser>lsthrottle fcs0
```

```
id 0
throttle_name throttle0
object_id 0
object_name fcs0
throttle_type vdisk
IOPs_limit 1000
bandwidth_limit_MB 100
```

Offload throttles

Starting with Spectrum Virtualize and Storwize version 7.7 we support an Offload I/O throttle.

Offload throttles are applied to xcopy, and write_same primitive scsi command that are used in the VMware environment with VAAI.

Throttle objects can have IOPs and/or bandwidth limits and can be created per volume. I/Os are queued if I/O flow exceeds configured limits and queuing has microsecond granularity.

Figure 4-6 on page 116 shows the throttles flow.

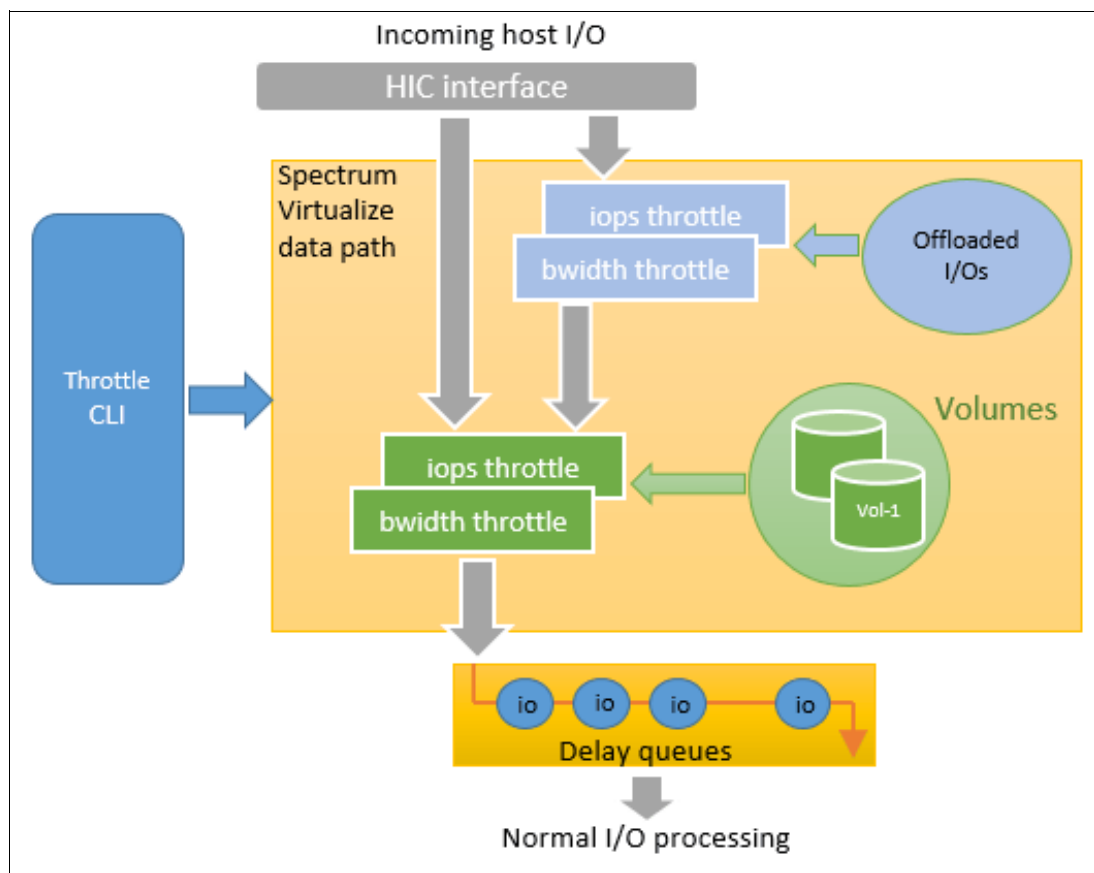


Figure 4-6 Throttles flow

To configure offload throttles use **mkthrottle**, **chthrottle** and **lsthrottle** commands as shown in Example 4-7.

Example 4-7 throttle command example

```
IBM_2145:SVC_ESC:superuser>mkthrottle -type offload -bandwidth 200 -iops 2000
-name OffThrottle
Throttle, id [1], successfully created.
```

```
IBM_2145:SVC_ESC:superuser>lsthrottle
throttle_id throttle_name object_id object_name throttle_type IOPs_limit
bandwidth_limit_MB
0          throttle0      0          fcs0          vdisk          1000          100
1          OffThrottle     0          fcs0          offload        2000          200
```

```
IBM_2145:SVC_ESC:superuser>chthrottle -iops 5000 OffThrottle
```

```
IBM_2145:SVC_ESC:superuser>lsthrottle OffThrottle
```

```
id 1
throttle_name OffThrottle
object_id
object_name
throttle_type offload
IOPs_limit 5000
bandwidth_limit_MB 200
```

Benefits of throttling

Throttling has these benefits:

- Manage performance impact of offloaded I/Os
 - Offloaded I/O for VM management
 - VMware uses XCOPY and WriteSame (EagerZeroedThick and Storage VMotion)
 - Microsoft HyperV uses ODX
 - Offloaded I/O commands have small footprint but can generate huge controller activity that can severely impact regular SCSI I/O performance. Using offloaded throttle:
 - Limit bandwidth used by offloaded I/Os
 - Reduce performance impact on regular SCSI I/Os

Figure 4-7 on page 117 shows an offloaded I/Os example.

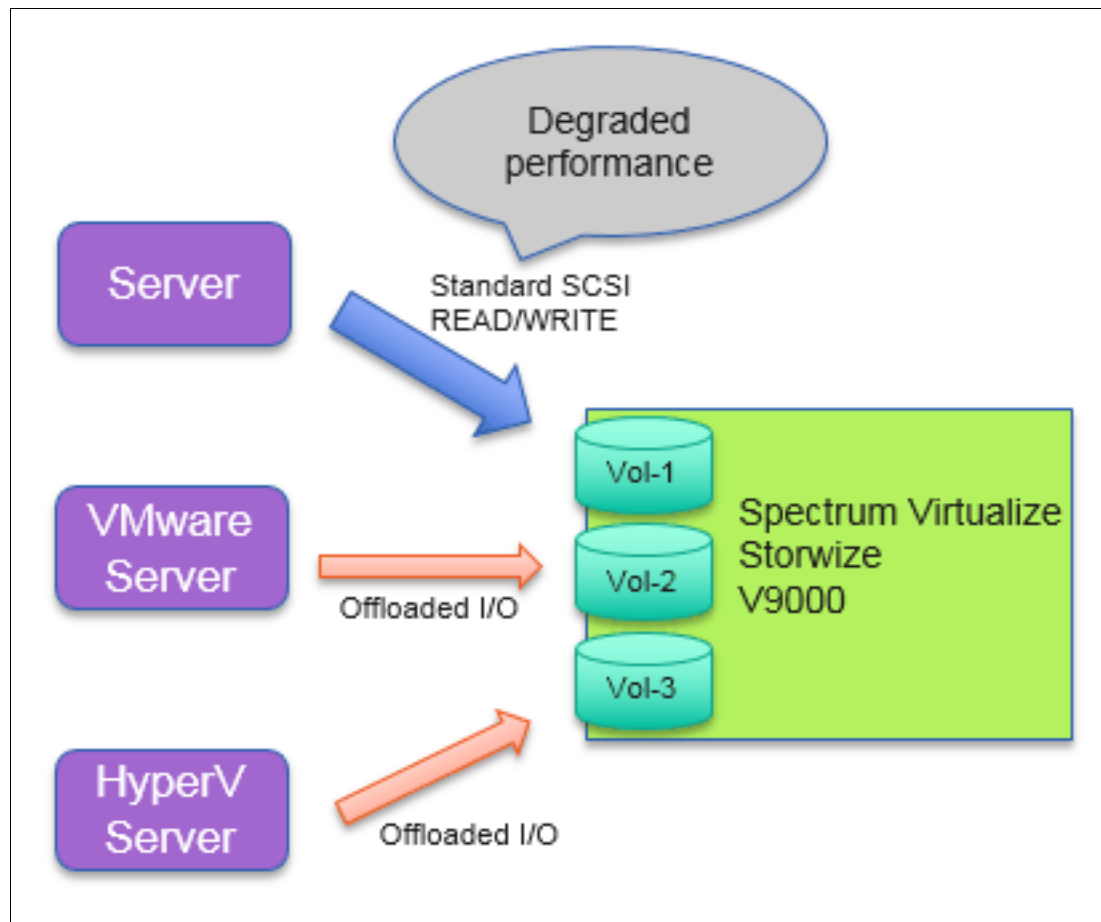


Figure 4-7 Offloaded I/O example

- Bandwidth consumed by secondary applications
 - Secondary applications like backup, data mining jobs generate bandwidth intensive workload
 - May adversely impact production job bandwidth and latency
 - Applying throttle on volume copy may improve production job performance utilizing the primary/source volume

- Fairness amongst large number of volumes
- Smoothing of I/O bursts
- Bandwidth and IOPs distribution amongst different applications
- Protection against rouge applications overloading storage controller
- Bandwidth distribution amongst large number of virtual machines

Figure 4-8 on page 118 shows the benefits of throttling.

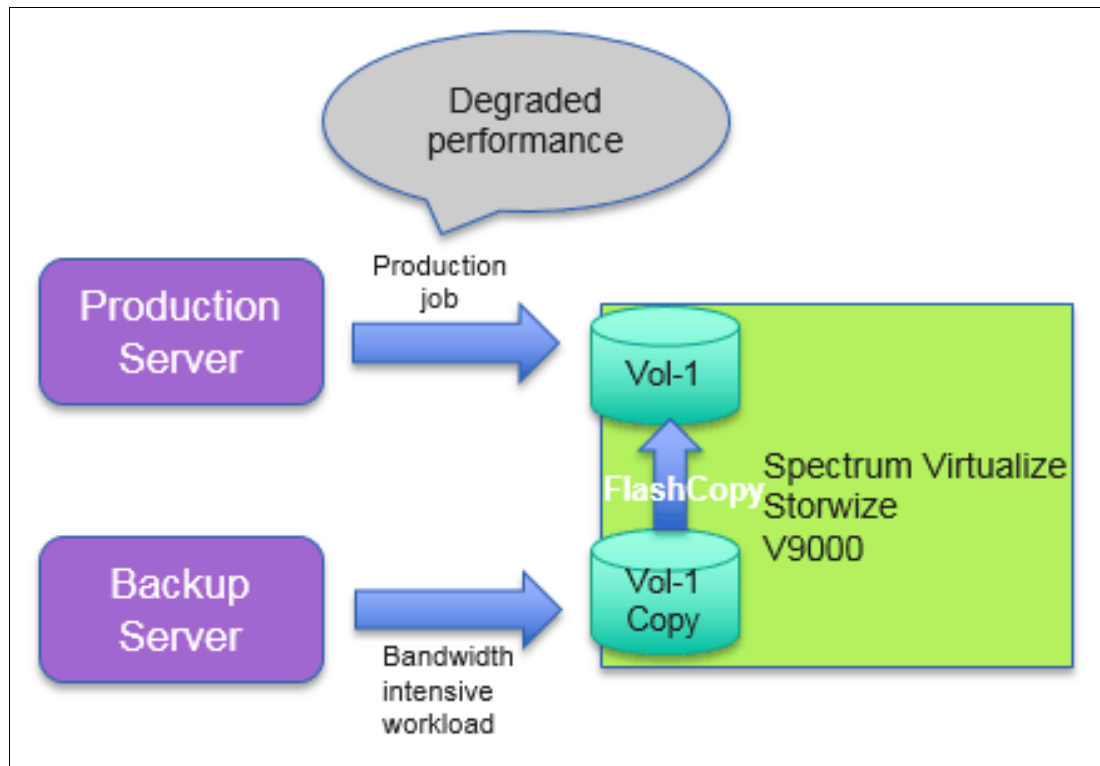


Figure 4-8 Benefits of throttling

4.6 Cache mode and cache-disabled volumes

Cache in 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 parameters is self-explanatory.

By default, when a Volume has been created the cache is set to *readwrite*.

You use cache-disabled (*none*) volumes primarily when you are virtualizing an existing storage infrastructure and you want to retain the existing storage system copy services. 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.

4.6.1 Underlying controller remote copy with 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 the Spectrum Virtualize and Storwize as image mode disks. 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 Spectrum Virtualize and Storwize. You can use 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 Spectrum Virtualize and Storwize copy services with the volume at the secondary site because the Spectrum Virtualize and Storwize does not detect the data that is flowing to this LUN through the controller.

Figure 4-9 shows the relationships between the Spectrum Virtualize and Storwize, the volume, and the underlying storage controller for a cache-disabled volume.

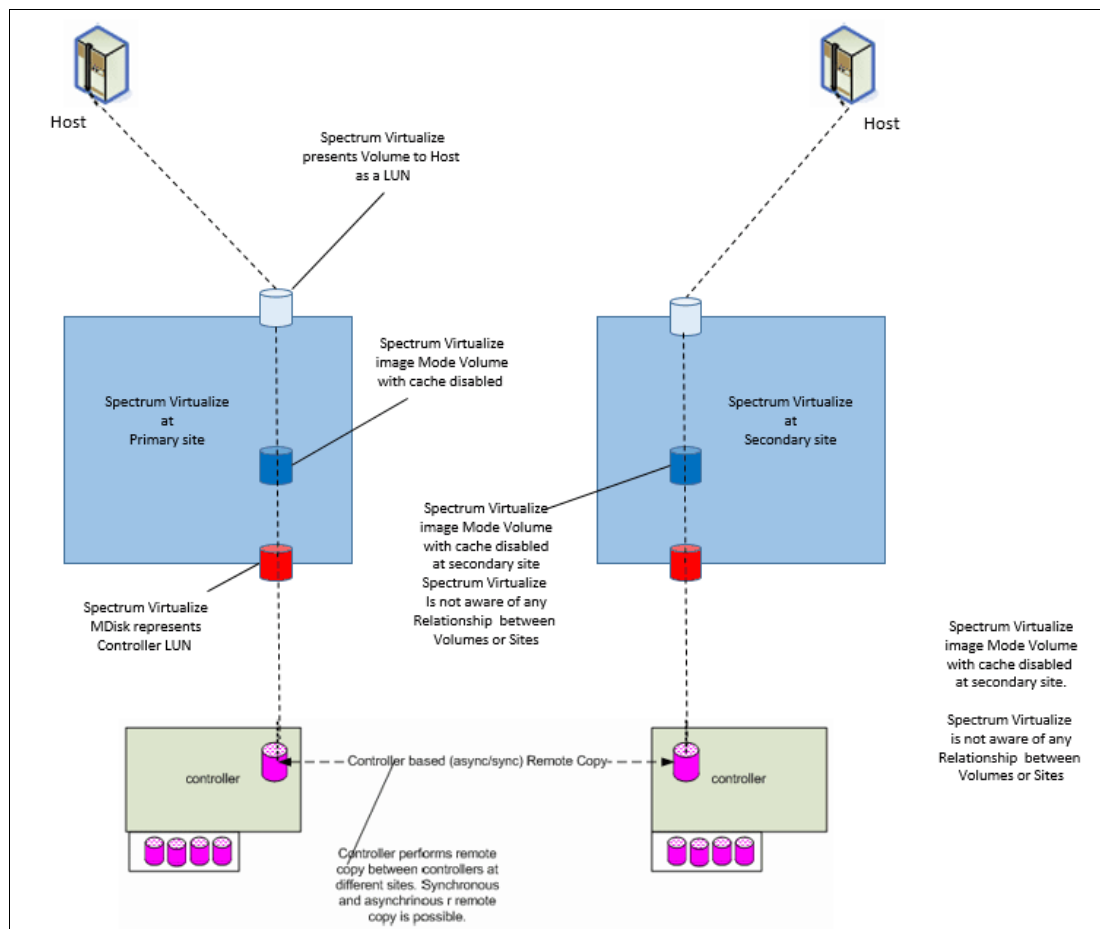


Figure 4-9 Cache-disabled volume in a remote copy relationship

4.6.2 Using underlying controller FlashCopy with 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 Spectrum Virtualize and Storwize as image mode disks, as shown in Figure 4-10. 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 the Spectrum Virtualize and Storwize.

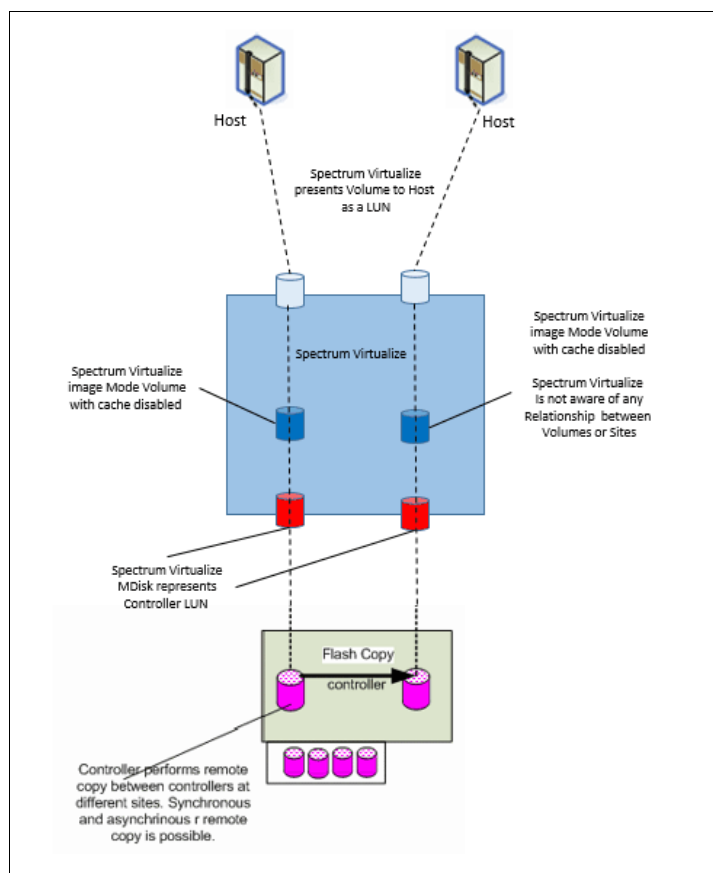


Figure 4-10 FlashCopy with cache-disabled volumes

4.6.3 Changing the cache mode of a volume

The cache mode of a volume can be concurrently (with I/O) changed by using the **svctask chvdisk** command. This command must 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 must not result in a corrupted volume. Therefore, the cache must be flush and discard cache data if the user disables the cache on a volume.

Example 4-8 on page 120 shows an image volume VDISK_IMAGE_1 that changed the cache parameter after it was created.

Example 4-8 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 the cache mode enabled (read/write), but you can specify the cache mode when the volume is created by using the **-cache** option.

4.7 Using Spectrum Virtualize or Storwize with FlashSystem

There can be some specific scenarios where you want to virtualize IBM or OEM All Flash Storage Subsystem (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 will be configured.

In this scenario it is suggested to perform some optimization acting on the Spectrum Virtualize or Storwize Volumes cache depending on the infrastructure you are building.

Figure 4-11 on page 122 shows write operation behavior when volume cache is activated (*readwrite*).

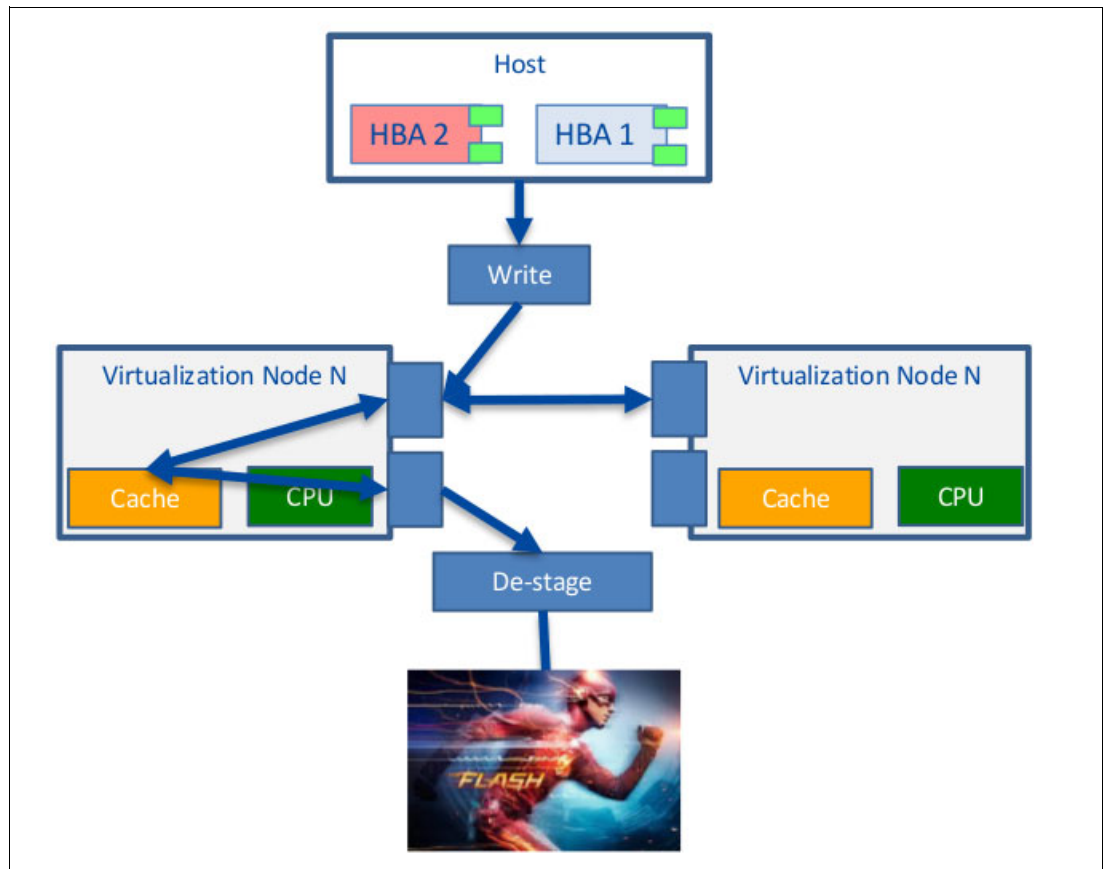


Figure 4-11 Cache activated

Figure 4-12 on page 123 shows a write operation behavior when volume cache is deactivated (*none*).

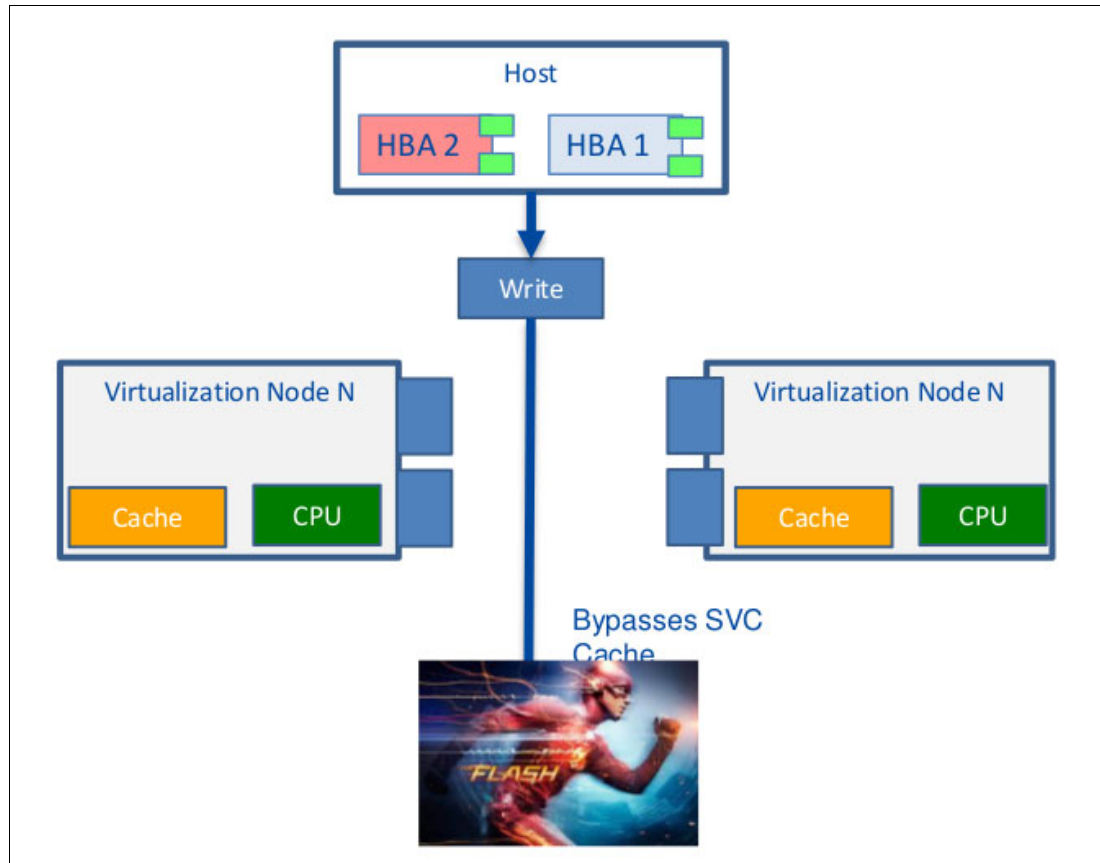


Figure 4-12 Cache deactivated

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 there are extremely high IOPS required, disabling the cache may help.

Attention: Carefully evaluate the impact to the entire ecosystem with quantitative analysis before and after making this change

4.8 FlashCopy services

Here is a short list of rules to apply when you implement Spectrum Virtualize or Storwize FlashCopy services.

4.8.1 FlashCopy rules summary

You must comply with the following rules for using FlashCopy:

- FlashCopy services can be provided only inside a SAN Volume Controller cluster. If you want to use FlashCopy for remote storage, you must define the remote storage locally to the SAN Volume Controller cluster.

- ▶ To maintain data integrity, ensure that all application I/Os and host I/Os are flushed from any application and operating system buffers.
- ▶ You might need to stop your application for it to be restarted with a copy of the volume that you make. Check with your application vendor if you have any doubts.
- ▶ Be careful if you want to map the target flash-copied volume to the same host that has the source volume mapped to it. Check that your operating system supports this configuration.
- ▶ The target volume must be the same size as the source volume. However, the target volume can be a different type (image, striped, or sequential mode) or have different cache settings (cache-enabled or cache-disabled).
- ▶ If you stop a FlashCopy mapping or a consistency group before it is completed, you lose access to the target volumes, and if the target volumes are mapped to hosts, they will have I/O errors.
- ▶ A volume can be the source for up to 256 targets.
- ▶ You can create a FlashCopy mapping by using a target volume that is part of a remote copy relationship. This way, you can use the reverse feature with a disaster recovery implementation. You can also use fast failback from a consistent copy that is held on a FlashCopy target volume at the auxiliary cluster to the master copy.

4.8.2 IBM Spectrum Protect Snapshot

The management of many large FlashCopy relationships and consistency groups is a complex task without a form of automation for assistance. IBM Spectrum Protect™ Snapshot provides integration between the Spectrum Virtualize or Storwize and Spectrum Protect for Advanced Copy Services. It provides application-aware backup and restore by using the Spectrum Virtualize or Storwize FlashCopy features and function.

For more information about IBM Spectrum Protect Snapshot, refer to:

<http://www.ibm.com/software/products/en/spectrum-protect-snapshot>

4.8.3 IBM System Storage Support for Microsoft Volume Shadow Copy Service

The Spectrum Virtualize and Storwize provides support for the Microsoft Volume Shadow Copy Service and Virtual Disk Service. The Microsoft Volume Shadow Copy Service can provide a point-in-time (shadow) copy of a Windows host volume when the volume is mounted and files are in use. The Microsoft Virtual Disk Service provides a single vendor and technology-neutral interface for managing block storage virtualization, whether done by operating system software, RAID storage hardware, or other storage virtualization engines.

The following components are used to provide support for the service:

- ▶ SAN Volume Controller
- ▶ The cluster Common Information Model (CIM) server
- ▶ IBM System Storage hardware provider, which is known as the *IBM System Storage Support*, for Microsoft Volume Shadow Copy Service and Virtual Disk Service software
- ▶ Microsoft Volume Shadow Copy Service
- ▶ The VMware vSphere Web Services when it is in a VMware virtual platform

The IBM System Storage hardware provider is installed on the Windows host. To provide the point-in-time shadow copy, the components complete the following process:

1. A backup application on the Windows host starts a snapshot backup.

2. The Volume Shadow Copy Service notifies the IBM System Storage hardware provider that a copy is needed.
3. The Spectrum Virtualize and Storwize prepares the volumes for a snapshot.
4. The Volume Shadow Copy Service quiesces the software applications that are writing data on the host and flushes file system buffers to prepare for the copy.
5. The Spectrum Virtualize and Storwize creates the shadow copy by using the FlashCopy Copy Service.
6. The Volume Shadow Copy Service notifies the writing applications that I/O operations can resume and notifies the backup application that the backup was successful.

The Volume Shadow Copy Service maintains a free pool of volumes for use as a FlashCopy target and a reserved pool of volumes. These pools are implemented as virtual host systems on the SAN Volume Controller.

For more information about how to implement and work with IBM System Storage Support for Microsoft Volume Shadow Copy Service, see:

<https://ibm.biz/BdsBqY>

4.9 Configuration Backup

To achieve the most benefit from a Spectrum Virtualize or Storwize systems implementation, post-installation planning must include several important steps. These steps ensure that your infrastructure can be recovered with either the same or a different configuration in one of the surviving sites with minimal impact for the client applications. Correct planning and configuration backup also help to minimize possible downtime.

Regardless of which failure scenario you face, apply the following guidelines.

To plan the Spectrum Virtualize or Storwize configuration backup, complete these steps:

1. Collect a detailed Spectrum Virtualize or Storwize configuration. To do so, run a daily configuration backup with the command-line interface (CLI) commands that are shown in Example 4-9. The configuration back up can be automated with your own script.

Example 4-9 Saving the Storwize V7000 configuration

```
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>svcconfig backup
.....
.....
CMMVC6155I SVCCONFIG processing completed successfully
IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsdumps
id filename
0 reinst.7836494-1.trc
1 svc.config.cron.bak_7836494-2
.
.lines removed for brevity
.40 svc.config.backup.xml_7836494-1
```

2. Save the .xml file that is produced in a safe place, as shown in Example 4-10.

Example 4-10 Copying the configuration

```
C:\Program Files\PuTTY>pscp -load V7K_HyperSwap
admin@10.17.89.251:/tmp/SVC.config.backup.xml_7836494-1
c:\temp\configbackup.xml
configbackup.xml | 97 kB | 97.2 kB/s | ETA: 00:00:00 | 100%
```

3. Save the output of the CLI commands that is shown in Example 4-11 in .txt format.

Example 4-11 List of Storwize V7000 commands to issue

```
lsystem
lsite
lsnodecanister
lsnodecanister <nodes name>
lsnodecanisterhw <nodes name>
lsiogrp
lsiogrp <iogrps name>
lscontroller
lscontroller <controllers name>
lsmdiskgrp
lsmdiskgrp <mdiskgrps name>
lsmdisk
lsquorum
lsquorum <quorum id>
lsvdisk
lshost
lshost <host name>
lshostvdiskmap
lsrcrelationship
lsrconsistgrp
```

Note: Example 4-11 on page 126 represent Storwize Systems but same commands can be applied to a Spectrum virtualize.

From the output of these commands and the .xml file, you have a complete picture of the Storwize V7000 HyperSwap infrastructure. Remember the Storwize V7000 HyperSwap ports' worldwide node names (WWNNs), so that you can reuse them during the recovery operation.

Example 4-12, which is contained in the .xml file, shows what you need to re-create a Storwize V7000 environment after a critical event.

Example 4-12 XML configuration file example

```
<xml>
  label="Configuration Back-up"
  version="750"
  file_version="1.206.9.169"
  timestamp="2015/08/12 13:20:30 PDT" >

  <!-- cluster section -->

  <object type="cluster" >
    <property name="id" value="00000100216001E0" />
    <property name="name" value="ITS0_V7K_HyperSwap" />
```

```

</object >
.
many lines omitted for brevity
.
<!-- controller section -->

<object type="controller" >
  <property name="id" value="0" />
  <property name="controller_name" value="ITS0_V7K_Q_N1" />
  <property name="WWNN" value="5005076802000EF" />
  <property name="mdisk_link_count" value="2" />
  <property name="max_mdisk_link_count" value="2" />
  <property name="degraded" value="no" />
  <property name="vendor_id" value="IBM      " />
  <property name="product_id_low" value="2145    " />
  <property name="product_id_high" value="        " />
  <property name="product_revision" value="0000" />
  <property name="ctrl_s/n" value="2076          " />
  <property name="allow_quorum" value="yes" />
  <property name="fabric_type" value="fc" />
  <property name="site_id" value="3" />
  <property name="site_name" value="ITS0_SITE_Q" />
  <property name="WWPN" value="50050768021000EF" />
  <property name="path_count" value="0" />
  <property name="max_path_count" value="0" />
  <property name="WWPN" value="50050768022000EF" />
  <property name="path_count" value="0" />
  <property name="max_path_count" value="0" />
</object >
<object type="controller" >
  <property name="id" value="1" />
  <property name="controller_name" value="ITS0_V7K_Q_N2" />
  <property name="WWNN" value="5005076802000F0" />
  <property name="mdisk_link_count" value="2" />
  <property name="max_mdisk_link_count" value="2" />
  <property name="degraded" value="no" />
  <property name="vendor_id" value="IBM      " />
  <property name="product_id_low" value="2145    " />
  <property name="product_id_high" value="        " />
  <property name="product_revision" value="0000" />
  <property name="ctrl_s/n" value="2076          " />
  <property name="allow_quorum" value="yes" />
  <property name="fabric_type" value="fc" />
  <property name="site_id" value="3" />
  <property name="site_name" value="ITS0_SITE_Q" />
  <property name="WWPN" value="50050768021000F0" />
  <property name="path_count" value="8" />
  <property name="max_path_count" value="8" />
  <property name="WWPN" value="50050768022000F0" />
  <property name="path_count" value="8" />
  <property name="max_path_count" value="8" />
</object >
many lines omitted for brevity

```

You can also get this information from the .txt command output that is shown in Example 4-13.

Example 4-13 lsnodcanister example output command

```

IBM_Storwize:ITS0_V7K_HyperSwap:superuser>lsnodcanister ITS0_HS_SITE_A_N1
id 8
name ITS0_HS_SITE_A_N1
UPS_serial_number
WWNN 500507680B0021A8
status online
IO_group_id 0
IO_group_name io_grp0_SITE_A
partner_node_id 9
partner_node_name ITS0_HS_SITE_A_N2
config_node yes
UPS_unique_id
port_id 500507680B2121A8
port_status active
port_speed 4Gb
port_id 500507680B2221A8
port_status active
port_speed 4Gb
port_id 500507680B2321A8
port_status active
port_speed 2Gb
port_id 500507680B2421A8
port_status active
port_speed 2Gb
hardware 400
iscsi_name iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsohssitean1
iscsi_alias
failover_active no
failover_name ITS0_HS_SITE_A_N2
failover_iscsi_name iqn.1986-03.com.ibm:2145.itsov7khyperswap.itsohssitean2
failover_iscsi_alias
panel_name 01-1
enclosure_id 1
canister_id 1
enclosure_serial_number 7836494
service_IP_address 10.18.228.55
service_gateway 10.18.228.1
service_subnet_mask 255.255.255.0
service_IP_address_6
service_gateway_6
service_prefix_6
service_IP_mode static
service_IP_mode_6
site_id 1
site_name ITS0_SITE_A
identify_LED off
product_mtm 2076-424
code_level 7.5.0.2 (build 115.51.1507081154000)

```

For more information about backing up your configuration, see the IBM Storwize V7000 Knowledge Center:

https://www.ibm.com/support/knowledgecenter/ST3FR7_7.8.0/com.ibm.storwize.v7000.780.doc/svc_configbackupovr_1e4imh.html

and

https://www.ibm.com/support/knowledgecenter/ST3FR7_7.8.0/com.ibm.storwize.v7000.780.doc/svc_clustconfbackuptsk_1e4k69.html

and for Spectrum Virtualize:

https://www.ibm.com/support/knowledgecenter/STVLF4_7.8.0/spectrum.virtualize.780.doc/svc_clustconfbackuptsk_1e4k69.html

4. Create an up-to-date, high-level copy of your configuration that describes all elements and connections.
5. Create a standard labeling schema and naming convention for your Fibre Channel (FC) or Ethernet (ETH) cabling, and ensure that it is fully documented.
6. Back up your storage area network (SAN) zoning by using your FC switch CLI or graphical user interface (GUI).

The essential zoning configuration data, domain ID, zoning, alias, configuration, and zone set can be saved in a .txt file by using the output from the CLI commands. You can also use the appropriate utility to back up the entire configuration.

The following IBM b-type/Brocade FC switch or director commands are helpful to collect the essential zoning configuration data:

- switchshow
- fabricshow
- cfgshow

During the implementation, use WWNN zoning. During the recovery phase after a critical event, reuse the same domain ID and same port number that were used in the failing site, if possible. Zoning is propagated on each switch because of the SAN extension with inter-switch link (ISL).

For more information about how to back up your FC switch or director zoning configuration, see your switch vendor's documentation.

7. Back up your back-end storage subsystems configuration.

In your Spectrum Virtualize or Storwize System implementation, potentially you can also virtualize the external storage controller.

If you virtualized the external storage controller, back up your storage subsystem configuration. This way, if a critical event occurs, you can re-create the same environment when you reestablish your infrastructure in a different site with new storage subsystems.

Back up your storage subsystem in one of the following ways:

- For the IBM DS8000 storage subsystem, save the output of the DS8000 CLI commands in .txt format, as shown in Example 4-14.

Example 4-14 DS8000 commands

```
lsarraysite -l
lsarray -l
lsrank -l
lsextpool -l
lsfbvol -l
lshostconnect -l
lsvolgrp -l
showvolgrp -lunmap <SVC vg_name>
```

- For the IBM XIV Storage System, save the output of the XCLI commands in .txt format, as shown in Example 4-15.

Example 4-15 XIV subsystem commands

```
host_list
host_list_ports
mapping_list
vol_mapping_list
pool_list
vol_list
```

- For IBM Storwize V7000, collect the configuration files and the output report as described previously.
- For any other supported storage vendor's products, refer to their documentation.



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 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
- ▶ IP Replication
- ▶ Volume Mirroring

5.1 Introduction to copy services

Spectrum Virtualize and Storwize family products offer a complete set of copy services functions which provide capabilities for Disaster Recovery, Business Continuity, data movement and data duplication solutions.

FlashCopy

FlashCopy is a function that allows you to create a point-in-time copy of one of your volumes. This might be helpful when performing backups or application testing. These copies may be cascaded upon 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.

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 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 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 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.

Global Mirror with Change Volumes

This function (also known as Cycle-Mode Global Mirror), introduced in SV V6.3, can best be described as “Continuous Remote FlashCopy.” If you use this feature, the system will essentially take 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.

Volume Mirroring function

Volume Mirroring is a function 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 may 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. In the event of 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 may 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. In this section, we describe the inner workings of FlashCopy, and provide some best practices and recommendation 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 frozen 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), which tracks changes to the data after the FlashCopy is started, and an indirection layer, which 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 some of the most common use cases is reported.

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; 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, you must 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, and 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 need to make the target available on a host. We suggest that you 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 using 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.

You 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.

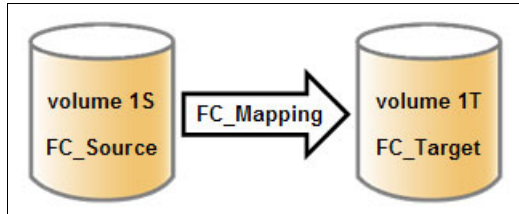
5.2.2 Flash Copy capabilities overview

FlashCopy occurs between a source volume and a target volume in the same storage system. The minimum granularity that 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 initiate 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.

Table 5-1 shows the concept of FlashCopy mapping.



A FlashCopy mapping has a set of attributes and settings which define the characteristics as well as the capabilities of the FlashCopy.

These characteristics are explained more in 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 will be 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 only remains a valid copy of the source data while the FlashCopy mapping remains in place.

Valid values for the background copy rate are between 0 and 100. 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

Value	Data copied per second
71 - 80	16 MB
81 - 90	32 MB
91 - 100	64 MB

FlashCopy Consistency Groups

Consistency Groups can be used to help create a consistent point-in-time copy across multiple volumes, thus managing the consistency of dependent writes that are executed in the application following the correct sequence.

When Consistency Groups are used, the FlashCopy commands are issued to the Consistency Groups, which perform the operation on all FlashCopy mappings contained within the Consistency Groups at the same time.

Figure 5-1 illustrates a Consistency Group consisting of two volume mappings.

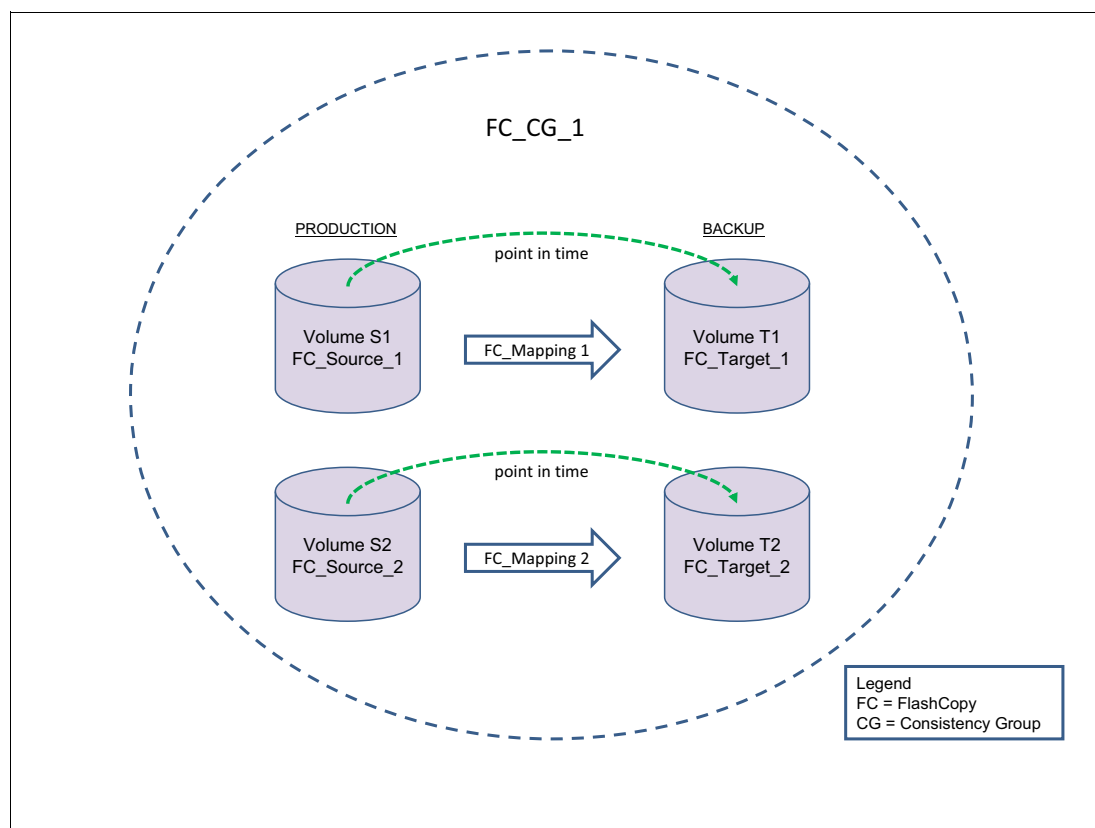


Figure 5-1 Multiple volume mapping in a Consistency Group

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 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 has to be copied, the workload put on the system and the back-end storage has been 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 5-2 on page 137.

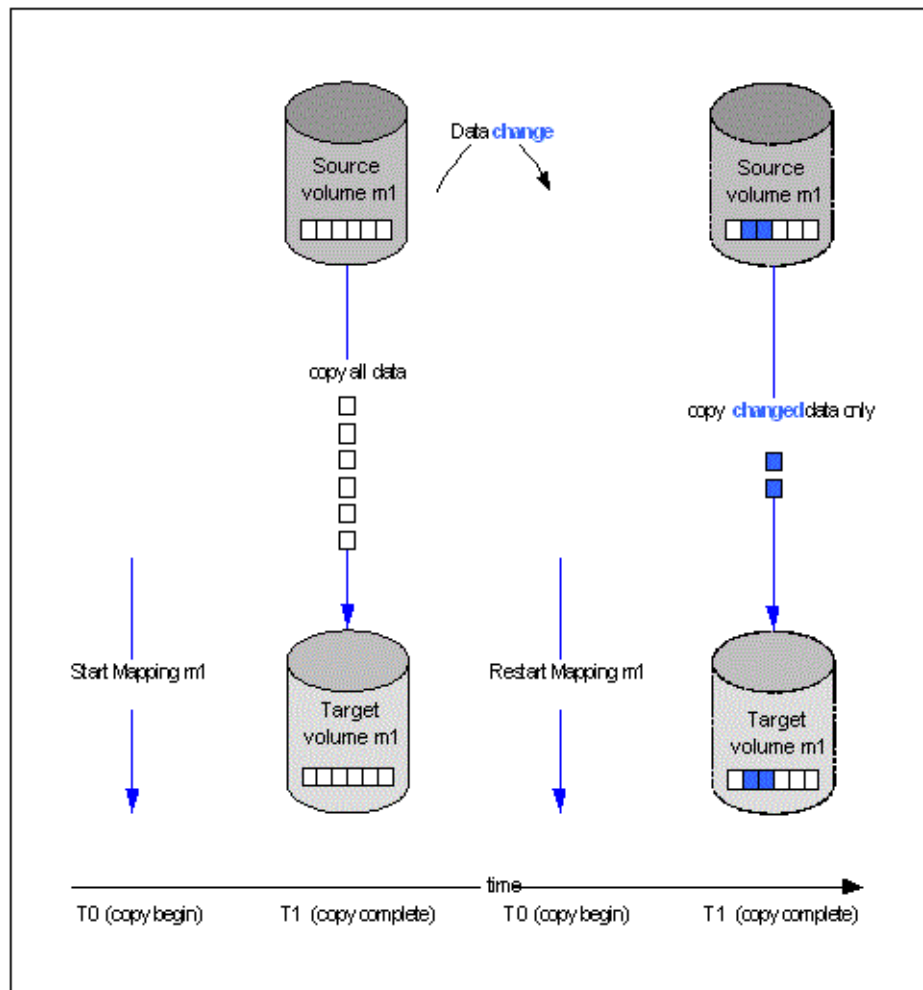


Figure 5-2 Incremental FlashCopy

If the FlashCopy mapping was stopped before the background copy had completed, then when the mapping is restarted, the data that had been 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 can 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 have been changed, and it 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 5-3.

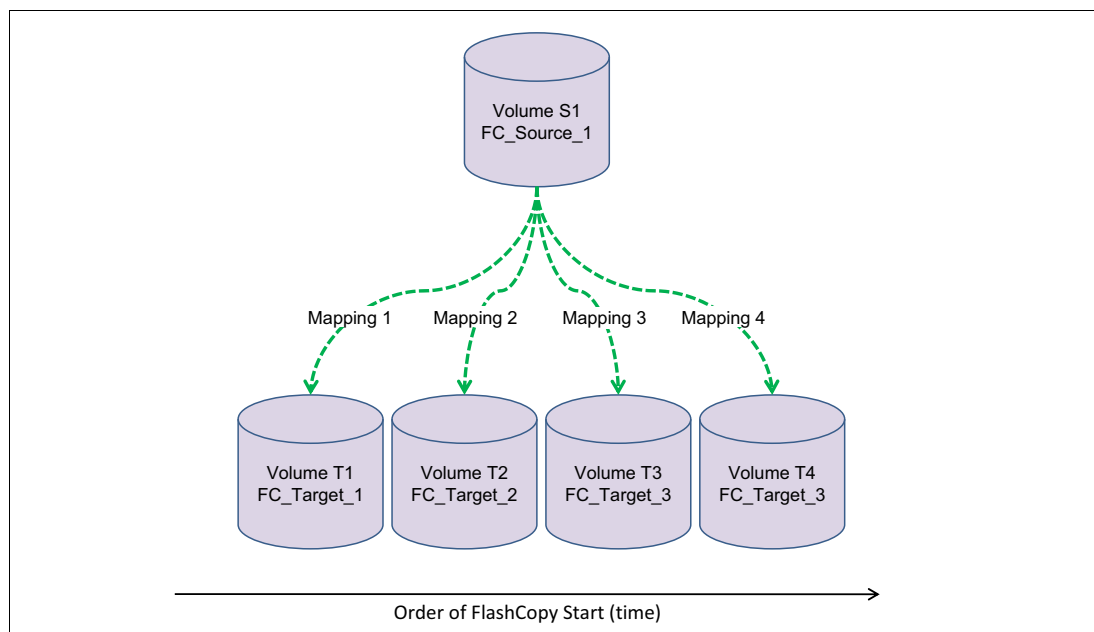


Figure 5-3 Multiple Target FlashCopy

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 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 5-4.

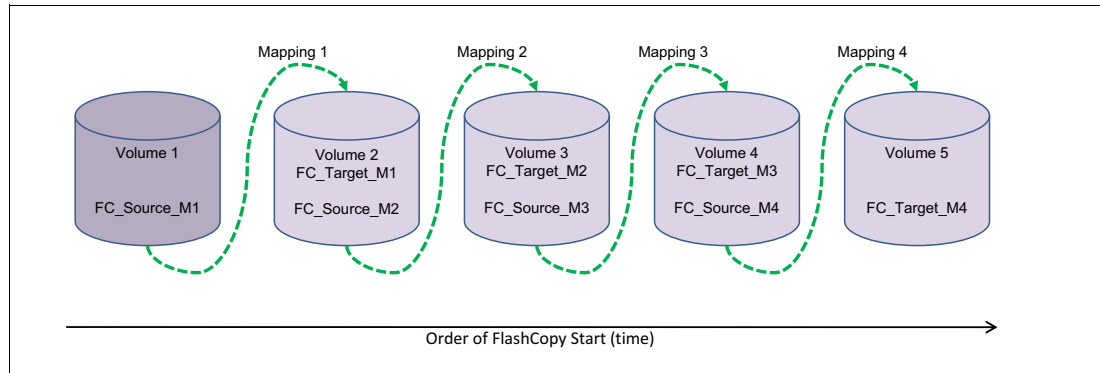


Figure 5-4 Cascaded FlashCopy

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. In a thin-provisioned volume, however, 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 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 and 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 will be registered in the event log when you use 80 percent of the total capacity. This 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 time, 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 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. It supports multiple targets (up to 256) and therefore multiple rollback points.

A key advantage of the Multiple Target Reverse FlashCopy function is that the reverse FlashCopy does not destroy the original target, which enables processes that are using the target, such as a tape backup, to continue uninterrupted.

Spectrum Virtualize and Storwize family systems also provides the ability 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 Flash Copy 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 have 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 which has to be shared with other features's 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, which is done 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 the FlashCopy has been started.

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 using 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; 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. 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.

Volume being accessed	Has the grain been copied?	Host I/O operation	
		Read	Write
Target	No	If any newer targets exists 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 if 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 accordingly. 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 5-5 shows the IBM Spectrum Virtualize cache architecture.

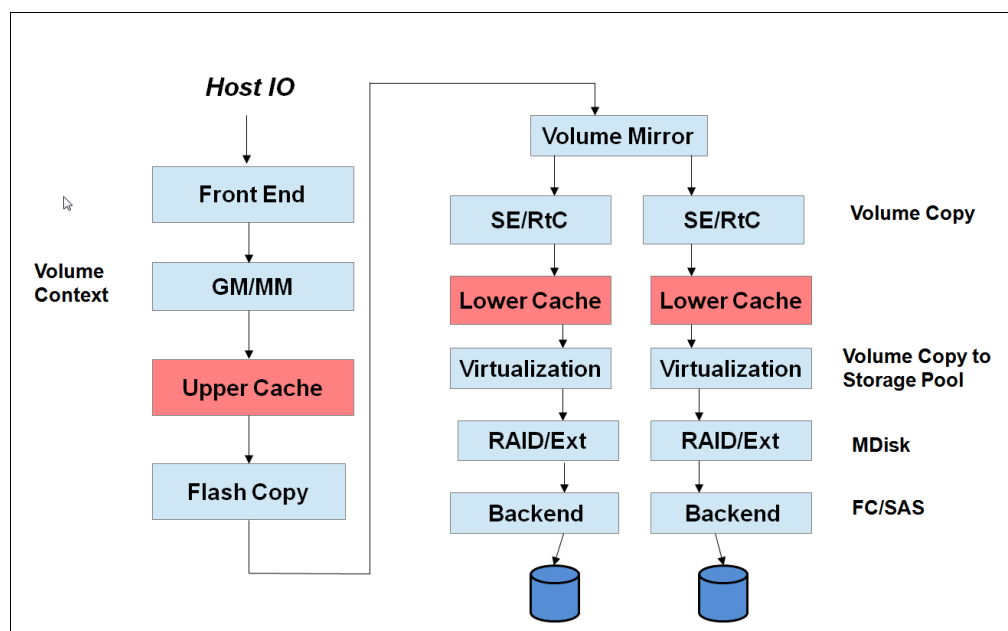


Figure 5-5 New cache architecture

This copy-on-write process introduces significant latency into write operations so to isolate the active application from this additional latency, the FlashCopy indirection layer is placed logically between 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 5-6.

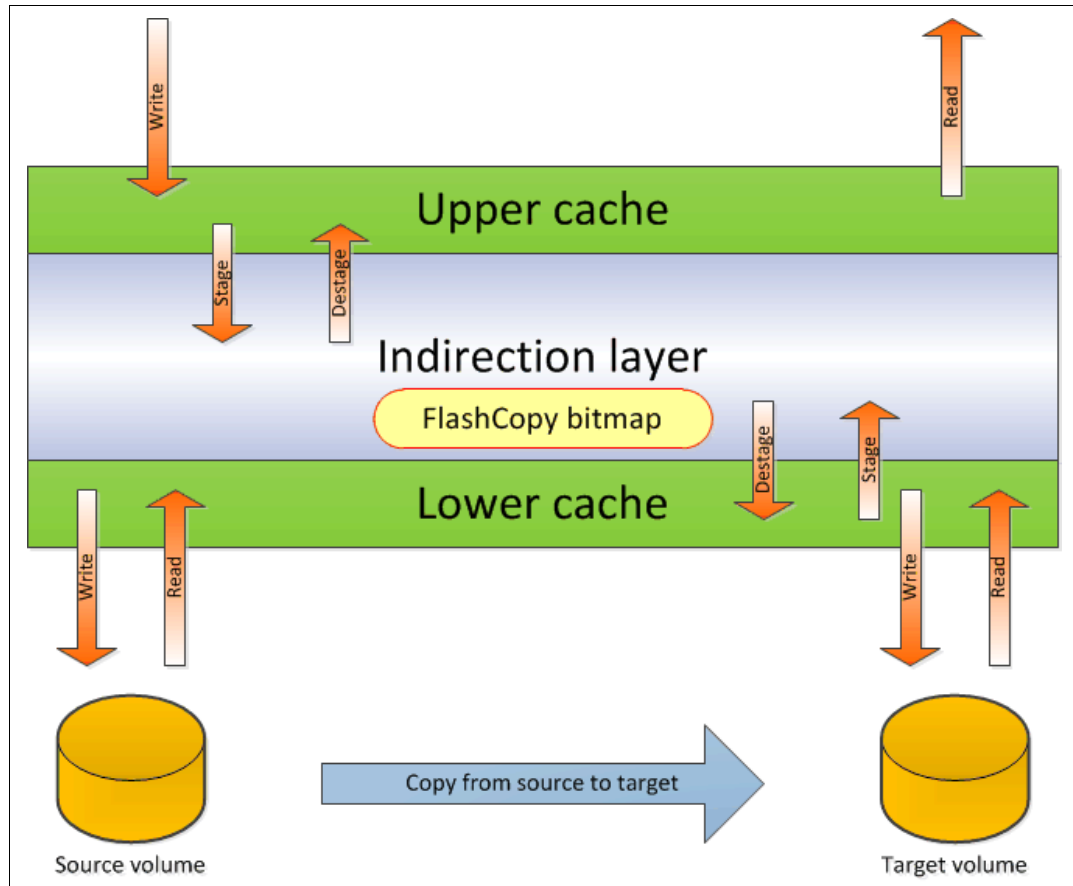


Figure 5-6 Logical placement of the FlashCopy indirection layer

Introduction of the two-level cache provides additional performance improvements to the FlashCopy mechanism. Because now the FlashCopy layer is above lower cache in the IBM Spectrum Virtualize software stack, it can benefit from read pre-fetching and coalescing writes to backend storage. Also, preparing FlashCopy is much faster, because upper cache write data does not have to go directly to backend storage but to lower cache layer.

Additionally, in the 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 5-7 on page 144 represents a set of four FlashCopy mappings that share a common source. The FlashCopy mappings will target volumes Target 0, Target 1, Target 2, and Target 3.

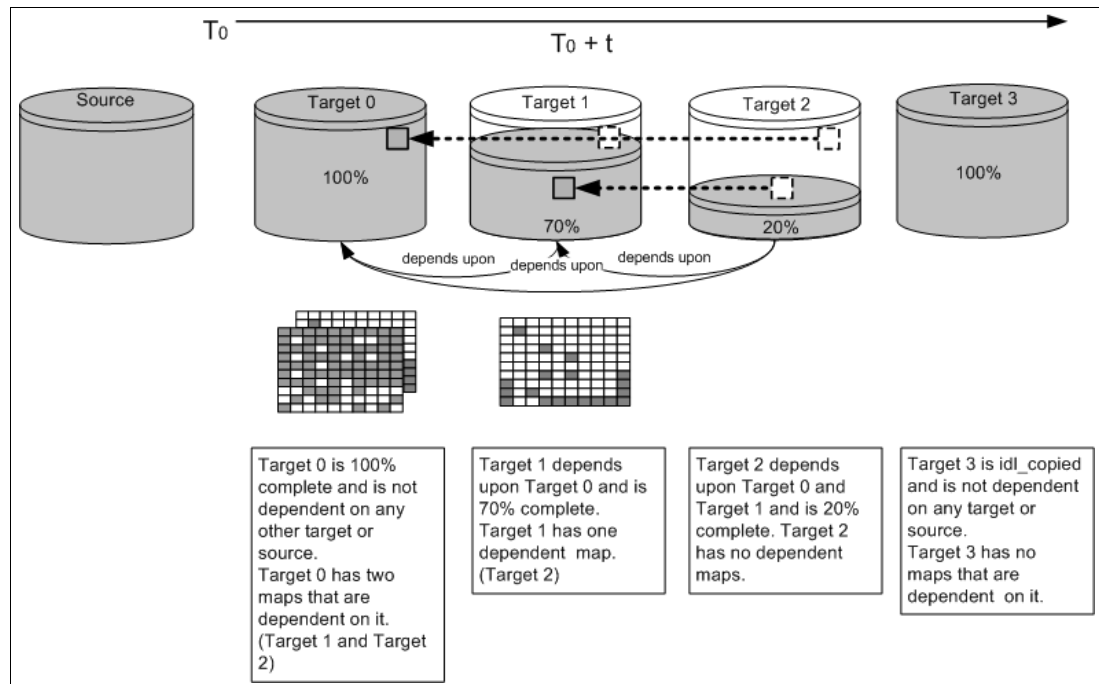


Figure 5-7 Interactions between multi target FlashCopy mappings

- ▶ Target 0 is not dependent on a source, because it has completed copying. Target 0 has two dependent mappings (Target 1 and Target 2).
- ▶ Target 1 is dependent upon Target 0. It will remain dependent until all of Target 1 has been copied. Target 2 is dependent on it, because Target 2 is 20% copy complete. After all of Target 1 has been copied, it can then move to the idle_copied state.
- ▶ Target 2 is dependent upon Target 0 and Target 1 and will remain dependent until all of Target 2 has been copied. No target is dependent on Target 2, so when all of the data has been copied to Target 2, it can move to the idle_copied state.
- ▶ Target 3 has actually completed copying, so it is not dependent on any other maps.

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 has been 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 Spectrum Virtualize and Storwize family products features, offers very useful capabilities but 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 limits for a system 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 applies to: ▶ 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	255	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 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 a large number of mappings.

Configuration Limits: the configuration limits always change with the introduction of new HW and SW capabilities. Please check in the 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 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 is 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 using the **chlogrp** 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 allocated to the FlashCopy bitmaps and no more free memory is available in the I/O group.

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 using the same target volume can be defined, but only one of these mapping can be started at a time. This 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 remove first the FlashCopy mapping
- ▶ In a cascading FlashCopy the grain size of all the FlashCopy mapping participating must be the same
- ▶ In a multi target FlashCopy the grain size of all the FlashCopy mapping participating must be the same
- ▶ In a reverse FlashCopy the grain size of all the FlashCopy mapping participating 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 invoking FlashCopy mappings at the same point in time. Within the *same* Consistency Group, it is not possible to have mappings where:
 - 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, which 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 (that is, 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. However if the target volume of a FlashCopy is used as primary in a Remote Copy relationship, the following rules apply:
 - The FlashCopy can not be started if the status of the Remote Copy relationship is different to *Idle* or *Stopped*.
 - The FlashCopy can not be started if the I/O group allocating the FlashCopy mapping bitmap is not the same as the FlashCopy target volume
- ▶ A FlashCopy can not be started if the target volume is the secondary volume of a Remote Copy relationship.

FlashCopy presets

The Spectrum Virtualize/Storwize GUI interface provides three FlashCopy presets (Snapshot, Clone, and Backup) to simplify the more common FlashCopy operations.

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.

In this section, we describe 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 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, but 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 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, as in the case of 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 (and 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 that 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 are identified using the bitmaps, and then the amount of data to be copied when refreshing the mapping depends on the grain size, as well. 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 the incremental FlashCopy the 64 KB grain size is recommended

Similarly to the FlashCopy, also the Thin Provisioned volumes have a grain size attribute that represent the size of chunk of storage to be added to used capacity.

The recommendation 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 recommended to provide 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 extremely 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 in order to minimize the impacts 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

For the copy-on-write process, the Table 5-5 on page 149 shows how the operations are distributed across the nodes.

Table 5-5 Workload distribution for the copy-on-write process

	Read from source	Read from target	Write to source	Write to target
Node that perform the backend I/O if the grain is copied	Preferred node in source volume's IO group	Preferred node in target volume's IO group	Preferred node in source volume's IO group	Preferred node in target volume's IO group
Node that perform the backend I/O if the grain is not yet copied	Preferred node in source volume's IO group	Preferred node in source volume's IO group	Preferred node in source volume's IO group will read and write; preferred node in target volume's IO group will write	Preferred node in source volume's IO group will read; preferred node in target volume's IO 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. Let's consider the following volume placement alternatives:

1. Source and target volumes uses the same preferred node.

In this scenario the node 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, but no node-to-node bandwidth is used.

2. Source and target volumes uses the different preferred node.

In this scenario both node acting as preferred nodes manages read and write FlashCopy operations according to the schemes described above. The data transferred between the two preferred nodes goes through the node-to-node network.

Both alternatives described have pros and cons then there is no a general best practice to apply. Anyway we can explore following scenarios:

1. Spectrum Virtualize or Storwize system with multiple I/O groups where the source volumes are evenly spread across all the nodes. Assuming that also the I/O workload is evenly distributed across the nodes, the alternative 1 is preferable. In fact the amount of read and write FlashCopy operations will be again evenly spread across the nodes without using any node-to-node bandwidth.
2. 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 the alternative 2 is recommended. In fact defining the target volumes's preferred node in the less utilized nodes will relieve the source volume's preferred node of some additional FlashCopy workload (especially during the background copy).
3. 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 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 volume's preferred node must be in Site B. Placing the target volume's preferred node on Site A will cause the re-direction of the FlashCopy write operation through the node-to-node network.

Regarding the placement on the backend storage, this is mainly driven by the availability requirements. Anyway it is recommended, if possible, to use different backend storage controllers or arrays for the source and target volumes.

Background Copy considerations

The background copy process makes use of internal resources such as CPU, memory and bandwidth. This copy process will try to reach the target copy data rate for every single volume according to the background copy rate parameter setting (as reported in Table 5-1 on page 135). If, for any reason, 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 of this, both background copy and foreground I/O will tend to see an increase in latency and consequential reduction in throughput with respect to the situation had the bandwidth not been limited. Degradation will be graceful: both background copy and foreground I/O will continue to make forward 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, we need a careful planning of 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, then usually the most affected component is the backend storage. Usually CPU and memory are not significantly impacted 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, will add a maximum throughput of 160 MB/s for the reads and 160 MB/s for the writes. The source and target volumes distribution on the backend 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, that is the total size of volumes in case of full background copy or the amount of data modified in case of incremental copy refresh. Performance monitoring tools like IBM Spectrum Control can be used to evaluate the existing workload on the backend storage in a specific time window and by adding this workload to the foreseen background copy workload we can have an estimation of the overall workload running toward the backend. Disk performance simulation tools, like Disk Magic, can finally be used to estimate the impacts, if any, of the added backend 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, a conservative and progressive approach is recommended. 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 setting the background copy rate to 0.

Then the recommendation is to initially set the background copy rate value in order to add a limited workload to the backend (for example less than 100 MB/s). If no impacts to hosts are noticed, the background copy rate value can be increased and so on. Note that background copy rate setting follows an exponential scale, so changing for instance from 50 to 60 it will double the data rate goal from 2 MB/s to 4 MB/s.

Cleaning rate: The Cleaning Rate is the rate at which the data are copied among dependant FlashCopies such as Cascaded and Multi Target FlashCopy. The cleaning process is a copy process similar to the background copy, so 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 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 unavailable, 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, 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

Spectrum Virtualize and Storwize technology offers a variety of remote copy services functions that addresses Disaster Recovery and Business Continuity needs.

Metro Mirror is designed for metropolitan distances in conjunction with a zero Recovery Point Objective (RPO); that is, zero data loss. This 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 will make the host wait for write I/Os to complete at both the local and remote storage systems and will include round-trip network latencies. Metro Mirror has the following characteristics:

- ▶ Zero RPO
- ▶ Synchronous
- ▶ Production application performance impacted 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 allows Global Mirror to be used over longer distances. By definition, any vendors' asynchronous replication will result in an RPO greater than zero. However, in the case of 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 quite 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 impacted by I/O sequencing preparation time

Global Mirror with Change Volumes provides an option to replicate point-in-time copies of volumes. This will generally require lower bandwidth because it is the average rather than the peak throughput that must be accommodated. The RPO for Global Mirror with Change Volumes will be higher than traditional Global Mirror. Global Mirror with Change Volumes has the following characteristics:

- ▶ Larger RPO
- ▶ Point-in-time copies
- ▶ Asynchronous

- Possible system performance overhead 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 represent 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 5-8 shows some of the concepts of remote copy.

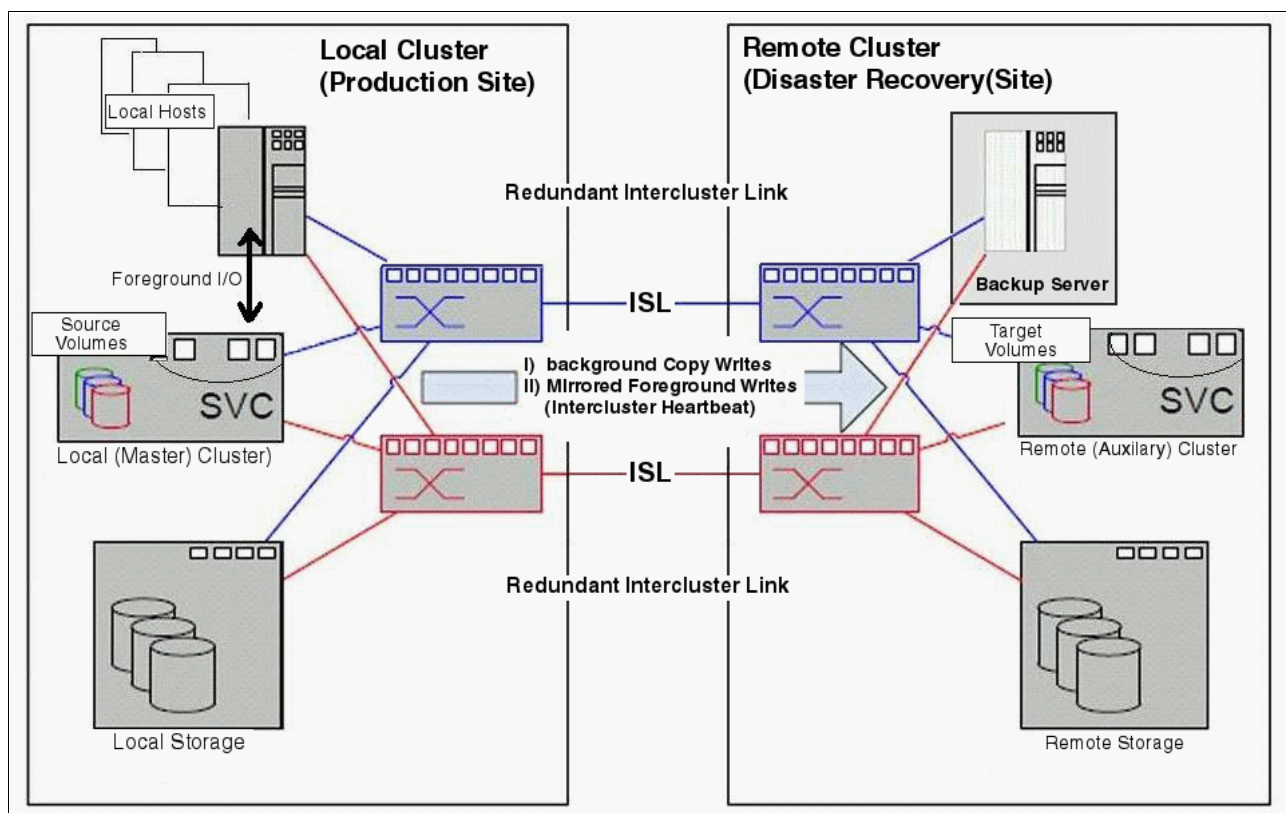


Figure 5-8 Remote copy components and applications

A successful implementation of an intersystem remote copy service depends on 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 5-9.

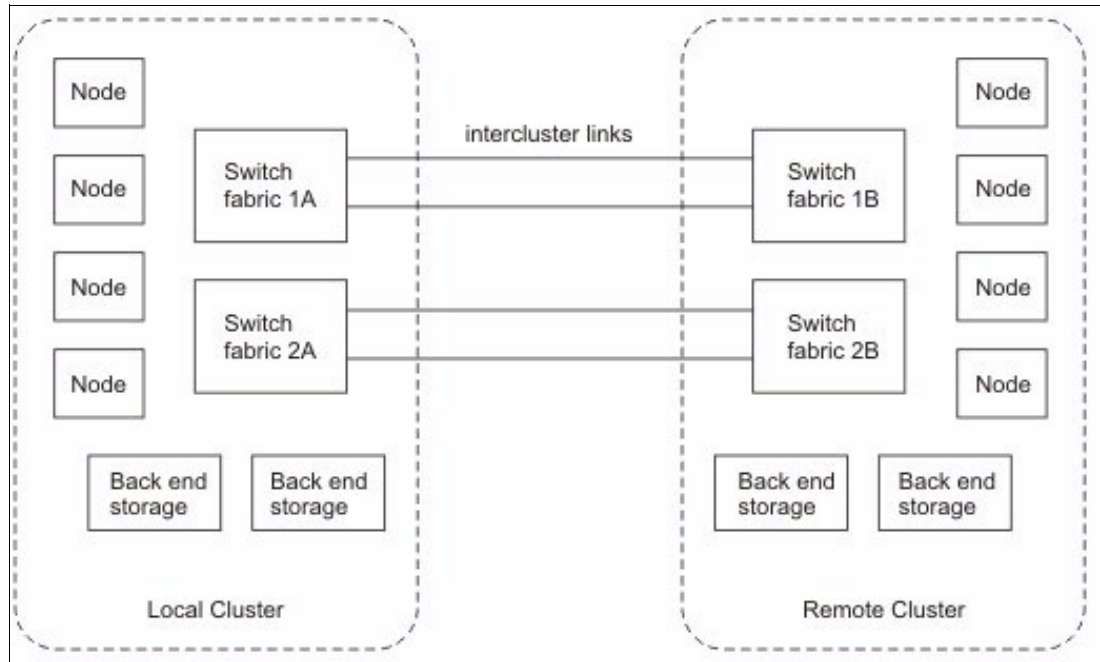


Figure 5-9 Remote copy partnership

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 will determine the maximum speed of the initial synchronization and resynchronization of the relationships.

Tip: To establish a fully functional Metro Mirror or Global Mirror partnership, you must 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 an target (secondary) volume in the remote system. Usually when a remote copy relationship is started a background copy process which copy the data from source to target volumes is started as well.

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 performed from master to auxiliary.

After the initial synchronization is complete, you can change the copy direction (see Figure 5-10 on page 156). The ability to change roles is used to facilitate disaster recovery.

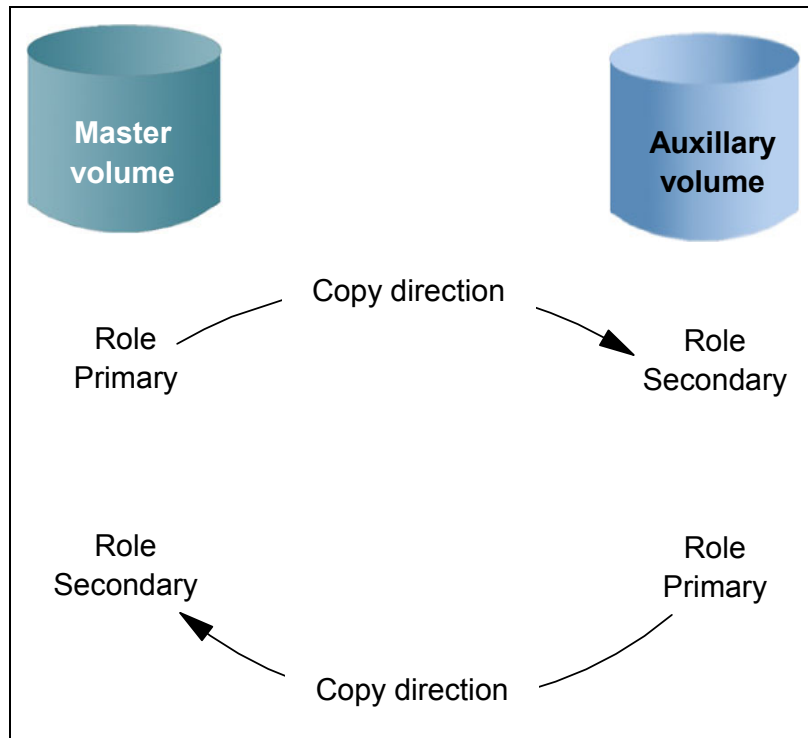


Figure 5-10 Role and direction changes

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, usually 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 **stoprcrelationship** 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 the same I/O groups.
- ▶ All relationships in a Consistency Group will also have the same copy direction and state.

- Each Consistency Group is either for Metro Mirror or for Global Mirror relationships, but not both. This 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 and thereby spread 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 will allocate traffic from each given Consistency Group in a round-robin fashion across 16 streams. That is, cg0 traffic will go into stream0 and cg1 traffic will go into stream1.

Any volume that is *not* in a Consistency Group will also go 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 will also delay the writes behind it in the stream. Having more streams (up to 16) reduces this kind of potential congestion. Also, 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 Spectrum Virtualize nodes/Storwize canisters, and ideally, perfect multiples of the node count.

Layer concept

Version 6.3 introduced the concept of *layers* which allows to create partnerships among Spectrum Virtualize and Storwize products. The key points concerning layers are listed here:

- Spectrum Virtualize is always in *Replication* layer.
- By default, Storwize products is in *Storage* layer.
- A system can only form partnerships with systems in the same layer.
- A Spectrum Virtualize can virtualize a Storwize system only if the Storwize is in Storage layer.
- With version 6.4, a Storwize system in Replication layer can virtualize a Storwize system in Storage layer.

Figure 5-11 illustrates the concept of layers.

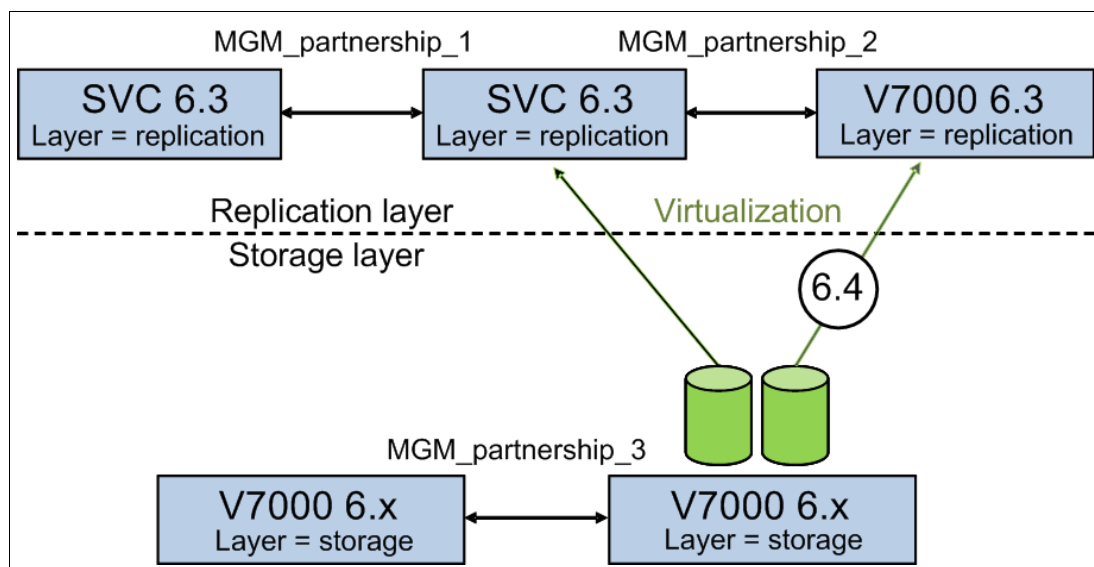


Figure 5-11 Conceptualization of layers

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:

1. The Storwize system must not have any Spectrum Virtualize or Storwize host objects defined, and must not be virtualizing any other Storwize controllers.
2. The Storwize system must not be visible to any other Spectrum Virtualize or Storwize system in the SAN fabric, and this might require SAN zoning changes.
3. 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 using the CLI. After you are certain that all of the pre-conditions have been met, issue the following commands:

```
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 5-12 shows examples of the principal supported topologies for Metro and Global Mirror partnerships. Each numbered box represents a Spectrum Virtualize or Storwize system.

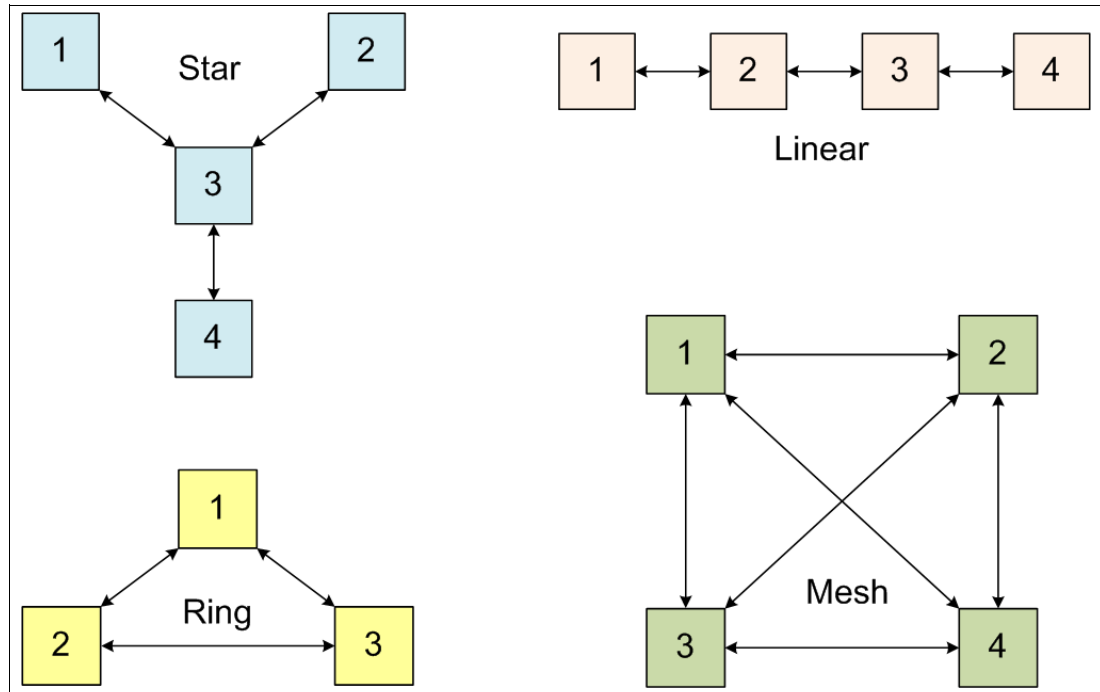


Figure 5-12 Supported topologies for Metro and Global Mirror

Star topology

A star topology can be used, for example, to share a centralized disaster recovery system (3, in our example) with up to three other systems, for example replicating $1 \rightarrow 3$, $2 \rightarrow 3$, and $4 \rightarrow 3$.

Ring topology

A ring topology (3 or more systems) can be used to establish a one-in, one-out implementation. for example, $1 \rightarrow 2$, $2 \rightarrow 3$, $3 \rightarrow 1$ so as to spread replication loads evenly among three systems.

Linear topology

A linear topology of two or more sites is also possible, although 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 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 a variety of topologies is supported, it is advisable to keep your partnerships as simple as possible, which in most cases will mean 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 Spectrum Virtualize/Storwize systems running 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 acknowledgement (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 will hold the same data as the primary did immediately prior to 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 additional latency, due to the network alone, on each I/O if you are running across 300 km separation.

Figure 5-13 shows the order of Metro Mirror write operations.

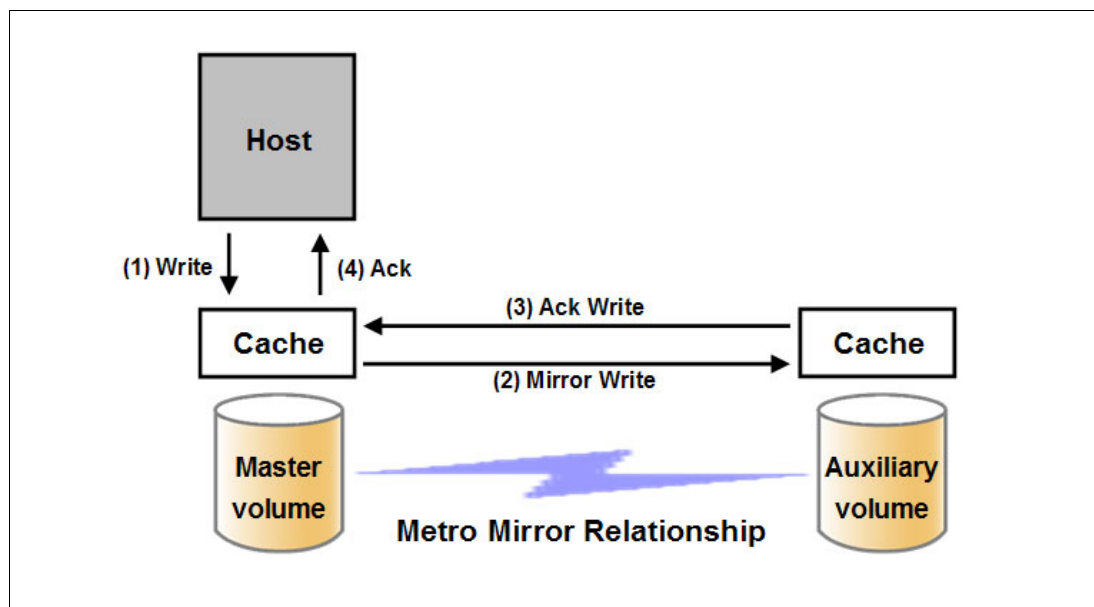


Figure 5-13 Metro Mirror write sequence

Note that a write into mirrored cache on an 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 acknowledgement.

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 will hold the same data as the primary did at a point a short time prior to the failure. That short period of data loss might typically be between 10 milliseconds and 10 seconds, but will vary 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 and with 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. 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 5-14 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.

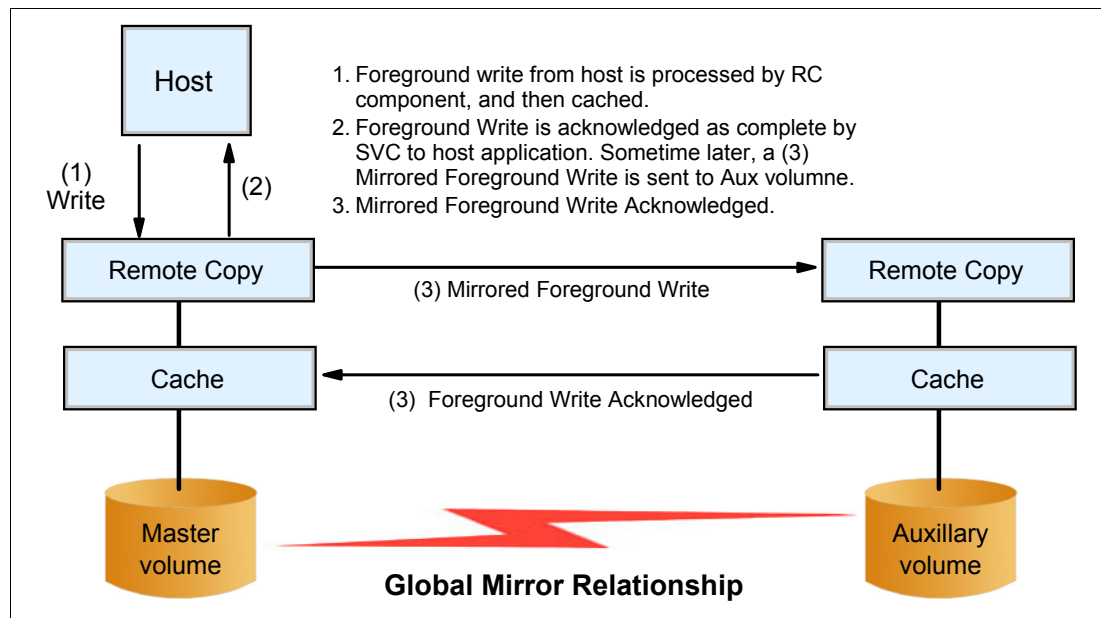


Figure 5-14 Global Mirror relationship write operation

With Global Mirror, a confirmation is sent to the host server before the host receives a confirmation of the completion at the auxiliary volume. When a write is sent to a master volume, it is assigned a sequence number. Mirror writes that are sent to the auxiliary volume are committed in sequential number order. If a write is issued when another write is outstanding, it might be given the same sequence number.

This function maintains a consistent image at the auxiliary volume all times. It identifies sets of I/Os that are active concurrently at the primary volume, assigning an order to those sets, and applying 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, and 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 often is designed with the concept of dependent writes. With dependent writes, ensure that an earlier write completed 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 logical block address (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, and 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 5-15 shows a colliding write sequence.

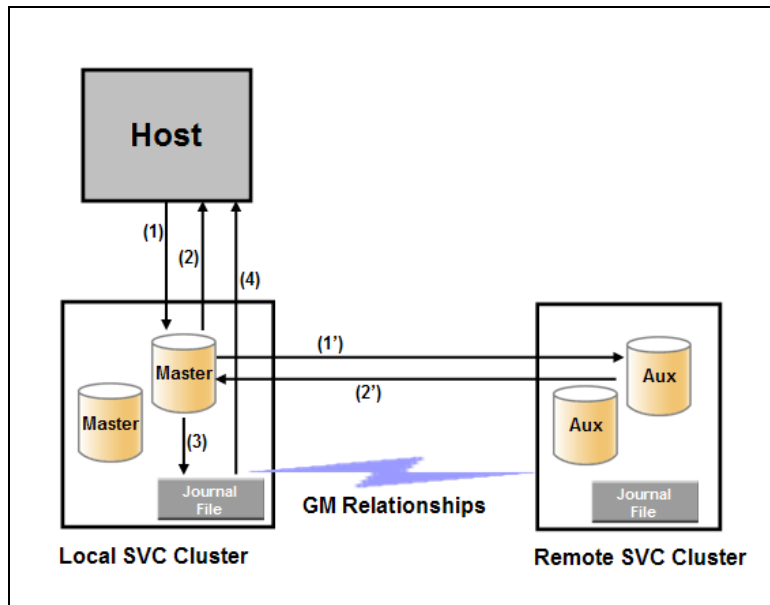


Figure 5-15 Colliding writes

The following numbers correspond to the numbers that are shown in Figure 5-15:

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 Volume functional overview

Global Mirror with Change Volumes 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, whereas Global Mirror with Change Volumes requires only to be sized to meet the average workload across a cycle period.

Figure 5-16 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.

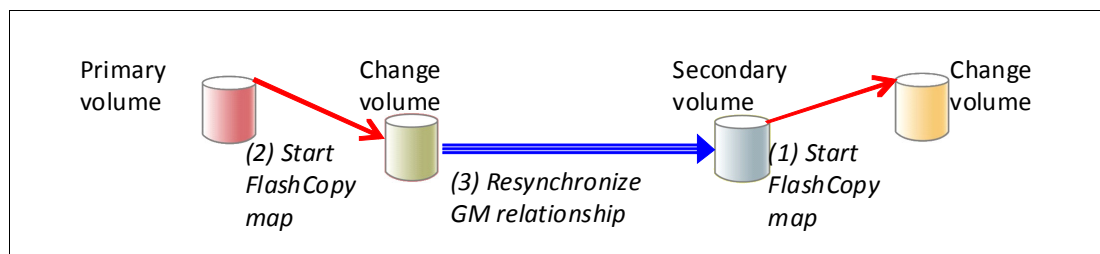


Figure 5-16 Global Mirror with Change Volumes

Global Mirror with Change Volumes will also only send one copy of a changed grain which might have been rewritten many times within the given cycle period.

If the primary volume fails completely for any reason, GM/CV is designed to ensure that the secondary volume will hold the same data as the primary did at a given point in time. That period of data loss might typically be between 5 minutes and 24 hours, but will vary according to the design choices you make.

Change Volumes hold point-in-time copies of 256 KB grains. If there is a change to any of the disk blocks in a given grain, that grain will be 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 5-17 on page 165 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.

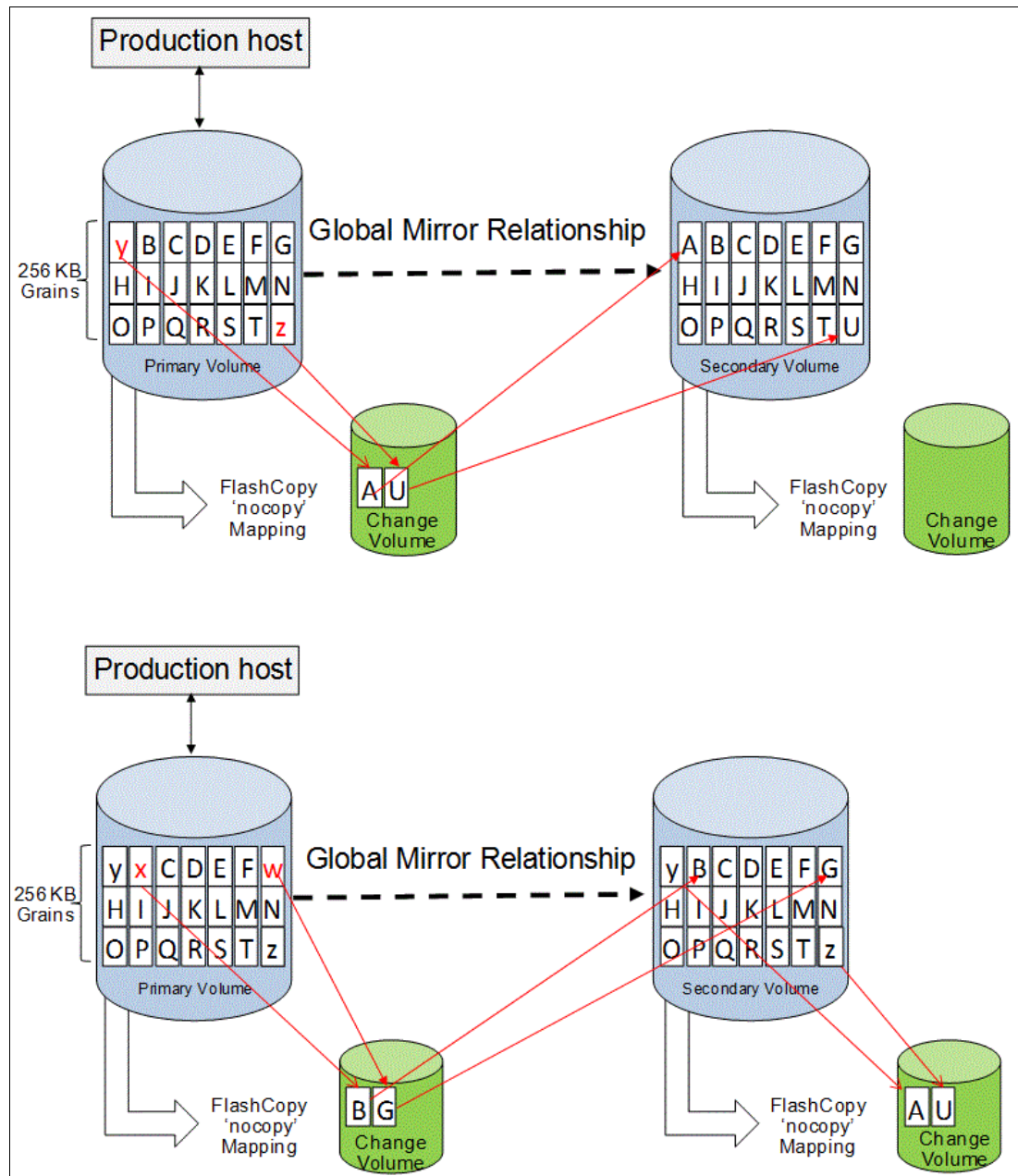


Figure 5-17 Global Mirror with Change Volumes uses FlashCopy point-in-time copy technology

FlashCopy mapping note: These FlashCopy mappings are not standard FlashCopy volumes and are not accessible for general use. They are internal structures dedicated to supporting Global Mirror with Change Volumes.

The options for **-cyclingmode** are **none** or **multi**.

Specifying or taking the default **none** will mean that Global Mirror acts in its traditional mode without Change Volumes.

Specifying **multi** will mean that Global Mirror will start 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 has finished. 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. Remember, however, that the shorter the cycle period, the less opportunity there is for peak write I/O smoothing, and the more bandwidth you will need.

Note that the `cyclingmode` can only be changed when the Global Mirror relationship is in a stopped state.

Recovery point objective using Change Volumes

Recovery point objective (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 will be not more than 2 x cycle period. 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 `lsrcrelationship` 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 a.m.
2. At 6:03 a.m., the cycle completes.
3. The freeze time would be 6:00 a.m. and the RPO at this point would be 3 minutes.
4. The cycle starts again at 6:05 a.m. and the RPO at this point would be 5 minutes.
5. The cycle is still running at 6:12 a.m. and the RPO is now up to 12 minutes, because 6:00 a.m. is still the freeze time of the last complete cycle.
6. At 6:13 a.m., the cycle completes and the RPO at this point would be 8 minutes because 6:05 a.m. is the freeze time of the last complete cycle.
7. 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 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 Spectrum Virtualize version 7.2 to implement intersystem networks 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 lists the maximum round-trip latency. This restriction applies to all variant of remote mirroring.

Table 5-6 Maximum round-trip

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.

Note: 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 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 174 for further zoning guidelines

Link bandwidth that is used by internode communication

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 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

An proper network sizing is essential for the remote copy services operations. Failing to estimate the network sizing requirements can lead a poor performance in remote copy services as well as in 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 re-synchronization, 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 will 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 will need to allow for unexpected peaks.

There are also unsupported tools available from IBM:

<http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD105947>

It is strongly advised that you 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, we suggest you 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 kilobytes ($= 8 \times 8 \text{ bits}/1024 = 0.0625 \text{ Mbits}$).
- ▶ For most of the day between 8 a.m. and 8 p.m. 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 a.m. to 8 p.m. 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 would require bandwidth on the instantaneous peak of 4500 writes per second as follows:

$$\begin{aligned} 4500 \times 0.0625 &= 282 \text{ Mbps} + 20\% \text{ resync allowance} + 5 \text{ Mbps heartbeat} \\ &= 343 \text{ Mbps dedicated plus any safety margin plus growth} \end{aligned}$$

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 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:

$$(4500 + 1500 + 1500) / 3 = 2500 \text{ writes/sec for a 30-minute cycle period}$$

The minimum bandwidth required for the cycle period of 30 minutes is:

$$\begin{aligned} 2500 \times 0.0625 &= 157 \text{ Mbps} + 20\% \text{ resync allowance} + 5 \text{ Mbps heartbeat} \\ &= 195 \text{ Mbps dedicated plus any safety margin plus growth} \end{aligned}$$

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:

$$(4500 + 5 \times 1500) / 6 = 2000 \text{ writes/sec for a 60-minute cycle period}$$

The minimum bandwidth required for a cycle period of 60 minutes is:

$$\begin{aligned} 2000 \times 0.0625 &= 125 \text{ Mbps} + 20\% \text{ resync allowance} + 5 \text{ Mbps heartbeat} \\ &= 155 \text{ Mbps dedicated plus any safety margin plus growth} \end{aligned}$$

Now consider if 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, so 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 p.m. each night and replicating until 8 a.m. at the latest would probably require bandwidth of at least:

$$\begin{aligned} (9000 + 70 \times 1500) / 72 &= 1584 \times 0.0625 = 99 \text{ Mbps} + 100\% + 5 \text{ Mbps heartbeat} \\ &= 203 \text{ Mbps at night plus any safety margin plus growth, non-dedicated,} \\ &\quad \text{time-shared with daytime traffic} \end{aligned}$$

Global Mirror with Change Volumes provides a way to maintain point-in-time copies of data at a secondary site where there is insufficient bandwidth to replicate the peak workloads in real-time.

Another factor that can reduce the bandwidth required for Global Mirror with Change Volumes is that it will also only send one copy of a changed grain, which might have been rewritten many times within the given cycle period.

Remember that these are simply examples. The central principle of sizing is that you need to know your data write rate, that 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 to be transferred to the remote site. In the real life this is unlikely to occur. Usually only a portion of a grain is changed during a GMCV cycle, but the transfer process always copy the whole grain to the remote site. This behavior can lead to an unforeseen overhead 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

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 the 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 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 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 37 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 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 ration *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 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 transceivers (SFPs) 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:

<http://www.ibm.com/storage/support/2145>

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 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 use transceivers that are designed for your particular SAN switch *only*.

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 and 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 time-out value and the resource allocation time-out value. It is considered that every extra hop adds about 1.2 microseconds of latency to the transmission.

Currently, 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 primary and secondary system.

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 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 must stop 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 a 8 Gbps link connecting two switches 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 acknowledgment (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 is 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 recommended configuration is described in the following Flash Alert:

<https://www.ibm.com/support/docview.wss?uid=ssg1S1003634>

The best practice for the 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 the remote port masking implementation is recommended.

This configuration of local node port masking is less of a requirement on Storwize family systems, where traffic between node canisters in an I/O group is serviced by the dedicated inter-canister link in the enclosure. The following guidelines also apply to the remote system connectivity:

- ▶ 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.

Specifically for the 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, SAN zoning should be applied to provide separate intrasystem zones for each local-remote I/O group pair that is used for replication, as shown in Figure 5-18 on page 175.

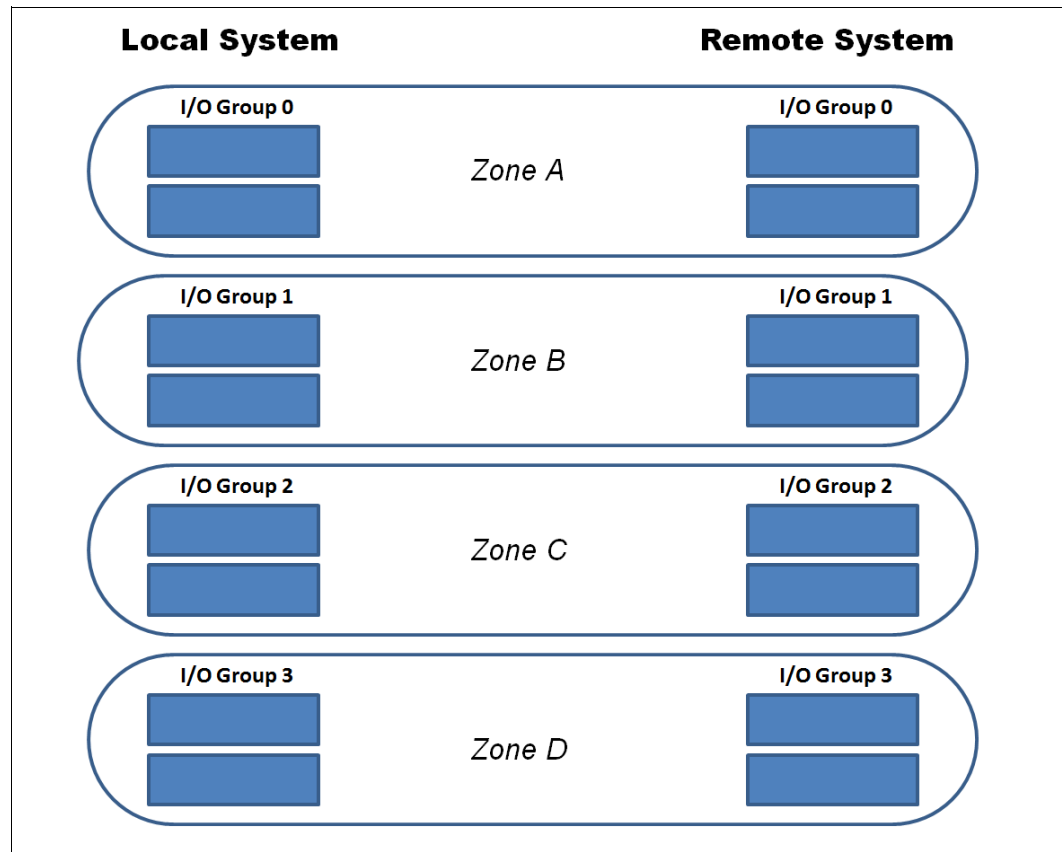


Figure 5-18 Zoning scheme for >80ms remote copy partnerships

Native IP connectivity

Remote Mirroring over IP communication is supported on the Spectrum Virtualize and Storwize Family systems by using Ethernet communication links. The 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 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 the section 5.4, "Native IP Replication" on page 199.

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 Spectrum Virtualize V7.8.

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	▶ SAN Volume Controller models SV1, DH8, CG8 and CF8 ▶ Storwize V7000 models 524 (Gen2) and 624 (Gen2+)	This can be any mix of Metro Mirror and Global Mirror relationships.
	8192	Any other Storwize model	This can be any mix of Metro Mirror and Global Mirror relationships. Maximum requires an 8-node system (volumes per I/O group limit applies).
Active-Active Relationships	1250	▶ 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 may 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.

Remote copy property	Maximum	Apply to	Comment
IP partnership Software Compression Limit	70 MB/s	▶ SAN Volume Controller models CG8 and CF8 ▶ Storwize V7000 model 124 (Gen1)	
	140 MB/s	▶ SAN Volume Controller models SV1 and DH8 ▶ Storwize V7000 models 524 (Gen2) and 624 (Gen2+)	

Similarly 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 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 restrictions

To use Metro Mirror and Global Mirror, you must adhere to the following rules:

- ▶ You must have the same target volume size as the source volume size. 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.
- ▶ You cannot resize Metro Mirror or Global Mirror volumes.
- ▶ 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 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.

Remote copy upgrade scenarios

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.

Pre-existing remote copy relationships are unaffected by a software upgrade that is performed correctly.

Remote copy compatibility cross-reference

Even if it is not a best practice, remote copy partnership can be established, with some restriction, among system with different Spectrum Virtualize version. For more information about a compatibility table for intersystem Metro Mirror and Global Mirror relationships between Spectrum Virtualize code levels, see *SAN Volume Controller Inter-system Metro Mirror and Global Mirror Compatibility Cross Reference*, S1003646, which 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 bandwidth (the rate of background copy) is then subdivided and distributed on a per-node basis. It is divided evenly between the nodes, which have volumes that perform a background copy for active copy relationships.

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 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 5-19 shows an example of remote copy resources that are not optimized. Volumes from the local system are replicated to the remote system, where 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.

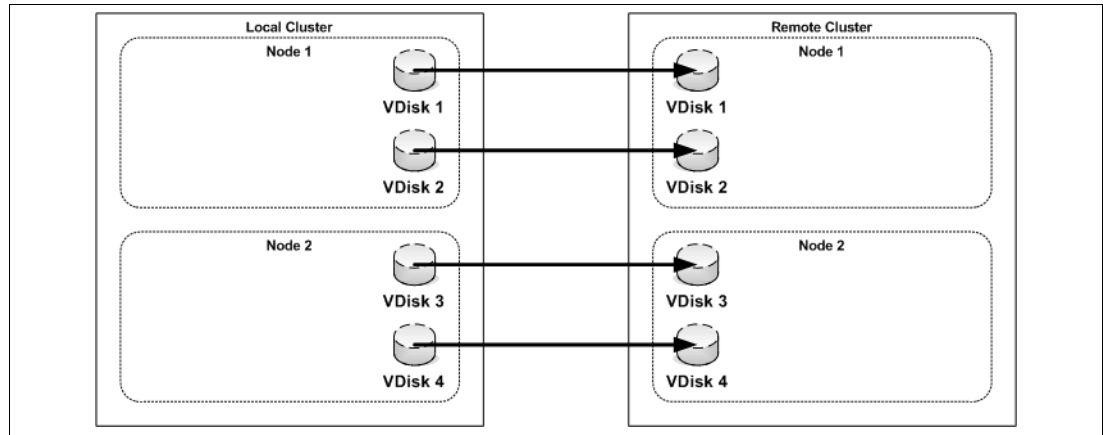


Figure 5-19 Remote copy resources that are not optimized

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 5-20, all remote copy resources for each node are used and remote copy operates with better performance than this configuration.

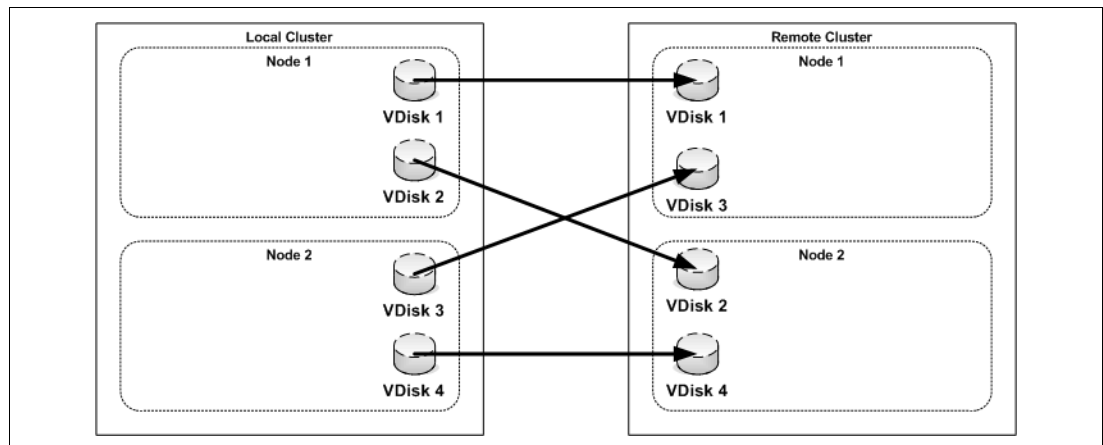


Figure 5-20 Optimized Global Mirror resources

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 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

To set the background copy bandwidth optimally, ensure that you consider all the above resources. 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.

Changes in the environment, or loading of it, can affect the foreground I/O. Spectrum Virtualize and Storwize technology provides the client with a means to monitor, and a parameter to control, how foreground I/O is affected by running remote copy processes. 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 188).

Finally note that with Global Mirror Change Volume the cycling process that transfer the data from the local to the remote system is a background copy task. For this reason the background copy rate setting affect the available bandwidth not only during the initial synchronization but also during the normal cycling process.

Backend storage considerations

In order to reduce the overall solution costs, it is a common practice to provide the remote systems with lower performance characteristics comparing 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 may indeed be limited by the performance of the remote system.

The recommendation is to perform an accurate backend resources sizing for the remote system in order 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 **chsystem** 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** apply also 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, and 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 - 86400 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 this before putting them into a live production environment.

Settings considerations about the **gmlinktolerance** and the **gmmaxhostdelay** parameters are described later.

rcbuffersize

rcbuffersize was introduced with the Version 6.2 code level so that systems with extremely 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 on page 182 shows how to change **rcbuffersize** to 64 MB 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 will be taken away from general cache.

Maxreplicationdelay and partnershipexclusionthreshold

Spectrum Virtualize version 7.6 has introduced two new parameters, **maxreplicationdelay** and **partnershipexclusionthreshold**, for the remote copy advanced tuning.

maxreplicationdelay is a system-wide parameter which defines a maximum latency (in seconds) for any individual write passing through the Global Mirror logic. If a write is completely hung for that time, for example due to a rebuilding array on the secondary system, Global Mirror will stop the relationship (and any containing consistency group) triggering a 1920 error.

The **partnershipexclusionthreshold** parameter has been introduced to allow users to set the timeout for an IO that will trigger a temporarily dropping of the link to the remote cluster. The value must be a number from 30 to 315.

Important: Do not change the **partnershipexclusionthreshold** parameter except under the direction of IBM Support.

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, as a testing and planning feature, Global Mirror allows you to simulate the effect of the round-trip delay between sites 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

This section describes the common uses cases of remote copy services.

Synchronizing a remote copy relationship

When creating a remote copy relationship two options regarding the initial synchronization process are available:

1. The *not synchronized* option is the default. With this option, when a remote copy relationship is started a full data synchronization occurs between the source and target volumes. It is the simplest in that it requires no other administrative activity apart from issuing the necessary Spectrum Virtualize commands. However, in some environments, the available bandwidth make this option unsuitable.
2. 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 make 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 issue 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, remote copy reports the relationship as being *consistent*, when it is not, which 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 as the following section describes.

Global Mirror relationships, saving bandwidth, and resizing volumes

Consider a situation where you have a large source volume (or many source volumes) 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 might be less expensive.

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. To increase the size of these volumes, you must delete the current mirror relationships and redefine the mirror relationships after you resize the volumes.

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 the volumes that become the source volumes in a Metro Mirror relationship or in a Global Mirror relationship.

2. Establish the Spectrum Virtualize system partnership with the target 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 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 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 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 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 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.

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.

Remote copy source as an FlashCopy target

Starting with V6.2 a FlashCopy target volume can be used as primary volume for a Metro or Global Mirror. The inclusion of Metro Mirror and Global Mirror source as an FlashCopy target helps in disaster recovery scenarios. You can have both the FlashCopy function and Metro Mirror or Global Mirror operating concurrently on the same volume.

However, the way that these functions can be used together has the following constraints:

- ▶ A FlashCopy mapping must be in the `idle_copied` state when its target volume is the secondary volume of a Metro Mirror or Global Mirror relationship.
- ▶ A FlashCopy mapping cannot be manipulated to change the contents of the target volume of that mapping when the target volume is the primary volume of a Metro Mirror or Global Mirror relationship that is actively mirroring. That is a FlashCopy mapping can not 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.

Native controller Advanced Copy Services functions

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>

Using backend controller's copy services

When the Spectrum Virtualize uses a LUN from a storage controller that is a source or target of Advanced Copy Services functions, you can use only that LUN as a cache-disabled image mode volume.

If you leave caching enabled on a volume, the underlying controller does not receive any write I/Os as the host writes them. The Spectrum Virtualize caches them and processes them later, which can have more ramifications if a target host depends on the write I/Os from the source host as they are written.

Performing cascading copy service functions

Cascading copy service functions that use 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 Spectrum Virtualize/Storwize copy services and by combining 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 5-21 describes the configuration for a three site cascading using the native storage controller copy services in combination with Spectrum Virtualize/Storwize remote copy functions.

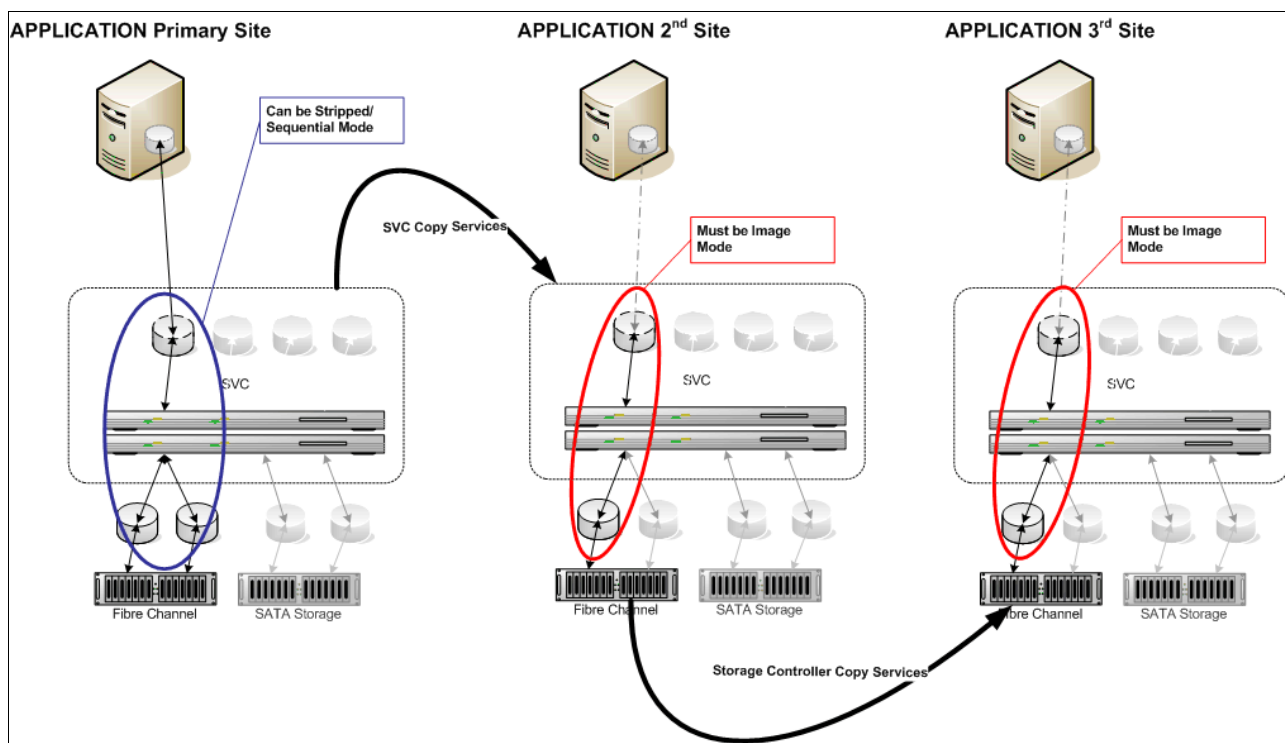


Figure 5-21 Using three-way copy services

In Figure 5-21, the primary site uses 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.

Cascading with Spectrum Virtualize and Storwize systems copy services

A cascading-like solution is possible also combining the Spectrum Virtualize/Storwize copy services. These remote copy services implementation are useful in three site disaster recovery solutions and data center moving scenarios.

In the configuration described in Figure 5-22, 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 the Site B, the primary disaster recover site. A third system, the 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.

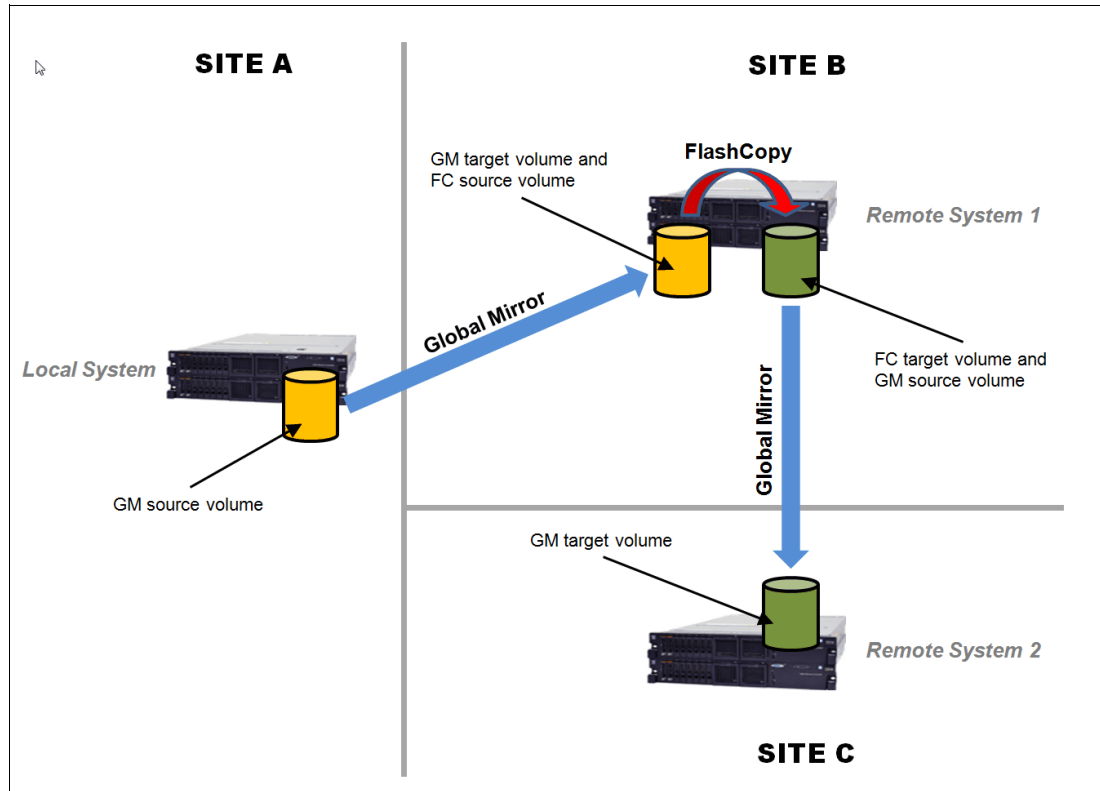


Figure 5-22 Cascading-like infrastructure

To implement a cascading-like solution, the following steps must be performed:

- Set up phase. Perform the following actions to initially set up the environment:
 - b. Create the Global Mirror relationships between the Local System and the Remote System 1.
 - c. Create the FlashCopy mappings in the Remote System 1 using the target Global Mirror volumes as FlashCopy source volumes. The FlashCopy must be incremental.
 - d. Create the Global Mirror relationships between the Remote System 1 and the Remote System 2 using the FlashCopy target volumes as Global Mirror source volumes.
 - e. Start the Global Mirror from Local System to the Remote System 1.

Once the Global Mirror is in ConsistentSynchronized state, we are ready to create the cascading.

- Consistency point creation phase. The following actions must be performed every time a consistency point creation in the Site C is required.
 - a. Check if the Global Mirror between the Remote System 1 to the 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 the Remote System 1.
 - c. Start the FlashCopy in the Remote Site 1.
 - d. Resume the Global Mirror between the Local System to the Remote System 1.
 - e. Start/resume the Global Mirror between the Remote System 1 to the Remote System 2.

The first time these operations are performed, a full copy between Remote System 1 and Remote System 2 will occur. Later executions of these operations will perform incremental

resynchronizations instead. Once the Global Mirror between the Remote System 1 to the Remote System 2 is in ConsistentSynchronized state, the consistency point in Site C is created. The Global Mirror the Remote System 1 to the Remote System 2 can now be stopped to be ready for the next consistency point creation.

5.3.5 1920 error

A 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 error 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 on the 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. Anyway the Spectrum Virtualize code implements many control mechanisms which mitigate the impacts of the Global Mirror to the foreground IOs.

gmmaxhostdelay 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 back logs that are 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 **gmlinktolerance** parameters to monitor Global Mirror protocol backlogs in the following ways:

- Users set the **gmmaxhostdelay** and **gmlinktolerance** 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 is greater than the **gmmaxhostdelay** setting, that sample period is marked as *bad*.
- Software keeps a running count of *bad periods*. Each time 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 **gminktolerance** parameter is defined in seconds. Bad periods are assessed at intervals of 10 seconds. The maximum bad period count is the **gminktolerance** parameter value that is divided by 10. For instance, with a **gminktolerance** 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, might result 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 **gminktolerance** 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 **gminktolerance** value, **gmmaxhostdelay** setting, or I/O load might reduce the risk of encountering this edge case.

maxreplicationdelay and partnershipexclusionthreshold

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 may be outstanding from the host before the relationship being stopped triggering a 1920 error. It can protect the hosts from seeing timeouts due to secondary hung IOs.

This parameter is mainly intended to protect from secondary system issues and it is not to 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, by setting the **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. Along with the 1920 error, the specific event ID 985004 is generated with the text “Maximum replication delay exceeded”.

The **maxreplicationdelay** parameter values can range from 0 to 360 seconds. Setting **maxreplicationdelay** to 0 will disable the feature.

The **partnershipexclusionthreshold** is a system-wide parameter that sets the timeout for an IO that will trigger a temporarily dropping of the link to the remote system. Similarly to **maxreplicationdelay**, the **partnershipexclusionthreshold** attribute provides some flexibility in a part of replication that tries to shield a production system from hung IOs on a secondary system.

In a Spectrum Virtualize/Storwize system, a node assert (restart with a 2030 error) occurs if any individual IO takes longer than 6 minutes. To avoid this situation, some actions are attempted to clean up anything that might be hanging IO before the IO gets to 6 minutes.

One of these actions is temporarily dropping (for 15 minutes) the link between systems if any IO takes longer than 5 minutes 15 seconds (315 seconds). This will often remove hang conditions caused by replication problems. The **partnershipexclusionthreshold** parameter introduced the ability to set this value to a time lower than 315 seconds in order to respond to hung IO more swiftly. The **partnershipexclusionthreshold** value must be a number from 30 to 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 will 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 have no interaction with **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, **IOstats**, in Spectrum Virtualize. Although not as powerful as IBM Spectrum Control, **IOstats** can provide valuable debug information if the **snap** command gathers system configuration data close to the time of failure.

The main performance statistics that should be investigated for the 1920 error are:

- *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 received per second, and the bandwidth that those writes consume. Summing the rate over many volumes you intend to replicate will give 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 the host is seeing: if a user complains that an application is slow, and the stats show the write response time leap from 1ms to 20ms, the two are most likely linked. Not always though: some applications with high queue depths and low to moderate workloads won't be impacted 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 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 1ms for a fast-performing system.

► *Global Mirror Write I/O Rate*

This 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 32kB or smaller tracks before replication, this might be quite different to the Write I/O Rate on the primary volume (magnified further because the samples on the two systems won't be aligned, so will capture a different set of writes).

► *Global Mirror Overlapping Write I/O Rate*

These monitor the amount of overlapping IO that the Global Mirror feature is handling for regular Global Mirror relationships. That's 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 IOs, a journaling feature has been implemented, as discussed in "Colliding writes" on page 162.

► *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, including the time to send the data to the remote system, consistently apply it to the secondary non-volatile cache, and send an acknowledgement back to the primary system. For secondary volumes, it will record only the time taken to consistently apply it to the system cache, which will normally be up to 20ms, with most of that time spent coordinating consistency across many nodes and volumes. Primary and secondary volumes for a relationship will tend to record times that differ by the round-trip time between systems. If this is high on the secondary system, look for congestion on the secondary system's fabrics, saturated secondary 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's a good indication that the write rate is faster than the storage can cope with. If this starts to increase on secondary storage while primary volumes suffer from increased Write Response Time, it's possible that the secondary storage isn't fast enough for the replicated workload.

► *Port to Local Node Send Response Time*

The time in milliseconds 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 IO group) and get an acknowledgement back. This should be well below 1ms, with values below 0.3ms essential for regular Global Mirror to provide a Write Response Time below 1ms (as up to 3 round-trip messages within the local system will happen before a write completes to the host). If 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 it takes to send a message to nodes in other systems and get an acknowledgement back. This is not separated out by remote system, but for environments that have replication to only one remote system, this should be very close to the low-level ping time between your sites. If this starts going significantly higher, it's likely that the link between your systems is saturated, which will usually cause high Zero Buffer Credit Percentage too.

► *Sum of Port-to-local node send response time and Port-to-local node send queue*

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 1-4)*

These values show how heavily loaded the nodes in the system are. If any core is has high utilization (say, over 90%) and there is an increase in write response time, it's 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 as the port had run out of buffer credits. If you have a long link from the node to the switch it is attached to, there may be benefit in getting the switch to grant more buffer credits on its port. It's more likely to be the result of congestion on the fabric, as running out of buffer credits is how Fibre Channel performs flow control. Normally, this value will be well under 1%. Between 1 and 10% is a concerning level of congestion, but you may find the performance acceptable. Over 10% is fairly extreme congestion. This is also called out on a port-by-port basis in the port-level statistics, which gives finer granularity of where any congestion may 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 used for messages to nodes in other systems.

► *Back-end Write Response Time*

This values is the average response time in milliseconds for write operations to the backend storage. This time might include several physical I/O operations, depending on the type of RAID architecture. Poor backend 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
- ▶ 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; for example, 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, to examine the internode communications, storage controller performance, or both. By using IBM Spectrum Control, you can check the storage metrics before 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 adjacent to your most bandwidth-constrained link between the systems. Exactly what's provided, and how to analyse it, will vary 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 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, as you will 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 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, which, 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. The problem similarly 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 (pilfering 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. For example, the algorithms that access the effect of the running Global Mirror incorrectly interpret slow foreground write activity (and the slow background write activity that is associated with it) as being slow as a consequence of running Global Mirror. Then, the Global Mirror relationships stop.

Node and canister

For the 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.

As Global Mirror needs to synchronize its IO processing across all nodes in the system to ensure data consistency, if any node is running out of CPU, it can impact all relationships. So check the CPU usage statistic. If it looks higher when there's a performance problem, then running out of CPU bandwidth could be causing the problem. Of course, CPU usage will go up when the IOPS going through a node goes up, so if the workload increases, you'd expect to see CPU usage increase.

If there's an increase in CPU usage on the secondary system but no increase in IOPS, and volume write latency increases too, it's 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 at that time) and consider moving that activity to a time with less workload. If there's an increase in both CPU usage and IOPS, and the CPU usage is close to 100%, then that node may be overloaded.

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 will potentially lower CPU usage, and will also lower the opportunities for each IO 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 additional IO 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 1.0ms is surprisingly high, and will cause high secondary volume write response time. Values greater than 1ms on primary system will guarantee an impact on write latency to Global Mirror primary volumes of 3ms or more.

If you've checked CPU utilization on all the nodes, and seen that there's no signs of it getting near 100% at the time, a high *Port to local node send response time* means it's likely that there's 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 173 for more information on Buffer Credit). If any port is seeing over 10% zero buffer credits, that's definitely going to cause a problem for all IO, not just Global Mirror writes. Values from 1-10% are moderately high and could 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, though this only really becomes an option on systems with more than 4 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, and therefore, their performance is impacted. To resolve this problem, start the FlashCopy mapping, which re-enables 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 backend storage, especially when the background copy is active (see “Background Copy considerations” on page 150). In case 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 leading to poor remote copy performances and then to a 1920 error. A careful planning of the backend resources is particularly recommended with this kind of scenarios. Reducing the FlashCopy background copy rate can also help to mitigate this situation.

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 using standard 64-byte ping packets.

In Figure 5-23, 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.

Before we proceed, we 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 5-23 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 5-23 Effect of packet size (in bytes) versus the link size

In Figure 5-23 on page 197, 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 a 1920 error occurs, the Global Mirror auxiliary volumes are no longer in the ConsistentSynchronized 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.

Tip: If the relationship stopped in a consistent state, you can use the data on the auxiliary volume at the remote system as backup. Creating a FlashCopy of this volume before you restart the relationship gives more data protection. The FlashCopy volume that is created maintains the current, consistent, image until the Metro Mirror or Global Mirror relationship is synchronized again and back in a consistent state.

To ensure that the system can handle the background copy load, you might want to 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 stops in an inconsistent state.

Adjusting the Global Mirror settings

Even though the default values are valid in most configurations, the settings of the **gmlinktolerance** 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 **gmlinktolerance** 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 (**gmlinktolerance 300**, **gmmaxhostdelay 5**) is saying “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 **gmlinktolerance 30**, **gmmaxhostdelay 100** says “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 could disrupt your server environment. The general recommendation is to change the **gmlinktolerance** and **gmmaxhostdelay** values progressively and evaluate the overall impact in order 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.

Note that the **maxreplicationdelay** settings does not mitigate the 1920 error occurrence, because it actually adds an additional trigger to the 1920 error itself. However the **maxreplicationdelay** provides users with a fine granularity mechanism to manage the hung IOs condition and it can be used in combination with **gminktolerance** and **gmmaxhostdelay** settings to better address particular environment conditions. In the aforementioned VMware example, an alternative option is to set the **maxreplicationdelay** to 30 seconds and let the **gminktolerance** and **gmmaxhostdelay** settings to their default. With these settings, the **maxreplicationdelay** timeout handles effectively the hung IOs conditions, while the **gminktolerance** 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 Spectrum Virtualize and Storwize family products running code version 7.2 or higher, using the built in networking ports or optional 1/10Gbit adapter cards. Following a recent partnership with IBM, it utilizes 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) Transmission Control Protocol/Internet Protocol (TCP/IP) link significantly improves the utilization of the link. It does this by applying patented Artificial Intelligence (AI) to hide latency normally associated with WANs. Therefore, 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 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.

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 5-24.

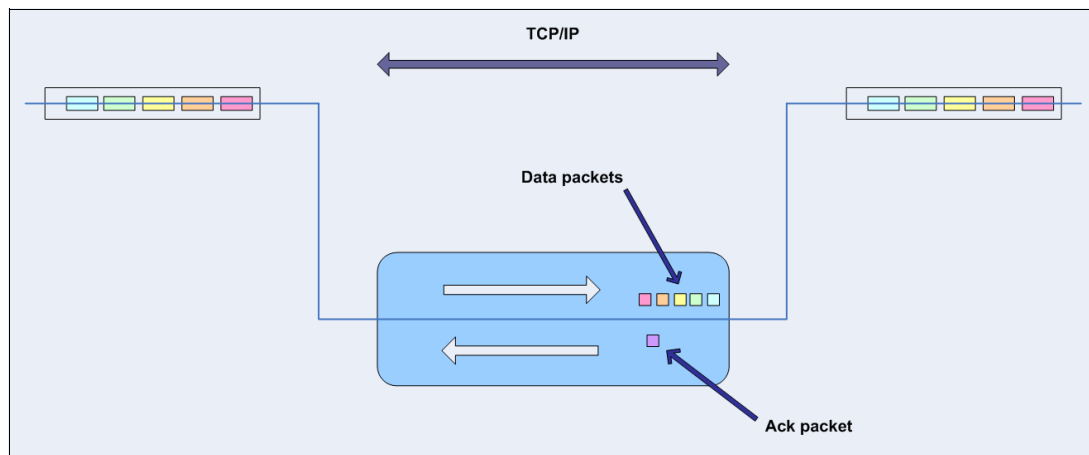


Figure 5-24 Typical Ethernet network data flow

However, by using the embedded IP replication, this behavior can be eliminated with the enhanced parallelism of the data flow by using multiple virtual connections (VC) 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 5-25.

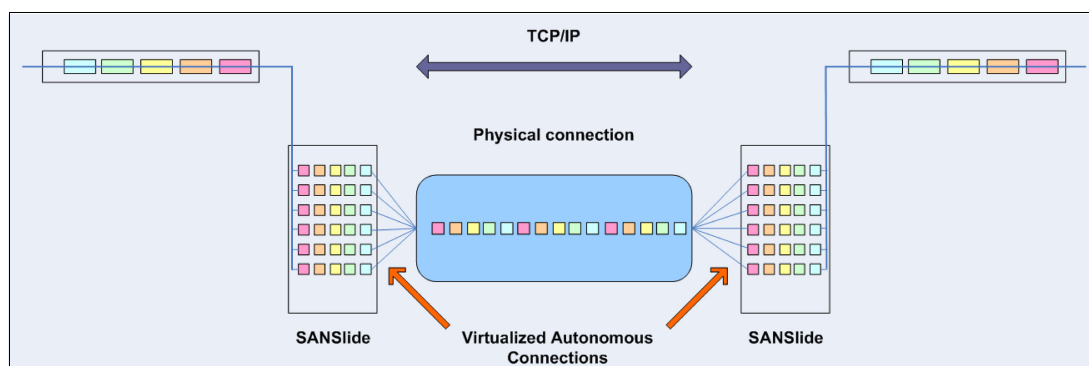


Figure 5-25 Optimized network data flow by using Bridgeworks SANSlide technology

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 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 restriction:
 - 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.
- ▶ NAT (Network Address Translation) 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 could be connected (by Ethernet switches) to this physical link. Similarly, multiple IP address on remote system B could be connected (by Ethernet switches) to the same physical link. At any point in time, 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, and varies based on distance (for example, round-trip latency) and quality of communication link (for example, packet loss):
 - One 1 Gbps port might transfer up to 110 megabytes per second (MBps) unidirectional, 190 MBps bidirectional
 - Two 1 Gbps ports might transfer up to 220 MBps unidirectional, 325 MBps bidirectional
 - One 10 Gbps port might transfer up to 240 MBps unidirectional, 350 MBps bidirectional
 - Two 10 Gbps port might transfer up to 440 MBps unidirectional, 600 MBps bidirectional

Note: The Bandwidth setting definition when the IP partnerships are created 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 the same 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 in order not to experience connectivity issues. After the VLANs are configured, changes to the VLAN settings will 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 will fail.

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 optionally 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

Spectrum Virtualize version 7.7 introduced the IP compression capability that speed up replication cycles or that may allow use of less bandwidth. This feature reduce the volume of data that must be transmitted during remote copy operations using with compression capabilities similar to those experienced with existing Real Time Compression implementations.

No License: IP compression feature doesn't require RtC software license.

The data compression is made within the IP replication component of the Spectrum Virtualize code and so it can be used with all the remote copy technology (Metro Mirror, Global Mirror and Global Mirror Change Volume). The IP compression is supported 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 feature provide two kind of compression mechanisms: the HW compression and the SW compression. The HW compression is active when compression accelerator cards are available, otherwise the SW compression is used.

The HW compression makes use of the currently under utilized cards with the internal resources shared between RACE and IP compression. The SW compression uses the system CPU and may have an impact on heavily utilized systems.

The 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 using the **mkippartnership** and **chippartnership** with **-compress** parameter. Furthermore, in systems with replication enabled in both directions, the IP compression may be enabled in only for one direction. IP compression is supported for IPv4 and IPv6 partnerships.

Figure 5-26 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 FlashSystem V9000	140 MB/s	500 MB/s
Storwize V5000 Gen2	140 MB/s	N/A
Storwize V7000 Gen1	70 MB/s	N/A
Storwize V7000 Gen2/Gen2+	140 MB/s	500 MB/s

Figure 5-26 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 associated with an IP port of system to indicate which physical IP link it is connected to. Multiple Spectrum Virtualize nodes could be connected to the same physical long-distance link, and must therefore share the same remote copy port group id.

In scenarios where there are 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 using the GUI or the **cfgport ip** 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 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 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 will failover 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 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, you should 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, there will be no connections in that remote-copy port group 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 re-established; as a result, the relationships might stop and a manual restart may be 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 so that the active IP partnership data path is made 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 recommended 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 5-27.

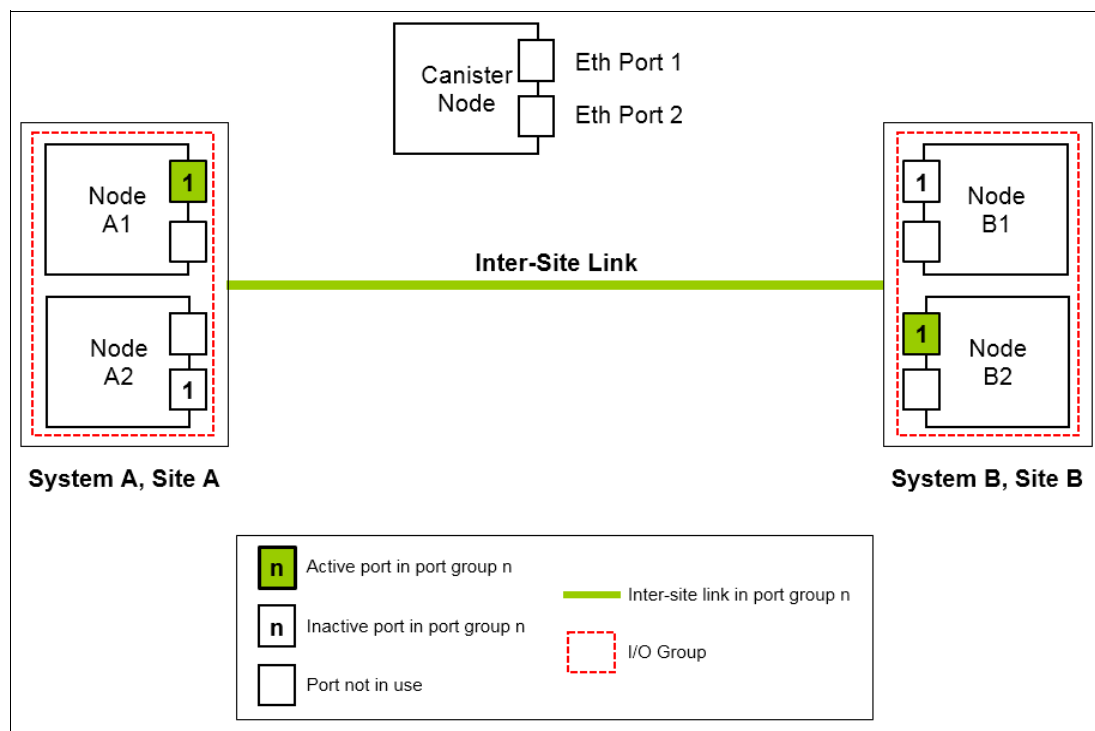


Figure 5-27 Only one remote copy group on each system and nodes with failover ports configured

As shown in Figure 5-27 on page 206, there are 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 there are two ports on each system that 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 if there is 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, and this 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 `1sportip` 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, but only one port in that remote copy port group is active at any time at each system.

If the Node A1 in System A or the 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 transition to the Not_Present state and recover.

In Figure 5-28 on page 207 a configuration with two 4-node systems in IP partnership over a single inter-site link (with failover ports configured) is depicted.

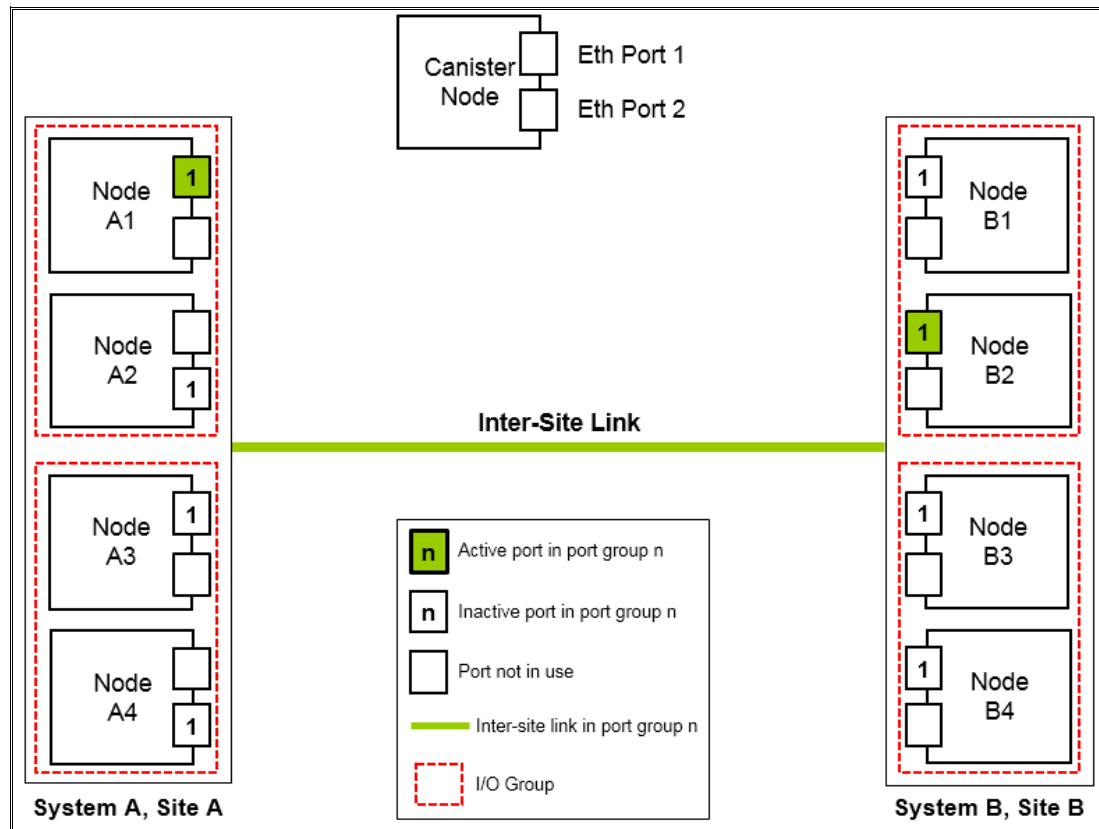


Figure 5-28 Multinode systems single inter-site link with only one remote copy port group

As shown in Figure 5-28, there are 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 there are four ports that 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, and this process can cause partnerships to transition to the Not_Present state.

This result causes remote copy relationships to stop, and 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 auto recover. The details of the particular IP port actively participating in the IP partnership process is provided in the **1sportip** 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 the Node A1 in System A or the Node B2 in System B were to encounter some failure in the system, IP partnerships discovery is triggered and it continues servicing the I/O from the failover port.
- ▶ The discovery mechanism that is triggered because of failover might introduce a delay wherein the partnerships momentarily transition 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 5-29.

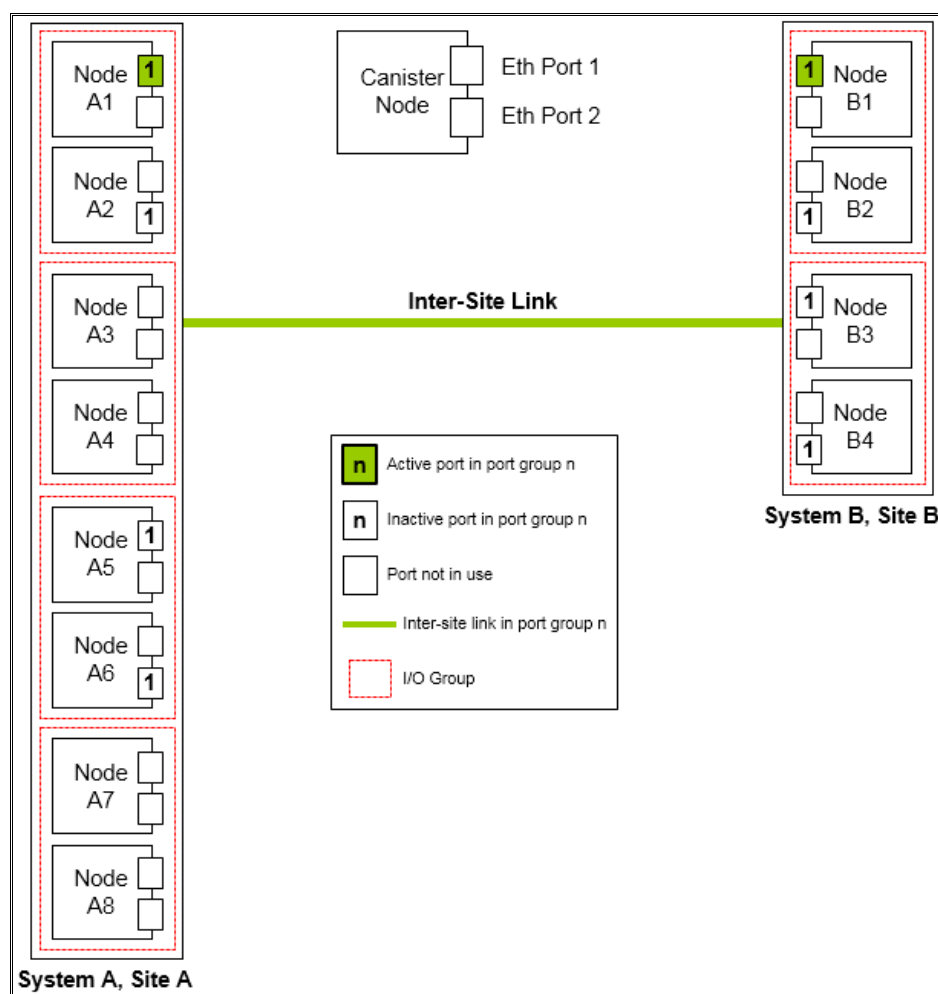


Figure 5-29 Multinode systems single inter-site link with only one remote copy port group

As shown in Figure 5-29, there is 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 and 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, and this delay might cause partnerships to transition to the Not_Present state.

This process can lead to remote copy relationships stopping, and 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 **1sportip** 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 wherein the partnerships momentarily transition 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 5-30.

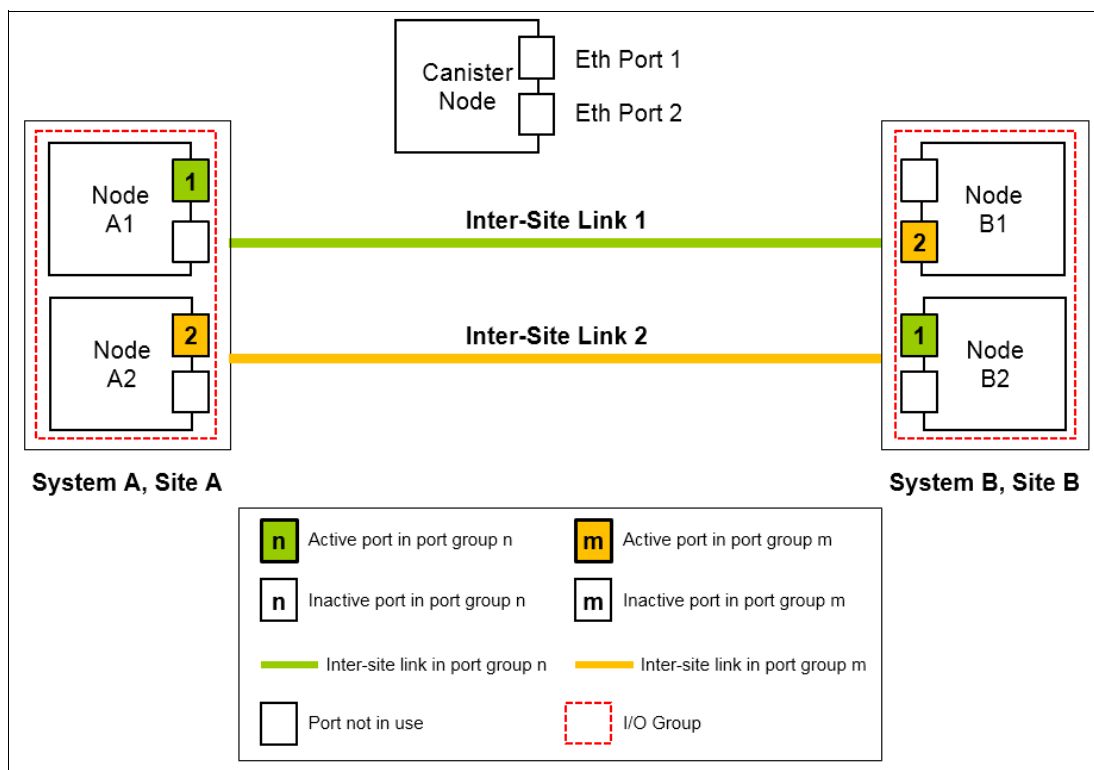


Figure 5-30 Dual links with two remote copy groups on each system configured

As shown in Figure 5-30 on page 210, remote copy port groups 1 and 2 are configured on the nodes in System A and System B because there are two inter-site links 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 nodes and they 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 5-31 on page 211.

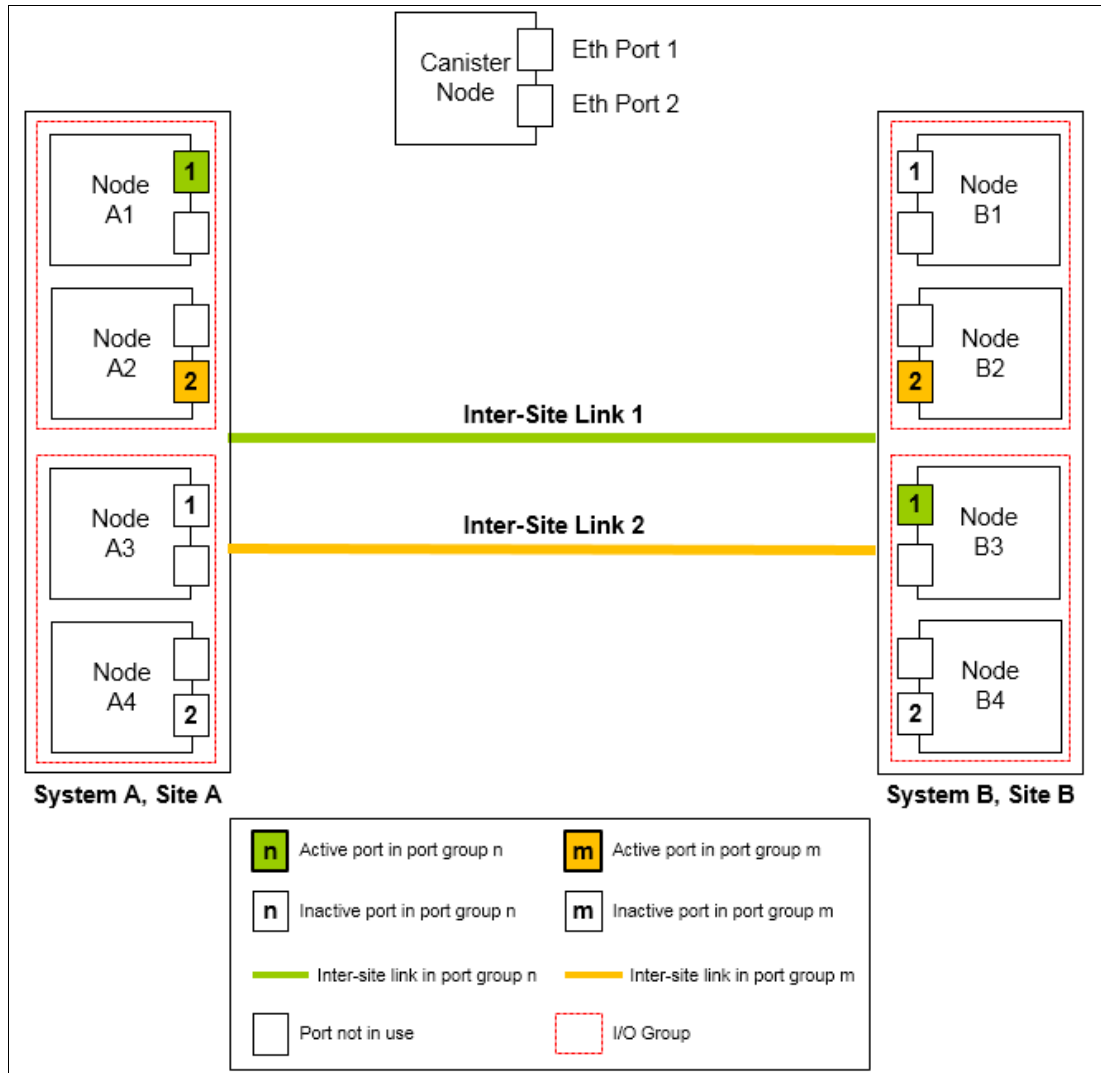


Figure 5-31 Multinode systems with dual inter-site links between the two systems

As shown in Figure 5-31, there are 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 and each node in the I/O group has a single port that is configured in remote copy port groups 1 or 2.

Although there are two ports that 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 in event of 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) and 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, which restores the total bandwidth, so both of the links are available to support IP partnership.

Finally a eight-node system in IP partnership with a four-node system over dual inter-site links is depicted in Figure 5-32.

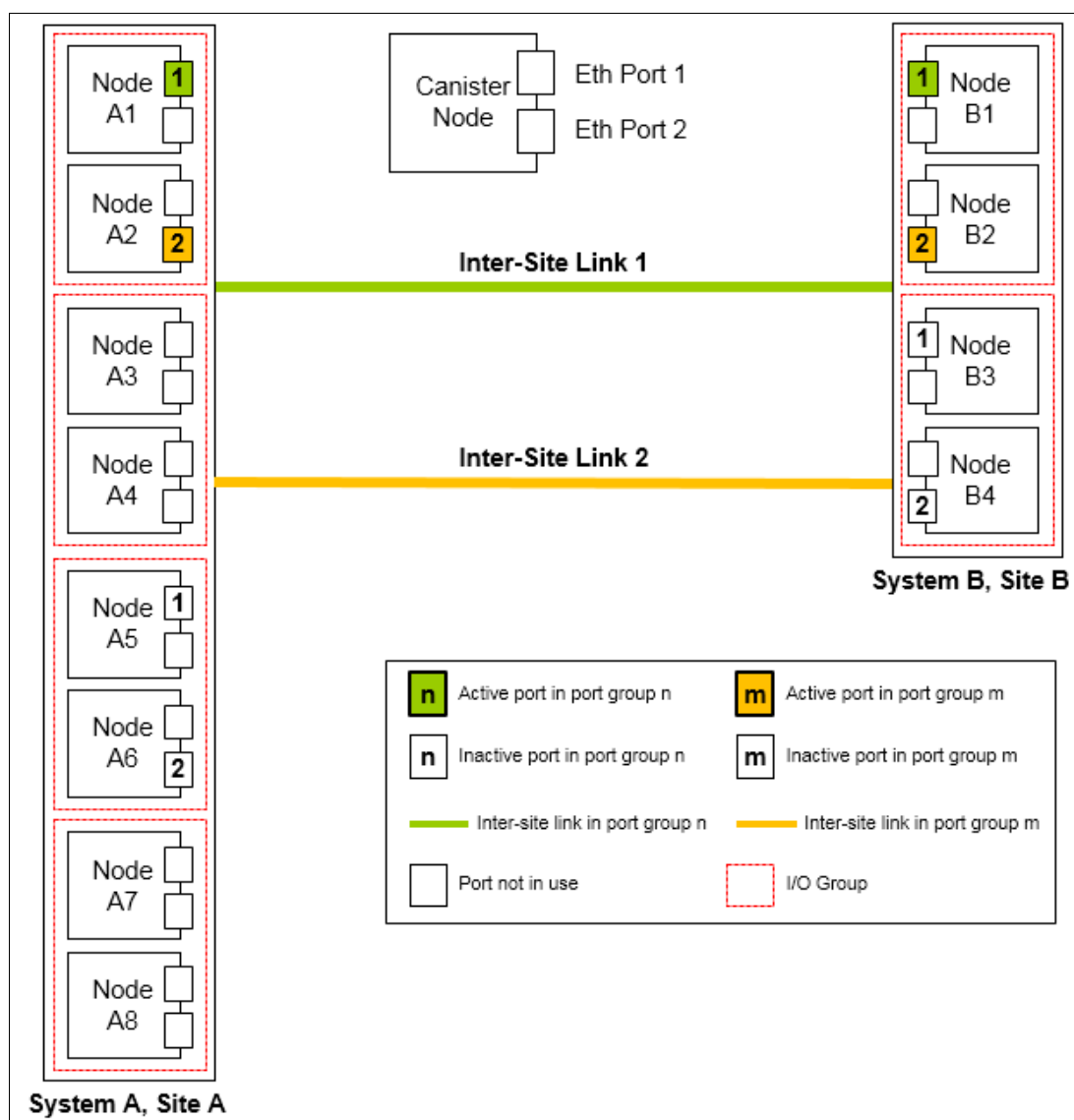


Figure 5-32 Multinode systems with dual inter-site links between the two systems

As shown Figure 5-32 on page 212, there is 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, but path selection logic is managed by an internal algorithm. Therefore, this configuration depends on the pathing algorithm to decide which of the nodes actively participates 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 might 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) and 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 **lsport 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, which 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 in order to make the IP replication more performing and reliable. With the version 7.7 a new workload distribution algorithm has been introduced that optimizes the usage of the 10Gbps ports. Nevertheless in presence of poor quality networks, that is with significant packet loss and high latency, the actual usable bandwidth may decrease considerably. For example, the Figure 5-33 on page 214 shows the throughput trend for a 1Gbps port in respect of the packet loss ratio and the latency.

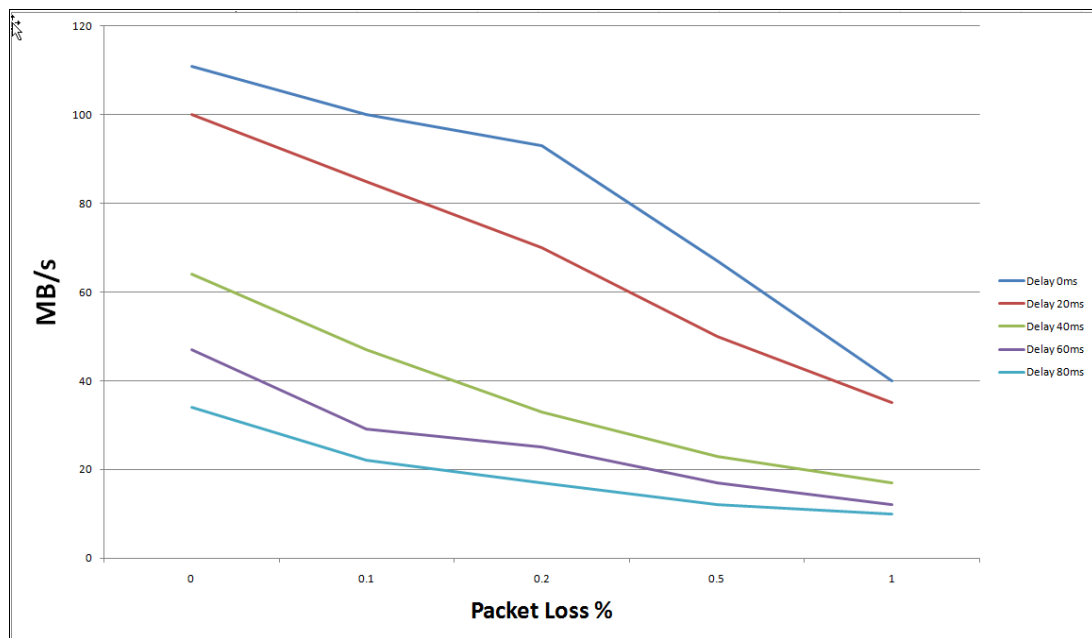


Figure 5-33 1Gbps port throughput trend

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 not be considered for the replication configuration requiring high quality and performing networks. Due to its characteristic of low bandwidth requirement, the Global Mirror Change Volume is the recommended 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, it is recommended to update 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 providing 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 will use the same virtualization policy. However, there is also a way to have two copies of a volume with different virtualization policies. In combination with *thin-provisioning*, each mirror of a volume can be thin-provisioned or fully allocated and in striped, sequential, or image mode.

A mirrored (secondary) volume has all of the capabilities of the primary volume copy and also 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* functionality that is achieved by “splitting” a copy from the volume. Note, however, that 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 5-34 provides an overview of Volume Mirroring.

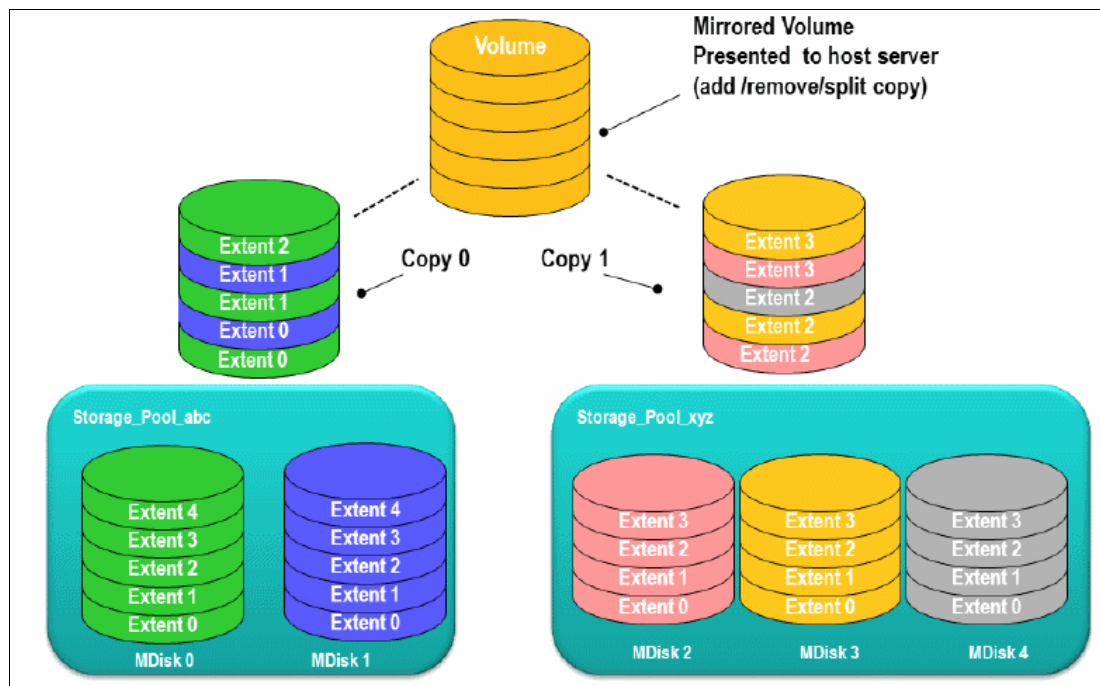


Figure 5-34 Volume Mirroring overview

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 copy is in synchronized status and so all the reads are directed to this copy. The write operations will be 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 these system topology and the enablement of site awareness capability, the concept of primary copy still exists but is not more relevant. The read operation will follow 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 an host I/O read is attempted to a mirrored disk through a Spectrum Virtualize Node in the Site A, then the I/O read will be directed to the copy in Site A, if available. Similarly, an host I/O read attempted through a node in Site B will go to Site B copy.

Important: For best performance it is still suggested to keep consistency between Hosts, Nodes and Storage Controller site affinity as long as possible.

In case of storage backend 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.

5.5.1 Volume mirroring use cases

Volume Mirroring offers the capability to provide extra copies of the data that can be used for High Availability solutions as well as 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.

Server 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 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 is useful if you are using the 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 using 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, so that the existing data remains intact. After you have imported it, all copy services functions can be used to migrate the storage to the other locations, while the data remains accessible to your hosts.

► Volume migration using Volume Mirroring and then using the Split into New Volume option

This option allows you to use the available RAID 1 functionality. You create two copies of data that initially have a set relationship (one primary and one secondary). You then break the relationship (both primary and no relationship) to make them independent copies of data.

You can use this option to migrate data between storage pools and devices. You might use this option if you want to move volumes to multiple storage pools.

► Volume migration using the Move to Another Pool option

This option allows any volume to be moved between storage pools without any interruption to the host access. This option is effectively a quicker version of the Volume Mirroring and Split into New Volume option. You might use this option if you want to move volumes in a single step, or you do not have a volume mirror copy already.

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, 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.

Other Volume Mirroring usage cases and characteristics follow:

► Creating a mirrored volume.

- The maximum number of copies is two.
- Both copies will be 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 specified will contain 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 again, 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:
 - Reports the first difference found.

It can iterate by starting at a specific Logical Block Address (LBA) using the **-startlba** parameter.
 - Creates virtual medium errors where there are differences.
 - Corrects the differences 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 will 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 now contain volume copy identifiers.

5.5.2 Mirrored volume components

Note the following points regarding mirrored volume components:

- ▶ A mirrored volume is always composed of two copies (copy0 and copy1).
- ▶ A volume that is not mirrored consists of a single copy (which for reference could 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 5-35 shows the attributes of a volume and Volume Mirroring.

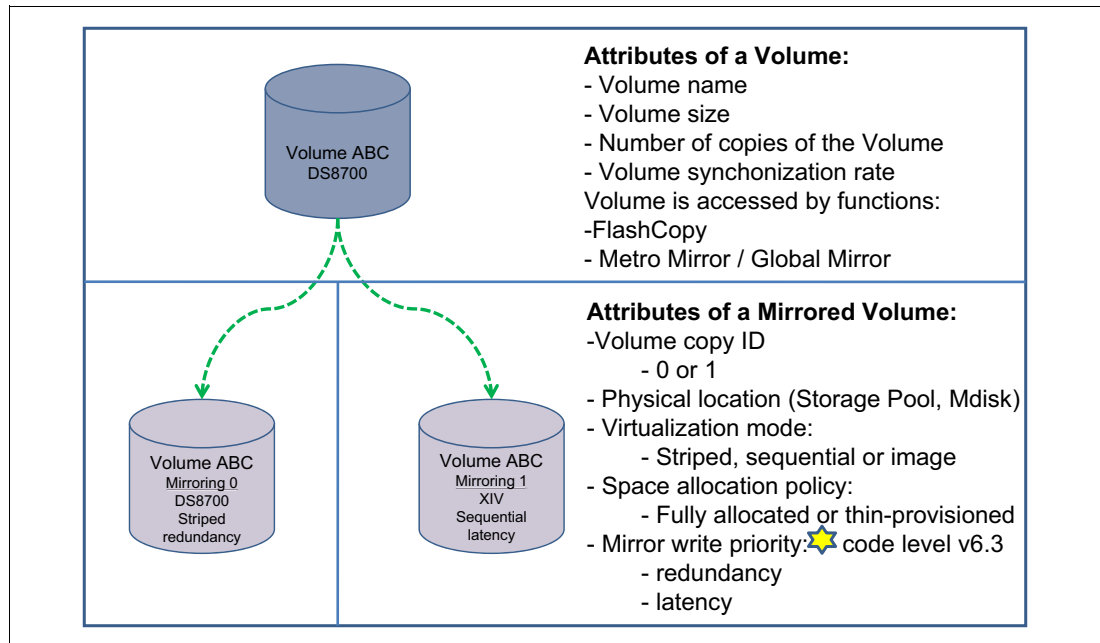


Figure 5-35 Attributes of a volume and Volume Mirroring

In Figure 5-35, XIV and DS8700 illustrate that a mirrored volume can utilize different storage devices.

5.5.3 Performance considerations of Volume Mirroring

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 located on Storage Pools with different performance characteristics, the slowest Storage Pool decides upon the performance of writes to the volume, which applies when writes have to be destaged to disk.

Recommendation: 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 will not be faster than without Volume Mirroring.

Be aware, however, that this is only true when both copies are synchronized. If the primary is out of sync, then reads will be submitted to the other copy. Finally, note that these considerations do not apply to 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 between 0 and 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 you want to mirror. The default value of a new volume mirror is 50%.

In large Spectrum Virtualize or Storwize system configurations the settings of the copy rate can affect considerably the performance in scenarios where a backend storage failure occurs. For instance consider a scenario where a failure of a backend storage controller is affecting one copy of 300 mirrored volumes. The host will continue the operations using the remaining copy. When the failed controller will come back online, immediately the resynchronization process for all the 300 mirrored volumes will start at the same time. With a copy rate of 100 for each volume this would add a theoretical workload of 18.75GB/s that will drastically overload the system.

The general recommendation for the copy rate settings is then to evaluate the impact of massive resynchronization and set the parameter accordingly.

Mirrored Volume and I/O Time-out Configuration

The source volume has pointers to two copies (mirrored volume copies) of data each 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 time-out has expired, then system has completed all available controller level ERPs (Error Recovery Procedure). 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 backend 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 IO, the system submits writes to both copies with a time-out of 10 seconds (20 seconds for stretched volumes). If one write succeeds and the other write takes longer than five seconds then the slow write is aborted. The fibre channel abort sequence can take around 25 seconds.

When the abort is done, one copy is marked as out of sync and the host write IO 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 using the command **chvdisk** and the **mirror_write_priority** attribute settings:

- ▶ *Latency* (default value): short time-out prioritizing low host latency. This option enable the fast failover feature
- ▶ *Redundancy*: long time-out prioritizing redundancy. This option indicates a copy that is slow to respond to a write I/O may 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 disable the fast failover feature.

Volume Mirroring ceases to use the slow copy for a period of between 4 to 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, shortly complete the synchronization.

If another I/O times out during the synchronisation 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 will get progressively more out of sync as more grains are written. Fast failovers occurring regularly suggests that there is an underlying performance problem with the copy's backend storage.

For mirrored volumes in Enhanced Stretched Cluster configurations the best practice is to set the **mirror_write_priority** field to *latency*.

5.5.4 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 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 there isn't enough space to allocate bitmaps in the target IO Group. To verify and change the space allocate and available on each IO Group with CLI see the Example 5-4.

Example 5-4 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 allocate and available on each IO Group with GUI see the Figure 5-36.

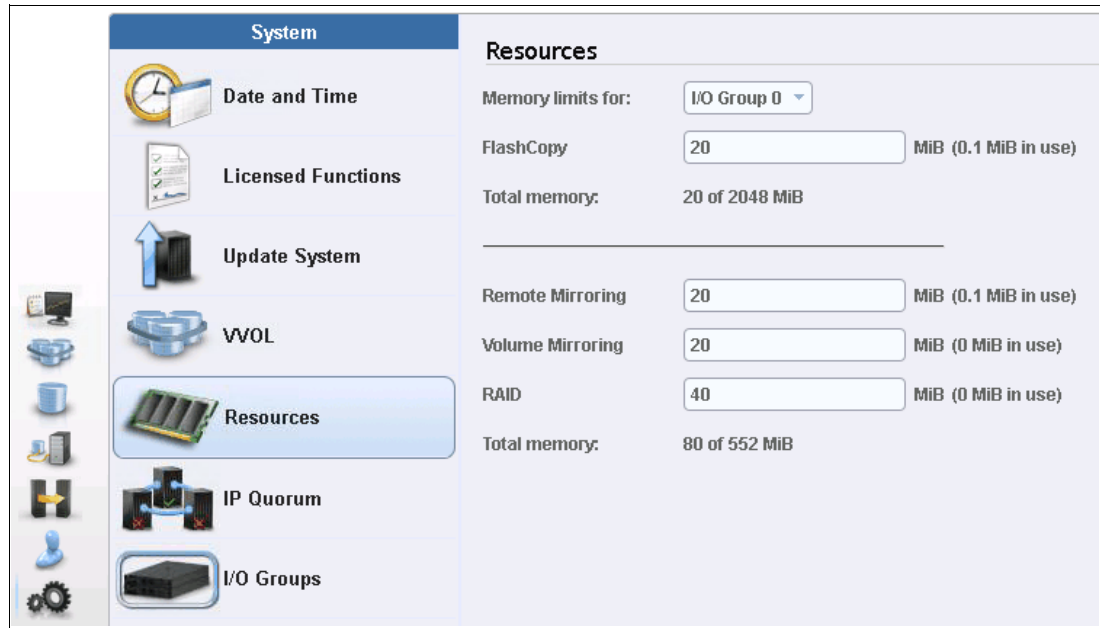


Figure 5-36 IOgrp feature example

5.5.5 Synchronization status of volume copies

As soon as a newly volume is created with two copies, the state of copies are in *out-of-synchronization*.

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 will be synchronized through the synchronization process, which runs at the default synchronization rate of 50 (see more details in Table 5-9 on page 220), or at the defined rate while creating or modifying the volume (see 5.5.3, “Performance considerations of Volume Mirroring” on page 219 for the impact of the copy rate setting)

There is a parameter **-fmtdisk** that ensures both copies are overwritten with zeros, and after this the volume comes online and they can be considered as *synchronized copies*. Both copies are filled with zeros, so they are exactly the same. Starting with version 7.5 the format process is initiated by default at the time of the volume creation.

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, and with the **-fmtdisk** parameter.

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 zeroes (0).

If the synchronization between the volume copies has been lost, the resynchronization process is incremental. This 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 `lsvdisksyncprogress` command.

6



Hosts

This chapter describes the guidelines on how to configure host systems on 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 a 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 V7.8 Information Center and Guides, which are available at this website:

<https://ibm.biz/Bdsvxb>

This chapter includes the following sections:

- ▶ Configuration guidelines
- ▶ N-Port Virtualization ID (NPV)
- ▶ 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 the number of paths through the fabric that are allocated to the host, the number of host ports to use, and the approach for spreading the hosts across I/O groups. They also apply to logical unit number (LUN) mapping and 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 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 bring a new host into the Spectrum Virtualize at the preferred levels, see *V7.8 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels*, which are available at these websites:

Spectrum Virtualize

<http://www.ibm.com/support/docview.wss?uid=ssg1S1009558>

Storwize V7000

<http://www.ibm.com/support/docview.wss?uid=ssg1S1009559>

When you are creating the host, use the host name from the host as the host object name in the Spectrum Virtualize to aid in configuration updates or problem determination in the future.

6.1.2 Host Cluster

Spectrum Virtualize software supports host clusters starting with version 7.7.1. The host cluster allows a user to create a group of hosts to form a cluster, which is treated as one single entity, thus allowing multiple hosts to have access to the same set of volumes.

Volumes mapped to that host cluster will be assigned to all members of the host cluster with the same SCSI ID.

A typical use-case is to define a host cluster containing all the WWPN's belonging to the hosts participating in a host operating system based cluster, such as IBM PowerHA® or Microsoft Cluster Server (MSCS) and such.

The following commands have been added to deal with host clusters:

- ▶ **lshostcluster**
- ▶ **lshostclustermember**
- ▶ **lshostclustervolumemap**
- ▶ **mkhost** (modified to put host in a host cluster on creation)
- ▶ **rmhostclustermember**
- ▶ **rmhostcluster**
- ▶ **rmvolumehostclustermmap**

Note: Host clusters allows for the creation of individual hosts and adding them to a host cluster. Care must be taken to make sure that no loss of access occurs when transitioning to host clusters.

CLI only at the time of writing: As of this writing, host cluster operations have not yet been incorporated into the Spectrum Virtualize GUI.

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 the 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 improves performance.

For the more information about maximum host configurations and restrictions, see *V7.8.0 Configuration Limits and Restrictions* which are available at these websites:

Spectrum Virtualize

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009560>

Storwize V7000

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009561>

The most important reason to limit the number of paths that are available to a host from Spectrum Virtualize is for error recovery, failover, and failback purposes. The overall time for handling errors by a host is reduced. Also, 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 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 release 7.8 and earlier, this information is now consolidated into the IBM System Storage Spectrum Virtualize Information Center, which is available at this website:

Spectrum Virtualize

http://www.ibm.com/support/knowledgecenter/STPVGU_7.8.0/com.ibm.storage.svc.console.780.doc/svc_ichome_780.html

Storwize V7000

http://www.ibm.com/support/knowledgecenter/ST3FR7_7.8.0/com.ibm.storwize.v7000.780.doc/v7000_ichome_780.html

6.1.4 Host ports

When you are using host ports that are connected to the 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 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 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.

Use one host object to represent a cluster of hosts and use multiple WWPNs to represent the ports from all the hosts that share a set of volumes.

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, and the switching between them can cause unwanted overhead 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, and using remote FC port masking you can set which ports can be used for replication communication.

Using port masking is strongly recommended by IBM, as a mixed traffic of host, back-end, intracluster and replication could cause congestion and buffer to buffer credit exhaustion, and could possibly result in heavy degradation of performance in your Spectrum Virtualize environment otherwise.

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. And 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 will be allowed. So, in a local FC port map, a 0 means no node-to-node traffic will happen, and a 0 on the remote FC port masking means no replication traffic will happen on that port. That means if a port has a 0 on both local and remote FC port masking, only host/back-end traffic will be allowed on it. The port mask can vary depending on the number of ports your Spectrum Virtualize HBA cards have. For an example of port distribution and masks on nodes 2145-CG8 and 2145-DH8, see Figure 6-1.

Slot/Port	Port #	SAN	8-port Nodes with 2 port cards	8-port Nodes with 4 port cards	12-port Nodes	16-port Nodes
S1P1	1	A	Host/Storage	Host/Storage	Host/Storage	Host/Storage
S1P2	2	B	Host/Storage	Host/Storage	Host/Storage	Host/Storage
S1P3	3	A	--	Inter-node	Host/Storage	Host/Storage
S1P4	4	B	--	Host/Storage or Replication*	Host/Storage	Host/Storage
S2P1	5	A	Host/Storage	Host/Storage	Inter-node	Inter-node
S2P2	6	B	Host/Storage	Host/Storage	Inter-node	Inter-node
S2P3	7	A	--	Host/Storage or Replication*	Host/Storage or Replication*	Host/Storage or Replication*
S2P4	8	B	--	Inter-node	Host/Storage	Host/Storage
S3P1	9	A	Inter-node		Host/Storage	Host/Storage
S3P2	10	B	Host/Storage or Replication*		Host/Storage or Replication*	Host/Storage or Replication*
S3P3	11	A	--		Inter-node or Host Storage	Inter-node or Host Storage
S3P4	12	B	--		Inter-node or Host Storage	Inter-node or Host Storage
S5P1	13	A	Host/Storage or Replication*			Host/Storage
S5P2	14	B	Inter-node			Host/Storage
S5P3	15	A	--			Host/Storage
S5P4	16	B	--			Host/Storage
localfcportmask			10010000	10010000	110000110000	110000110000
remotefcportmask			1100000	1100000	1001000000	1001000000
* Use for Host/Storage in case no replication is in place.						
** This applies to CG8 and DH8 nodes.						
*** Do not mix intracluster, replication and Host/Storage traffic. Use dedicated ports for each purpose.						

Figure 6-1 Port masking configuration on 2145-CG8 and 2145-DH8 nodes

For an example of port masking on 2145-SV1 nodes, see Figure 6-2.

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 11/ No Rep 1111	10000001	110010000001	110010000001
remotefcportmask			1100	10010	10010	10010

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

Figure 6-2 Port Masking for 2145-SV1

How to set portmask 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, go to **Settings** → **Network** → **Fibre Channel Ports**. Then you can select the use of a port. Setting *none* means no node-to-node and no replication traffic is allowed, only host and storage traffic will be allowed; *local* means only node-to-node traffic will be allowed, *remote* means only replication traffic will be allowed. See Figure 6-3.

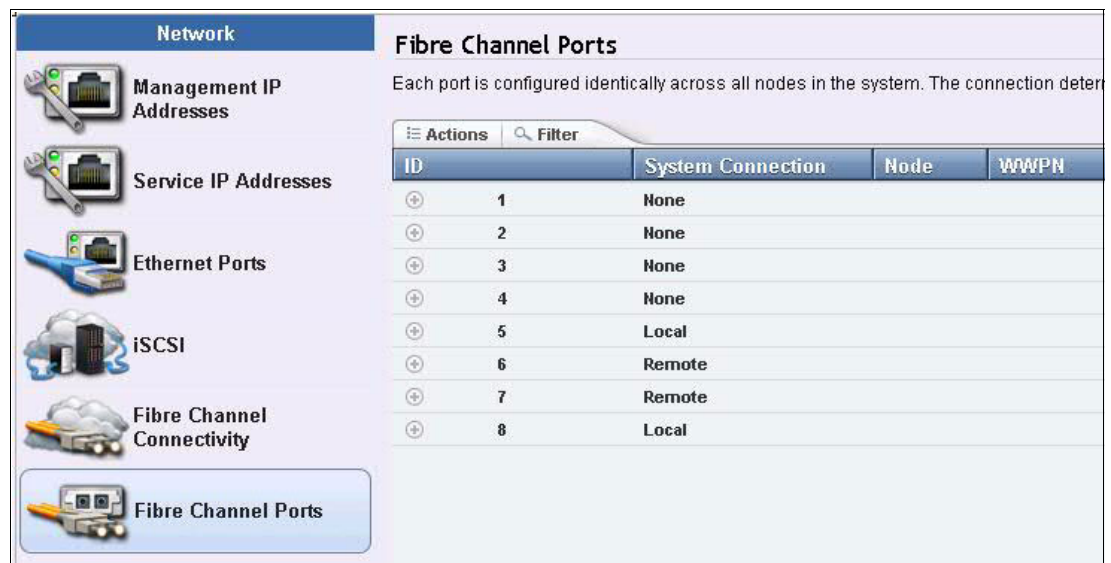


Figure 6-3 Fibre Channel ports system connection

6.1.6 Host to I/O group mapping

An *I/O grouping* consists of two 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 the 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 upon the switch fabric type. For more information about the maximum configurations, see V7.8 Configuration Limits and Restrictions, which are available at:

Spectrum Virtualize

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009560>

Storwize V7000

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009561>

Occasionally, a powerful host can benefit from spreading its volumes across I/O groups for load balancing. Start with a single I/O group, and use the performance monitoring tools, such as Spectrum Control, to determine whether the host is I/O group-limited. If more I/O groups are needed for bandwidth, you can use more host ports to allocate to the other I/O group.

For example, start with two HBAs zoned to one I/O group. To add bandwidth, add two more HBAs and zone to the other I/O group. The host object in the Spectrum Virtualize contains both sets of HBAs. The load can be balanced by selecting which host volumes are allocated to each volume. Because volumes are allocated to only a single I/O group, the load is then spread across both I/O groups that are based on the volume allocation spread.

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 Spectrum Virtualize software automatically assigns the next available ID if none is specified. Also, 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.

If you are using image mode to migrate a host to the 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 lshostvdiskmap command

```
svcinfo lshostvdiskmap -delim
```

Example 6-2 shows the results of using the **lshostvdiskmap** command.

Example 6-2 Output of using the lshostvdiskmap command

```
svcinfo lsvdiskhostmap -delim : EEXCLS_HBin01 id:name:SCSI_id:host_id:host_name:wwpn:vdisk_UID
950:EEXCLS_HBin01:14:109:HDMCENTEX1N1:10000000C938CFDF:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:109:HDMCENTEX1N1:10000000C938D01F:600507680191011D4800000000000466
950:EEXCLS_HBin01:13:110:HDMCENTEX1N2:10000000C938D65B:600507680191011D4800000000000466
950:EEXCLS_HBin01:13:110:HDMCENTEX1N2:10000000C938D3D3:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:111:HDMCENTEX1N3:10000000C938D615:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:111:HDMCENTEX1N3:10000000C938D612:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:112:HDMCENTEX1N4:10000000C938CFBD:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:112:HDMCENTEX1N4:10000000C938CE29:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:113:HDMCENTEX1N5:10000000C92EE1D8:600507680191011D4800000000000466
950:EEXCLS_HBin01:14:113:HDMCENTEX1N5:10000000C92EDFFE:600507680191011D4800000000000466
```

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 used 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 (SDDDSM), the **datapath query device** command shows the `vdisk_UID` (unique identifier), which enables easier management of volumes. The equivalent command for 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:ITSOCL1:admin>svcinfo lshostvdiskmap senegal
```

id	name	SCSI_id	vdisk_id	vdisk_name	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 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: 60050768018101BF28000000000000B5
=====
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: 60050768018101BF28000000000000B1
=====
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: 60050768018101BF28000000000000B2
=====
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: 60050768018101BF28000000000000B3

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: 60050768018101BF28000000000000B4

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: 60050768018101BF28000000000000A8

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: 60050768018101BF28000000000000A9

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 Spectrum Virtualize cluster access on two different I/O busses to maximize the availability and performance. When purchasing a server, always prefer to have two cards instead of one, for instance having two dual port HBA cards instead of one quad port HBA card, as in this way you can spread the I/O and keep the redundancy.

6.2 N-Port Virtualization ID (NPIV)

The usage model for all 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 given volume. These nodes each have their own Fibre Channel WWNN, and thus 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 will go offline, and it is up to the native OS multipathing software to failover from using both sets of WWPNS to just those that remain online. While this 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.

NPIV on Spectrum Virtualize is a feature released in V7.7. The NPIV feature aims to provide an availability improvement for the hosts that are connected to the Spectrum Virtualize nodes. It creates a virtual WWPN that will be available only for host connection, and in case of a node assert, failure, reboot or service mode, the virtual WWPN from that node will be transferred to the other node in the iogrp, to the same port. That means instead of having the host lose the connection to the Spectrum Virtualize node, the connection will remain active and the multipath software will 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 Spectrum Virtualize.

NPIV will work in a symmetric way, that means the NPIV port from node 1 port 1 will have the failover on node 2 port 1. For NPIV to work properly, you must have a symmetric cabling of Spectrum Virtualize in your switch, for instance: odd ports on Fabric 1, even Ports on Fabric 2 or vice versa. In short, you must have the ports that will 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.

NPIV is native on new deployments running V7.8. You can disable the NPIV feature at installation, although IBM strongly recommends you do not do this. If you have an existing 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 will report up to three virtual WWPNs as shown in Table 6-1.

Table 6-1 Spectrum Virtualize NPIV Ports

NPIV port	Port description
Primary NPIV Port	This is the WWPN that will communicate with backend storage only.
Primary Host Attach Port	This is the WWPN that will communicate with hosts. It is a target port only, and this is the primary port, therefore represents this local nodes WWNN.
Failover Host Attach Port	This is a standby WWPN that will communicate with hosts and will only be brought online on this node, if the partner node in this I/O Group goes offline. This will be the same as the Primary Host Attach WWPN on the partner node.

Then, when NPIV effectively goes into action, we can see a situation such as that illustrated in Example 6-6:

Example 6-6 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: 10online
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; 0x000000003
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; 0x000000003
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: 10online
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 might exist across the SAN fabric from the hosts to the Spectrum Virtualize nodes that present the volume. Most operating systems present each path to a volume as a separate storage device. Therefore 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 *V7.8 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels*, which are available at these websites:

Spectrum Virtualize
<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009560>

Storwize V7000
<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009561>

6.3.2 Preferred path algorithm

I/O traffic for a particular volume is, at any one time, 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 a 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. Keep in mind performance is enhanced by accessing the volume through the preferred node.

IBM multipathing software (SDDPCM, or SDDDSM) uses 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 Asymmetric Logical Unit Access (ALUA) compliant. That means, the host multipath driver must understand ALUA in order to achieve the best performance, resilience and availability from Spectrum Virtualize/Storwize. As long as the multipathing driver understands ALUA, it will apply the load balance multipath policy only in the paths belonging to the preferred node. If the multipath driver does not understand ALUA, it will spread 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, Spectrum Virtualize/Storwize sends the data using the intracluster/node-to-node connection in order to have the operation executed by the preferred node. The specific behavior can take place when Spectrum Virtualize Enhanced Stretched Cluster or HyperSwap is implemented. For more detailed information refer to:

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 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 16 device random 4 Kb read miss response time (4.2 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).

Table 6-3 16 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 (4.1 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 (4.2 nodes, microseconds)

Preferred node (owner)	Nonpreferred node	Delta
5,346	5,433	87

IBM SDDDSM, and SDDPCM software recognize the preferred nodes and use the preferred paths.

6.3.4 Path management

Spectrum Virtualize design is based on multiple path access from the host to both 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 just before any SAN maintenance or software upgrades. With IBM multipathing software (SDDPCM, and SDDDSM), this monitoring is done by using the **datapath query device** or **pcmpath query device** commands.

Fast node reset

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 improved from several minutes to approximately 30 seconds for the standard node reset.

Node reset behavior in Spectrum Virtualize V4.2 and later

When a Spectrum Virtualize node is reset, the node ports do *not* disappear from the fabric. Instead, the node keeps the ports alive. From a host perspective, Spectrum Virtualize stops responding to any SCSI traffic. Any query to the switch name server finds that the Spectrum Virtualize ports for the node are still present, but any FC login attempts (for example, PLOGI) are ignored. This state persists for about 30 - 45 seconds.

This improvement is a major enhancement for host path management of potential double failures. Such failures can include a software failure of one node where the other node in the I/O group is being serviced or software failures during a code upgrade. This new feature also enhances path management when host paths are misconfigured and includes only a single Spectrum Virtualize node.

6.3.5 Non disruptive Volume migration between I/O groups

Attention: These migration tasks can be nondisruptive if they are performed correctly and hosts that are mapped to the volume support nondisruptive 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 nondisruptive 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 can be done 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 moved 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, select **Volumes** → **Volumes**. In the Volumes panel, select the volume that you want to move and select **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 panels.

In the following example, we move VDisk ndvm to another I/O group nondescriptively by using Redhat Enterprise Linux 6.5 (Default Kernel).

Example 6-7 shows the Redhat Enterprise Linux 6.5 before I/O group migration. For this example, the Storwize V7000/Spectrum Virtualize caching I/O group is io_grp0.

Example 6-7 Native Linux multipath display before I/O group migration

```
[root@RHEL_65 ~]# multipath -ll
mpathb (3600507680281005500000000000000fd) 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 movevdisk -iogrp io_grp1 ndvm
svctask addvdiskaccess -iogrp io_grp1 ndvm
```

2. Detect the new paths to the volume in the destination I/O group, as shown in Example 6-8.

Example 6-8 SCSI rescan command on Redhat 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 Redhat Enterprise Linux 6.5, as shown in Example 6-9.

Example 6-9 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 sdl 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 we validate that the new paths are detected, we can safely remove access from the old I/O group by running the following command:

```
svctask rmvdiskaccess -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 successfully removed, as shown in Example 6-10.

Example 6-10 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

```

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 the 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 groups. 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 57 and “Queue depth in Fibre Channel hosts” topic in the 7.8 Knowledge Center, which are available at these websites:

Spectrum Virtualize

http://www.ibm.com/support/knowledgecenter/STPVGU_7.8.0/com.ibm.storage.svc.console.780.doc/svc_ichome_780.html

Storwize V7000

http://www.ibm.com/support/knowledgecenter/ST3FR7_7.8.0/com.ibm.storwize.v7000.780.doc/v7000_ichome_780.html

6.5 Host clustering and reserves

To prevent hosts from sharing storage inadvertently, establish a storage reservation mechanism. The mechanisms for restricting access to Spectrum Virtualize volumes use the SCSI-3 persistent reserve commands or the SCSI-2 legacy reserve and release commands.

The host software uses several methods to implement host clusters. These methods require sharing the volumes on 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 a legacy 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 functionality that is provided by the persistent reserve commands is a superset of the legacy reserve or release commands. The persistent reserve commands are incompatible with the legacy reserve or release mechanism. Also, target devices can support only reservations from the legacy mechanism or the new mechanism. Attempting to mix persistent reserve commands with legacy reserve or release commands results in the target device returning a reservation conflict error.

Legacy 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**: Legacy 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 ODM for a **reserve_policy** and a **PR_key_value** and then opens the device appropriately. 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 Spectrum Virtualize volume or on the Spectrum Virtualize MDisk during migration into 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 Spectrum Virtualize volume reserves by removing all the host mappings when Spectrum Virtualize code is at V4.1 or later.

Example 6-11 shows a failing `pcmquerypr` command to clear the reserve and the error.

Example 6-11 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 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 2.6.8.0 FOR AIX*, which is available at this website:

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S4001363#ESS>

6.5.2 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 Spectrum Virtualize. You cannot clear a leftover reserve on a Spectrum Virtualize MDisk from 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 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 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 a large number of 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 a 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 in size, and for other allowable values of the max_xfer_size attribute, the memory area is 128 MB in size.

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-01.ibm.com/support/docview.wss?uid=ssg1S4001363#SVC>

SDDPCM and AIX MPIO are continually improved since their release. You must be at the latest release levels of this software.

You do 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

The SDDPCM code of 2.1.3.0 and later features improvements in failed path reclamation by a health checker, a fallback error recovery algorithm, FC dynamic device tracking, and support for a SAN boot device on MPIO-supported storage devices.

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 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 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 VCSI disk 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 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 with the latest Spectrum Virtualize-supported levels, use MPIO with SDDPCM. For more information, see *V7.8 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels*, available at:

Spectrum Virtualize

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009558>

Storwize V7000

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009559>

For more information about VIOS, see this website:

<http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/faq.html>

One common question is 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.” That means, 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. The device cannot be used “as is.” 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 for new disk installations only.

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 SDDDSM from this website:

<http://www-01.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 that are described in this article are translated 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 the spectrum Virtualize Information Center on the link below:

<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 95.

6.8.3 Guidelines for disk alignment by using Windows with Spectrum Virtualize volumes

You can set preferred settings for best performance with Spectrum Virtualize when you use Microsoft Windows prior to 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 “V7.8 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels, available at:

Spectrum Virtualize

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009558>

Storwize V7000

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009559>

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 at this website:

http://www.ibm.com/support/knowledgecenter/STPVGU_7.8.0/com.ibm.storage.svc.console.780.doc/svc_dmpsetuplinux431_3yd13v.html

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 57. In addition, the Spectrum Virtualize Information Center provides maximums per HBA adapter or type. For more information, see this website:

Spectrum Virtualize

http://www.ibm.com/support/knowledgecenter/STPVGU_7.8.0/com.ibm.storage.svc.console.780.doc/svc_ichome_780.html

Storwize V7000

http://www.ibm.com/support/knowledgecenter/ST3FR7_7.8.0/com.ibm.storwize.v7000.780.doc/v7000_ichome_780.html

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 Information Center, which is available at this website:

Spectrum Virtualize

http://www.ibm.com/support/knowledgecenter/STPVGU_7.8.0/com.ibm.storage.svc.console.780.doc/svc_ichome_780.html

Storwize V7000

http://www.ibm.com/support/knowledgecenter/ST3FR7_7.8.0/com.ibm.storwize.v7000.780.doc/v7000_ichome_780.html

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:

<http://www.ibm.com/developerworks/linux/library/l-lpic1-v3-104-2/index.html?ca=dgr-lnxw06TracjLXFilesystems>

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 “V7.8 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels, available at:

Spectrum Virtualize

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009558>

Storwize V7000

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009559>

IBM SDD is no longer supported because its features are now available natively in the multipathing driver 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 a combined file system and logical volume manager designed by Sun Microsystems, called ZFS. It uses MPxIO and is inbound to the Solaris 11, being an native option to Veritas Volume Manager. For more information about Oracle ZFS see the links below:

<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 Spectrum Virtualize code level V4.3.

If you want to run MPxIO on your Sun SPARC host, configure your 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 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 Spectrum Virtualize requires for error recovery.

Use the commands that are shown in Example 6-12 on page 252 to determine the basic configuration of a Symantec Veritas server.

Example 6-12 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=cxtxdxs2' (lists paths by LUN)
```

The commands that are shown in Example 6-13 and Example 6-14 determine whether the Spectrum Virtualize is properly connected and show at a glance, which ASL is used (native DMP ASL or SDD ASL).

Example 6-13 shows what you see when Symantec Volume Manager correctly accesses Spectrum Virtualize by using the SDD pass-through mode ASL.

Example 6-13 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_SANVC0 VPATH_SANVC 0200628002faXX00 CONNECTED
```

Example 6-14 shows what you see when Spectrum Virtualize is configured by using native DMP ASL.

Example 6-14 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_VC0 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 **vxdtl enable** command. To list the ASLs that are active, enter the following command:

```
vxddladm listsupport
```

6.10.4 Troubleshooting configuration issues

Example 6-15 shows that the appropriate ASL is not installed or the system is enabling the ASL. The key is the enclosure type **OTHER_DISKS**.

Example 6-15 Troubleshooting ASL errors

vxddmpadm 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

vxddmpadm 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 “V7.8 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels, available at:

Spectrum Virtualize

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009558>

Storwize V7000

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009559>

On this website, you can also find information about the available support in V7.8 of VMware vStorage API for Array Integration (VAAI).

Spectrum Virtualize V7.2 and later supports VMware vStorage APIs. 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 and 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 7.8 Information Center at this website:

Spectrum Virtualize

http://www.ibm.com/support/knowledgecenter/STPVGU_7.8.0/com.ibm.storage.svc.console.780.doc/svc_ichome_780.html

Storwize V7000

http://www.ibm.com/support/knowledgecenter/ST3FR7_7.8.0/com.ibm.storwize.v7000.780.doc/v7000_ichome_780.html

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 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 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 failback 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 on, VMware multipath driver fully supports Spectrum Virtualize/Storwize V7000 ALUA preferred path algorithms. VMware administrators should choose Round Robin and validate that VMW_SATP_ALUA is displayed. This reduces operational overhead 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 are supported

- 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 maxima:

<https://www.vmware.com/pdf/vsphere5/r55/vsphere-55-configuration-maximums.pdf>

<https://www.vmware.com/pdf/vsphere6/r60/vsphere-60-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 6-16 on page 255 shows the output of the **datapath query devstats** command for two devices.

Example 6-16 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:	<= 512 271	<= 4k 2337858	<= 16K 104	<= 64K 1166537	> 64K 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:	<= 512 296	<= 4k 27128331	<= 16K 215	<= 64K 3108902	> 64K 0

Also, the **datapath query adaptstats** adapter-level statistics command is available (mapped to the **pcmpath query adaptstats** command). Example 6-17 shows the use of two adapters.

Example 6-17 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; Windows has **perfmom.msc /s** tool.

Industry standard performance benchmarking tools are available by joining the Storage Performance Council. For more information about this council, see the Storage Performance Council page at this website:

<http://www.storageperformance.org/home>

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 function that is provided by the IBM Easy Tier feature of the 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 that 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 intelligently aligns the system with current workload requirements and optimizes the usage of SSDs or flash arrays.

This functionality 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 to achieve 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, but 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.

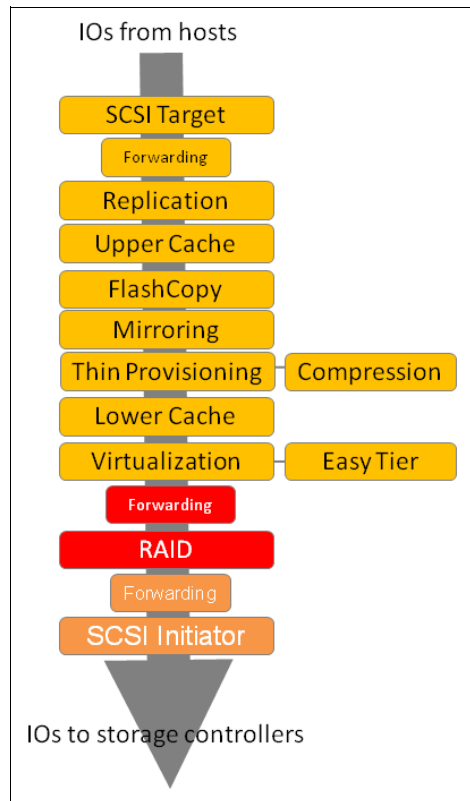


Figure 7-1 Easy Tier in the software stack

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, because it automatically migrates (or moves) extents that belong to a volume between different storage tiers (see Figure 7-2 on page 260) or the same storage tier (see Figure 7-7 on page 264). As 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 unnoticed 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 on page 260 shows the basic Easy Tier principle of operation.

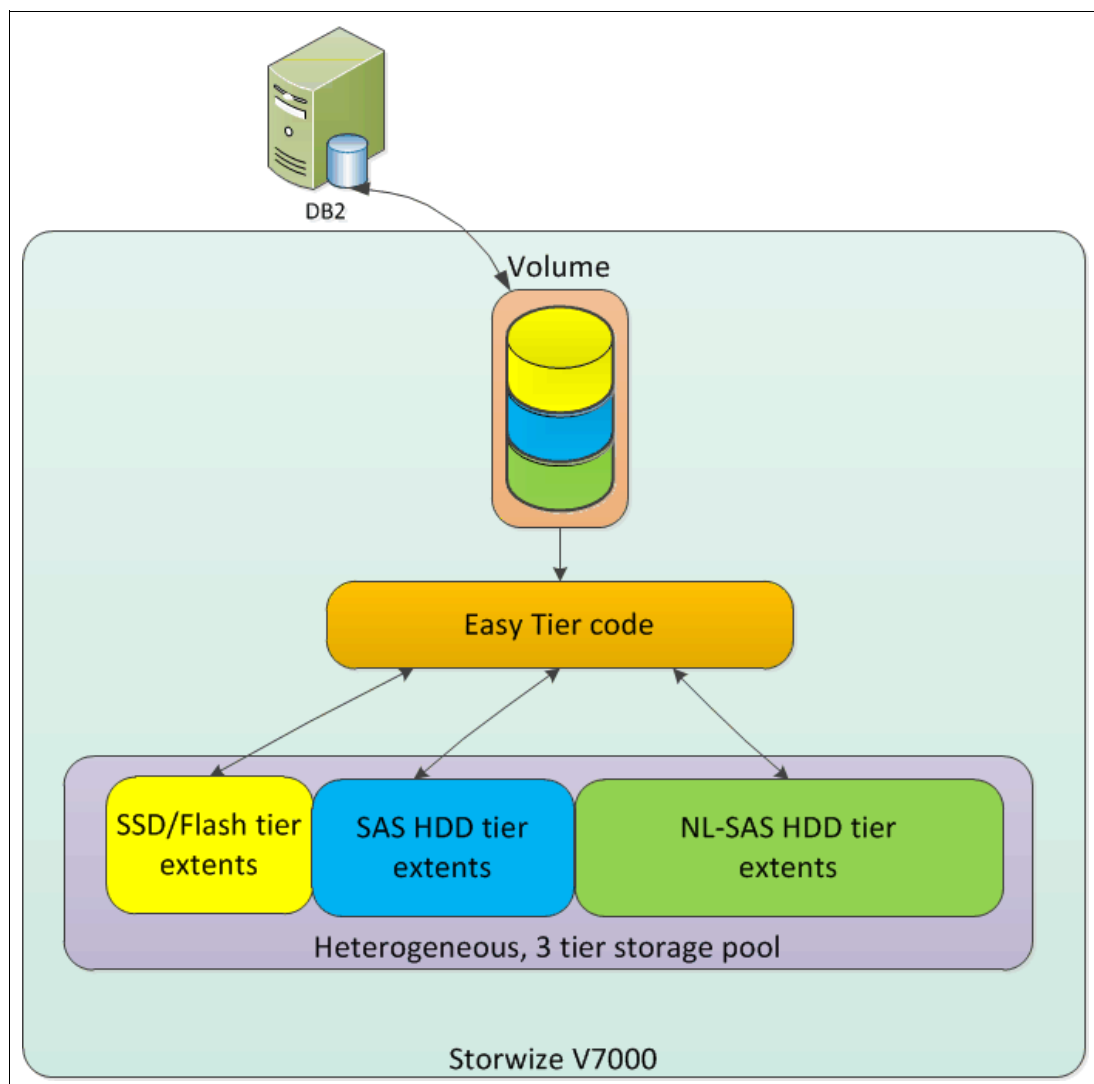


Figure 7-2 Easy Tier

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

Easy Tier tiering model has been modified with V7.8 adding a new tier to support Read-Intensive (RI) flash drives.

One of the reasons why flash technology is still quite expensive when compared to traditional HDD, is that an over provisioning of the physical memory is provided to mitigate the Write Amplification issue (https://en.wikipedia.org/wiki/Write_amplification). Read-Intensive

flash drives are lower-cost flash drives with the cost reduction being achieved by having less redundant flash material.

Read Intensive support for 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. The table in Figure 7-3 shows the possible combinations for the pool configuration with regard to the four user tiers (in orange we have highlighted the configurations containing the RI user tier).

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 reports in which Easy Tier tier each user tier is mapped. 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 *tier1_flash* 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 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 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.

Starting with SVC 2145-DH8 nodes and software version 7.3, up to two expansion drawers can be connected to the one 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 MDisks presented from external storage systems or RAID arrays made of HDDs. Because Spectrum Virtualize/Storwize does not know what kind of physical disks external MDisks 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 MDisks 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

[illegible]

```
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
6005076802880102c00000000000002000000000000000000000000000000 tier3_nearline no
2 SITE_B no no
2 mdisk2 online managed 1 POOL_V7K_SITEB 250.0GB 0000000000000002 V7K_SITEB_C2
6005076802880102c00000000000002100000000000000000000000000000 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 MDisks. To change the tier, complete the following steps:

1. Go to **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. Then, right-click the wanted MDisk and select **Modify Tier** (Figure 7-4).

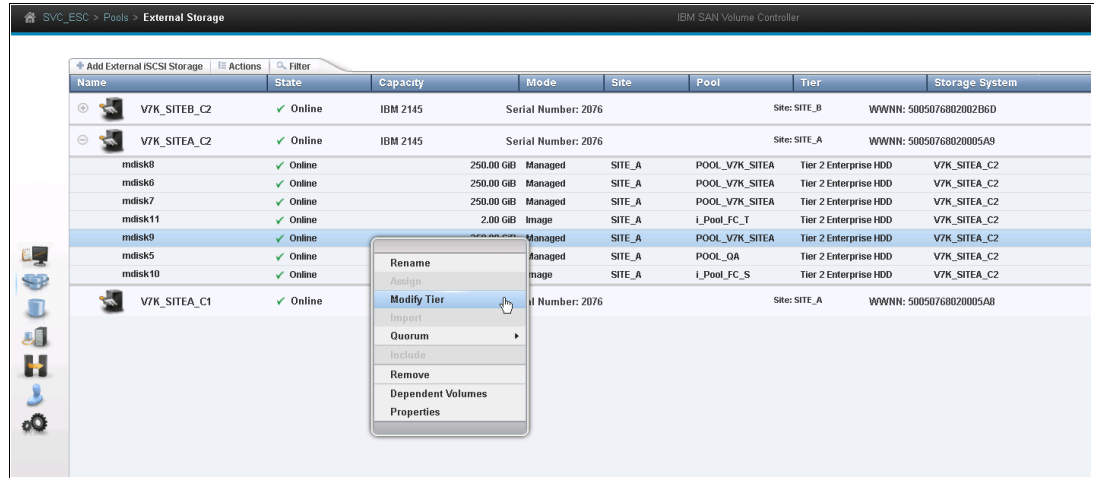


Figure 7-4 Change the MDisk tier

3. 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.

4. If you do not see the *Tier* column, right-click the blue title row and select the **Tier** checkbox, as shown in Figure 7-6.

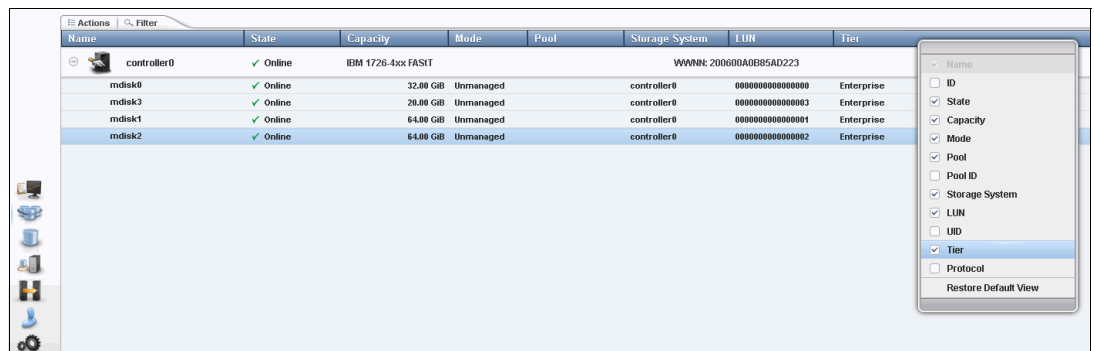


Figure 7-6 Customizing the title row to show the tier column

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 reside. 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 261, Spectrum Virtualize and Storwize systems do not automatically detect the type of external MDisks. Instead, all external MDisks initially are put into the enterprise tier by default. Then, the administrator must manually change the tier of MDisks and add them to storage pools. Depending on what type of disks are gathered to form a storage pool, we distinguish two types of storage pools: *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.

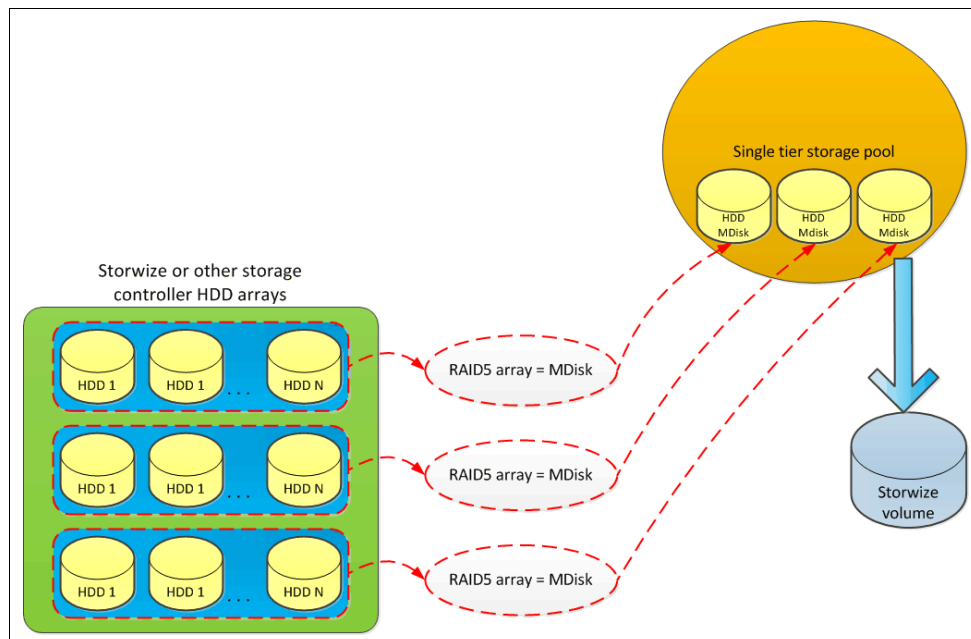


Figure 7-7 Single tier storage pool with striped volume

MDisks that are used in a single-tier storage pool should have the same hardware characteristics, for example, 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, for example, a storage pool that contains a mix of enterprise and SSD MDisks or enterprise and NL-SAS MDisks.

Figure 7-8 on page 265 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.

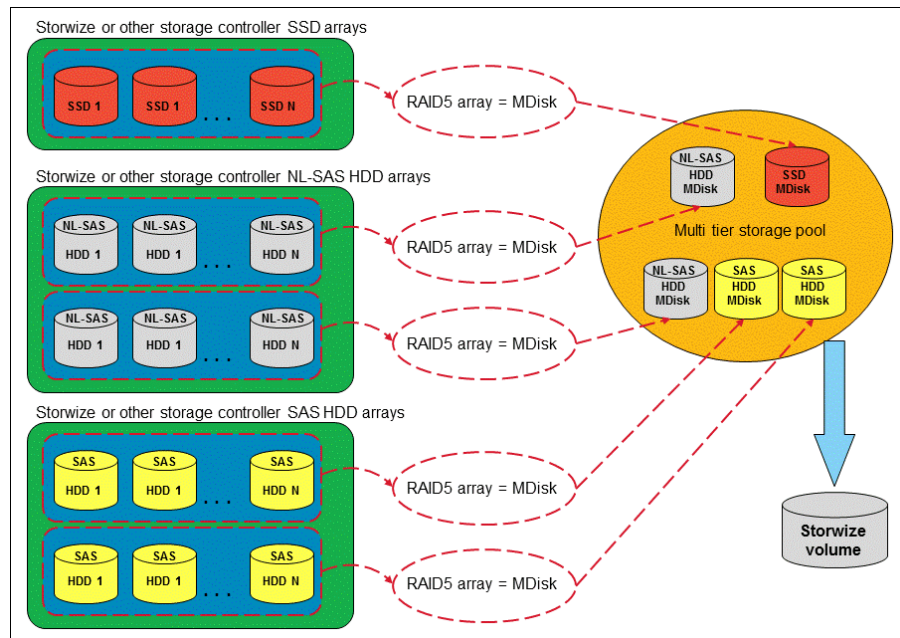


Figure 7-8 Multitier storage pool with striped volume

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, 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 to four	off	inactive (see note 2)
Off	Two to four	on	inactive (see note 2)
Measure	One	off	measured (see note 3)
Measure	One	on	measured (see note 3)
Measure	Two to four	off	measured (see note 3)
Measure	Two to four	on	measured (see note 3)
Auto	One	off	measured (see note 3)
Auto	One	on	balanced (see note 4)
Auto	Two to four	off	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
Auto	Two to four	on	active (see note 5)
On	One	off	measured (see note 3)
On	One	on	balanced (see note 4)
On	Two to four	off	measured (see note 3)
On	Two to 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 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 must 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:

- Automatic Data Relocation (ADR mode) plan to migrate extents across adjacent tiers

- 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 the higher tier to the lower tier. Demote occurs within adjacent tiers.
- ▶ Swap:

Exchanges cold extent in upper tier with hot extent in lower tier.
- ▶ Warm Demote:
 - Prevents performance overload of a tier by demoting a warm extent to the lower tier.
 - 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 five 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.
 - Moves hot extents from high used MDisks to low used MDisks.
 - 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 nearline tier and vice versa without moving them first to enterprise tier.

Easy Tier extent migration types are presented on Figure 7-9.

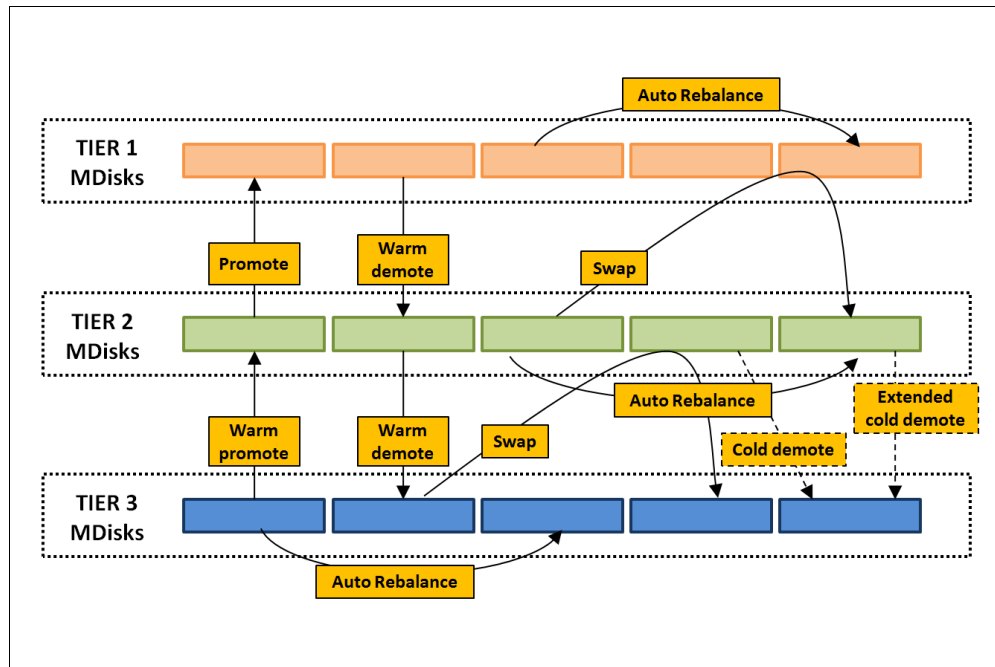


Figure 7-9 Easy Tier extent migration types

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 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 275. 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 269.

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 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 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.

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 re-enabled when the migration is complete.

7.2.2 Limitations

When you use Easy Tier on the 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 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** 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 new capacity to the pool, accelerating Easy Tier can quickly spread existing volumes onto the new MDisk.
- ▶ 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 under utilizing or over utilizing an external MDisk(s). This setting doesn't apply to internal MDisk(s) (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 that means that they don't take into account the actual backend configuration. For instance, the same performance profile is used for a DS8000 with 300GB 15K rpm and 1.8TB 10K rpm. This is why the user is allowed to change the Easy Tier load setting to better align it with 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 MDisk. 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 under utilized 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 MDisk.

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, that is the size of 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 contained in an extent have 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.

Also 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 and then 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 bandwidth consumption is.

On the other hand, Easy Tier efficiency should not be considered the only factor when choosing the extent size. Also manageability and capacity requirements considerations must be taken in account.

In general we recommend an extent size of either 256MB or 512MB for Easy Tier enabled configurations. With these extent sizes the maximum configurable capacity for a Spectrum Virtualize/Storwize system is respectively 1PB and 2PB. For systems with larger capacity requirements, bigger extent size must be used.

External controller tiering considerations

IBM Easy Tier is an algorithm that is developed by IBM Almaden Research and made available to many members of the IBM storage family, such as the DS8000, 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 Spectrum Virtualize and Storwize technology, such as Easy Tier Application, Easy Tier Heat Map Transfer and Easy Tier Control.

Before V7.3, 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 the Spectrum Virtualize/Storwize Easy Tier disabled.

With V7.3 and the introduction of the autorebalance function, Easy Tier can be effectively enabled at the Spectrum Virtualize/Storwize systems level. However we have to consider that Spectrum Virtualize is not aware about the tiering functions of an external controller, and vice versa. So each tiering function will make its decisions independently. What we need to avoid is rebalance over rebalance because this situation can cancel any performance benefits.

We can consider the following two options:

1. Easy Tier is done at Spectrum Virtualize/Storwize level. In this case at the backend level we must perform the following actions:

- Set up homogeneous pools according to the tier technology available.
- Create volumes to present to Spectrum Virtualize/Storwize from the homogeneous pool.
- Disable tiering functions.

At a Spectrum Virtualize/Storwize level we need to perform the following actions:

- Discover the MDisk provided by the backend storage and set the tier properly
- Create hybrid pools aggregating the MDisk
- Enable the Easy Tier function

2. Easy Tier is done at backend level. In this case at the backend level we must perform the following actions:

- Set up hybrid pools according to the tier technology available
- Create volumes to present to Spectrum Virtualize/Storwize from the hybrid pools
- Enable the tiering functions.

At Spectrum Virtualize/Storwize level we need to perform the following actions:

- Discover the MDisk provided by the backend storage and set the same tier for all
- Create standard pools aggregating the MDisk
- Disable the Easy Tier function

Even though both these options provide benefits in term of performance, they have different characteristics.

Option 1 provides some advantages when compared to option 2. One advantage is that Easy Tier can be enabled or disabled at volume level. This allows users to decide which volumes will benefit from Easy Tier and which will not. With option 2 this cannot be achieved. Another advantage of option 1 is that the volume heat map matches directly the host workload profile running toward the volumes. With option 2 the volume heat map on the backend storage is based on the Spectrum Virtualize/Storwize workload and this doesn't represent exactly the hosts workload profile because of the effects of the Spectrum Virtualize/Storwize caching. Finally, with option 1 we have the chance to change the extent size to improve the overall Easy Tier efficiency (as described in "Extent size considerations" on page 272).

On the other hand option 2, especially with DS8000 as backend, offers some advantages when compared to option 1. For example, when using external storage, 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 271. These profiles may not match exactly the actual backend capabilities leading sometimes to a not fully optimized resource utilization. With option 2 this mostly doesn't happen because the performance profiles are based on the real backend 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 before eventually achieving 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.

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 invoked just before the beginning of the disaster recovery test and very likely 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.

Tier sizing considerations

Tier sizing is a complex task that always requires an environment workload analysis to match the performance and costs expectations.

We can consider the following sample configurations that address some of 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 is a configuration with improved performance compared to a single tier solution and where 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 is 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 command prompt 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` 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

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 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 obmitted.....
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
```

Next, perform the normal PSCP **-load** download process, as shown in Example 7-6.

Example 7-6 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, click **Settings** → **Support** page, as shown in Figure 7-10. If the page is not displaying a list of individual log files, click **Show full log listing**.

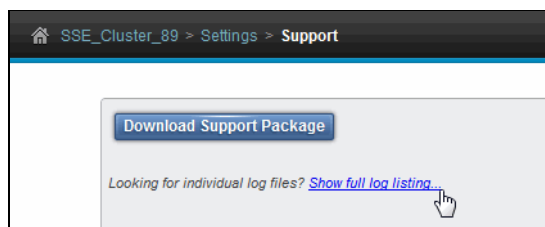


Figure 7-10 Accessing the “dpa_heat” file in the GUI

Next, right-click the row for the dpa_heat file and choose **Download**, as shown in Figure 7-11.

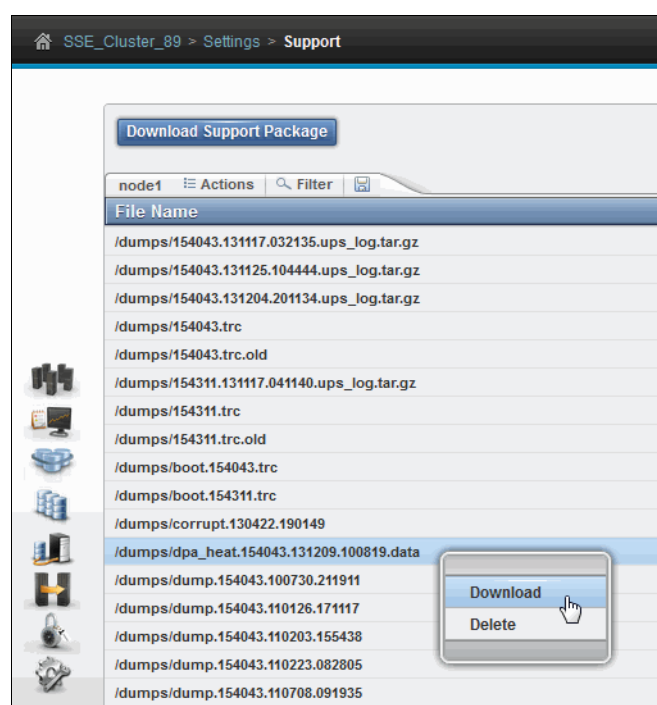


Figure 7-11 Downloading the dpa_heat file from the GUI

The file is downloaded to your local workstation.

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 page 277.

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.

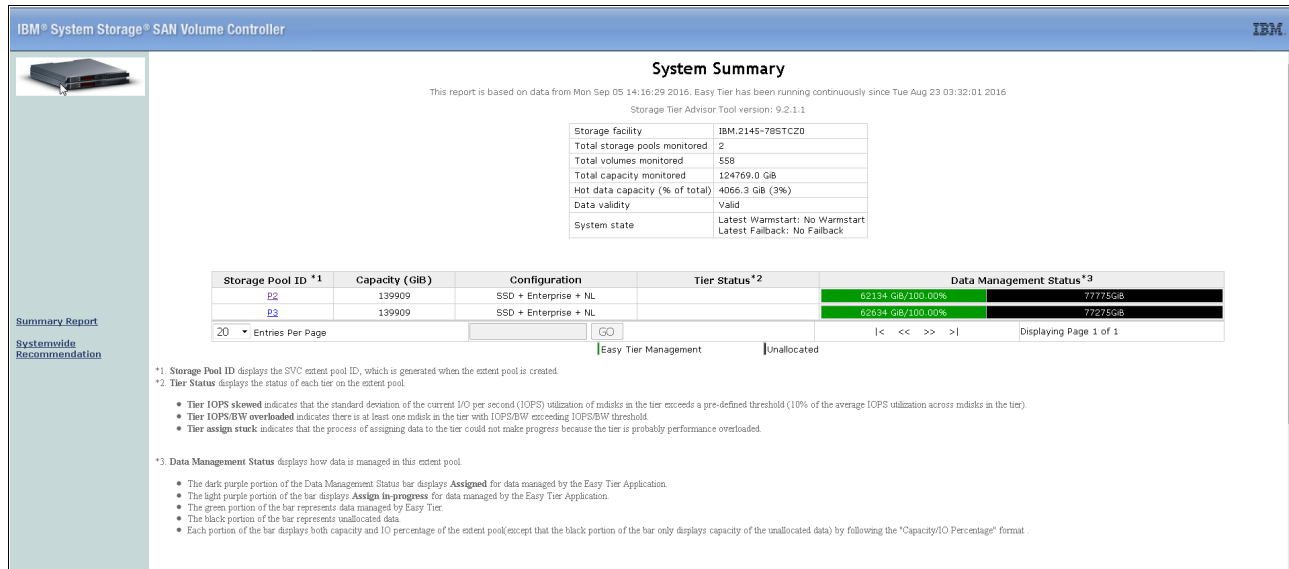


Figure 7-12 STAT main window

System Report

The System Summary window contains data that the Easy Tier monitor previously collected.

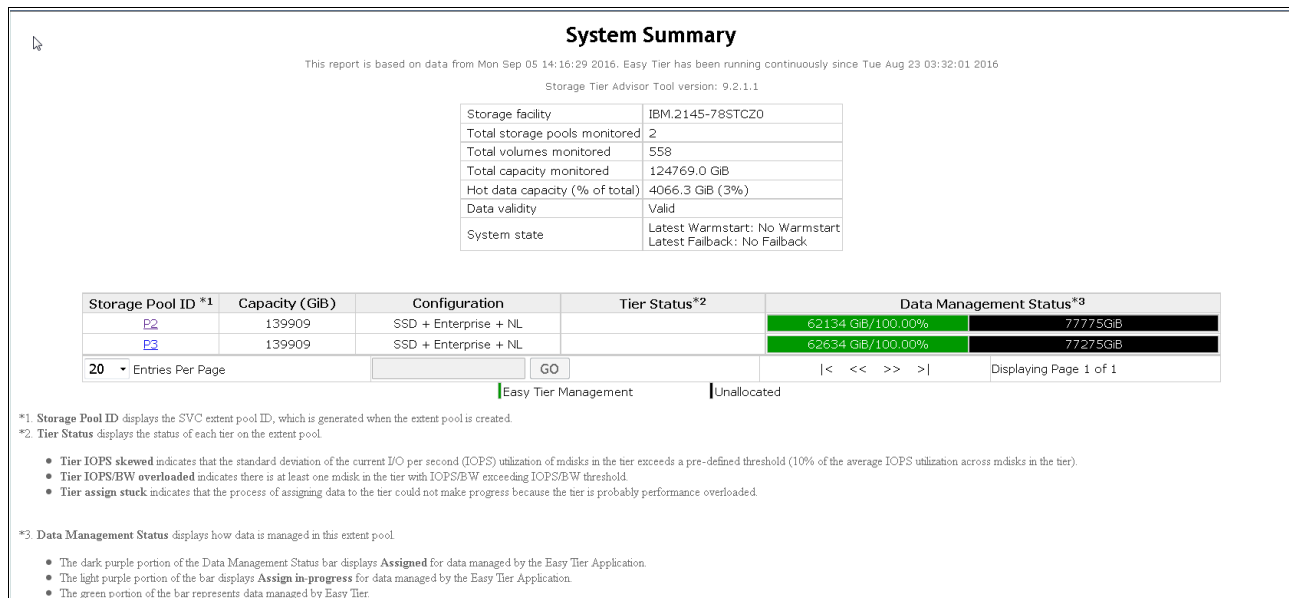


Figure 7-13 System Summary window

The System Summary window (Figure 7-13 on page 277) 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 unallocated 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 unallocated 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 off-load.
- 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 on page 279 shows the Systemwide Recommendation window, which is opened by selecting the Systemwide Recommendation link in the left panel.

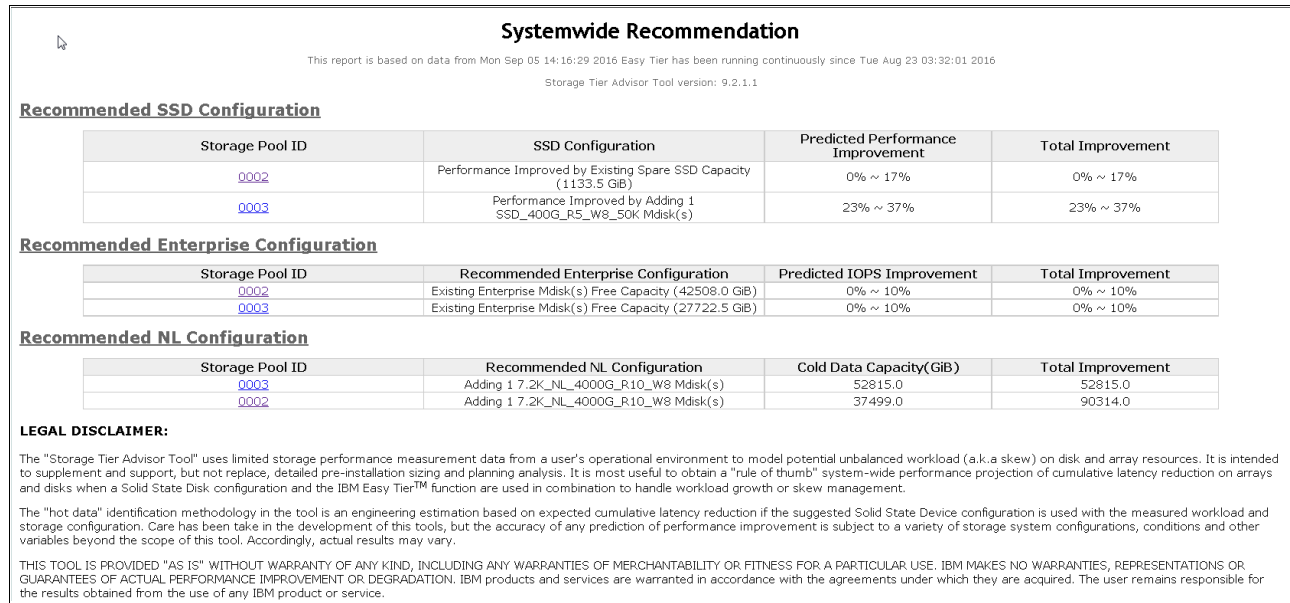


Figure 7-14 Systemwide Recommendation window

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 a 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 8 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 clicking on a pool ID the Storage Pool Performance Statistics and Improvement Recommendation window opens, as shown in Figure 7-15.

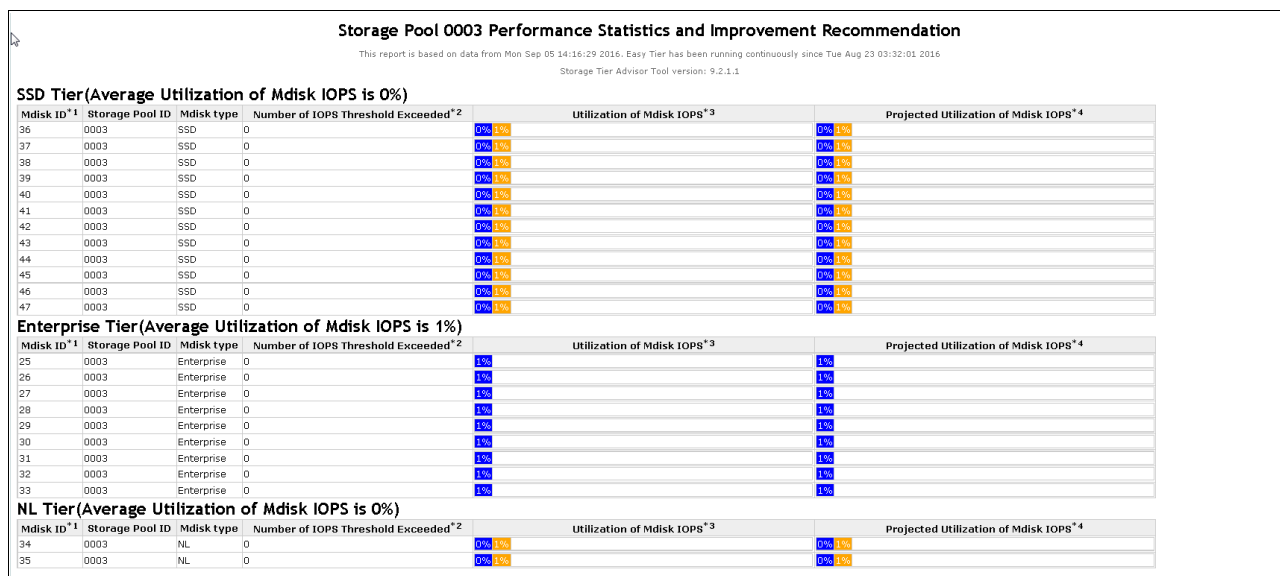


Figure 7-15 Storage Pool Performance Statistics and Improvement Recommendation for pool ID 0003

The table reported in the window shows the following characteristics for each MDisk in this pool:

- ▶ The MDisk ID and type.
- ▶ The number of IOPS thresholds exceeded, which 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, which represents, in three possible colored bars (blue, orange, and red), the current percentage of the maximum allowed IOPS threshold for the MDisk device type. The blue portion represents the percentage of IOPS below the tier average usage of the MDisk IOPS. The orange portion represents the percentage between the average and the maximum allowed IOPS for the MDisk IOPS. And, the red portion represents the percentage above the maximum allowed IOPS for the MDisk IOPS.
- ▶ The projected usage of the MDisk IOPS, which represents 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 on page 281. The Workload Distribution Across Tiers window includes the following components:

- ▶ The X-axis denotes the top x-intensive data that is 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.

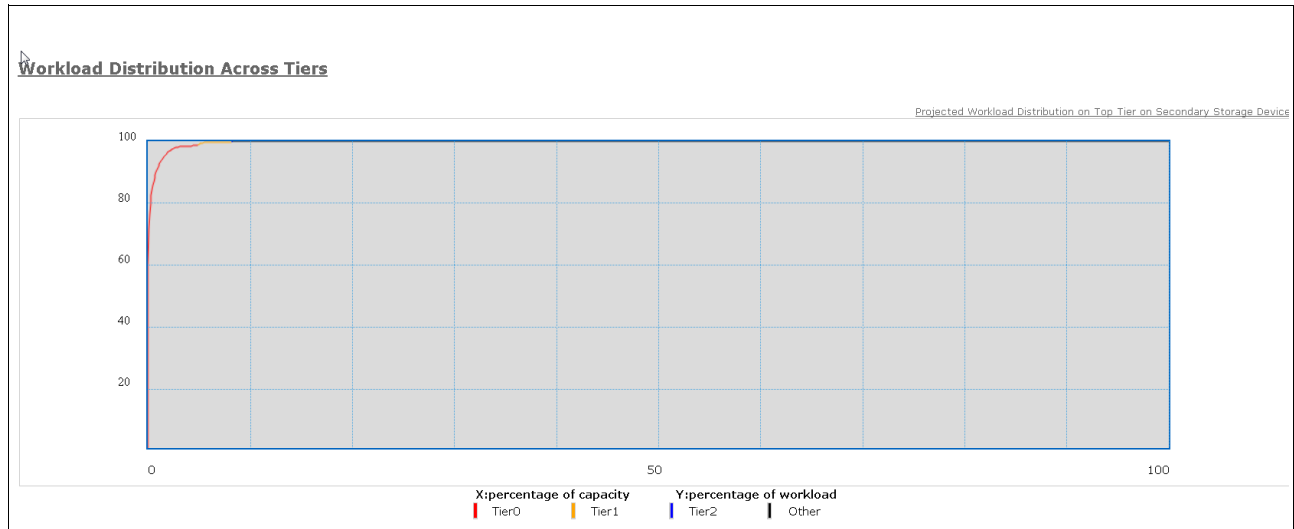


Figure 7-16 Workload Distribution Across Tier section

This report (pool workload distribution) uses the moving average of the small-block I/O only.

Recommended configurations

By selecting Recommended SSD/Enterprise/NL Configuration, 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_800G_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_800G_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:

- ▶ 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 279).

- ▶ 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.

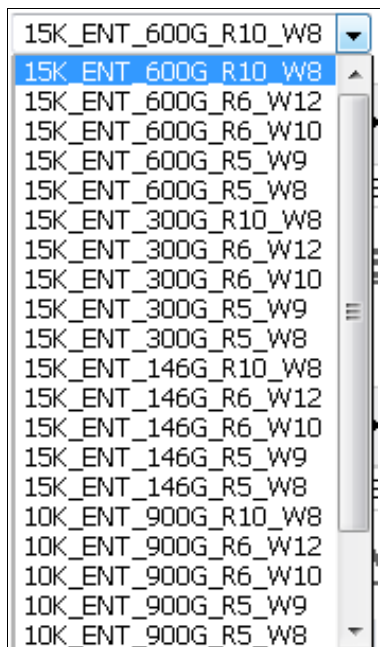


Figure 7-18 Drop-down enterprise menu

Volume Heat Distribution

By clicking Volume Heat Distribution, the heat distribution table opens, as shown on Figure 7-19 on page 283. 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 [nearline]), 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

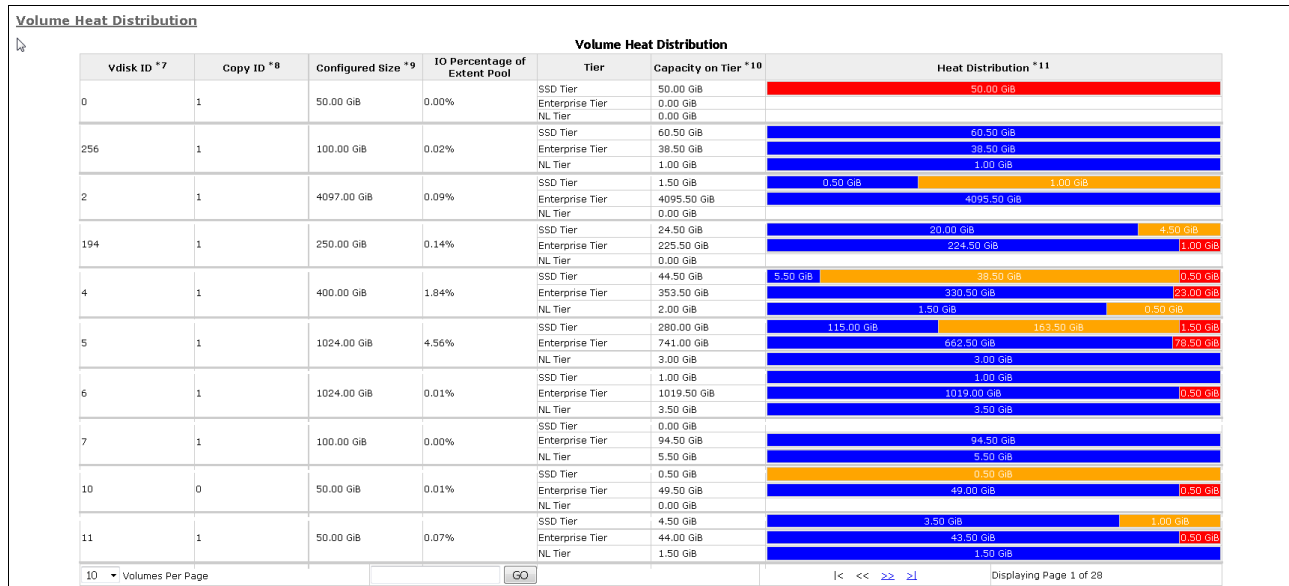


Figure 7-19 Volume Heat Distribution section

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 corresponds to data that is promoted to the higher tier if it is on a lower tier, after the hot data is promoted and the capacity is 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, but they might be moved onto nearline ranks (cold demotion), if applicable.

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       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
05/22/2014  10:15 AM      51,468 pool_rec_p00000.html
05/22/2014  10:15 AM      59,218 pool_rec_p00001.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 for Easy Tier

In addition to the STAT tool, IBM Spectrum Virtualize has another utility, which 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 Chart which 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.

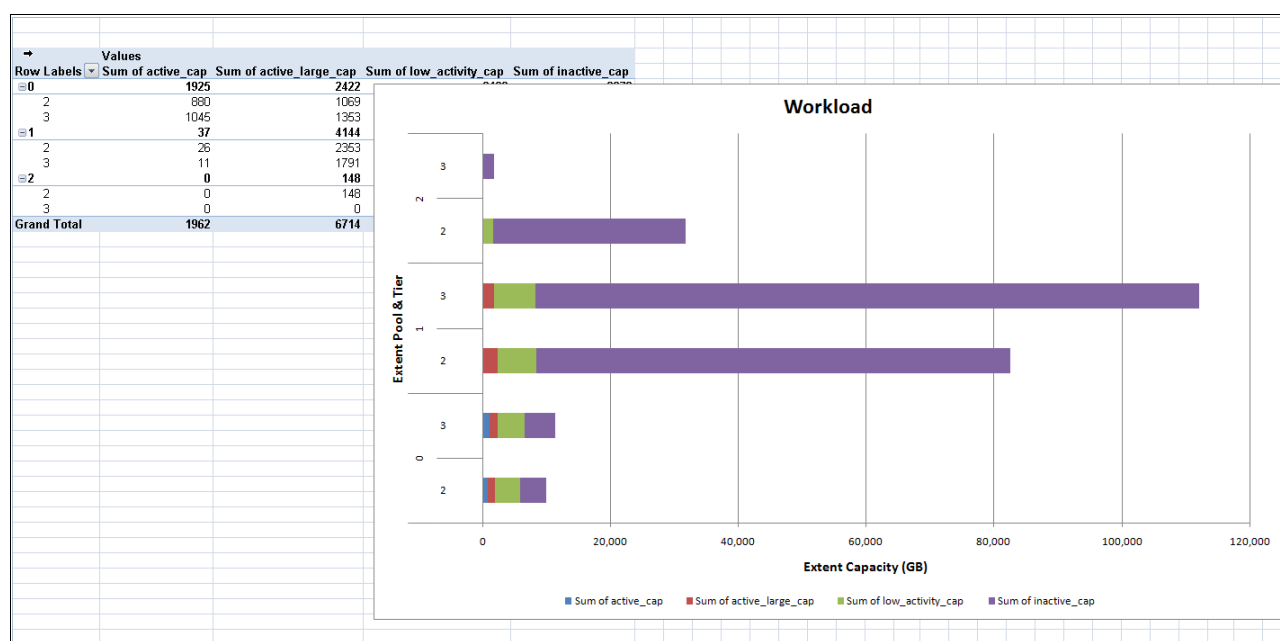


Figure 7-21 STAT Charting Utility Workload Categorization report

Using the pivot table features, you can obtain information at tier, pool and even single volume level. The data are classified in 4 type 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 < 57KiB for CKD and <64KiB for FB)
- *Active_large*: all data not classified above (transfer size >= 57KiB for CKD and >=64KiB for FB)

For each of these data type many statistics are available which can be used for a detailed workload analysis.

► Data Movement report

The data movement reports provides information about the extents moving activity in 5-minute intervals. The output is illustrated in Figure 7-22.

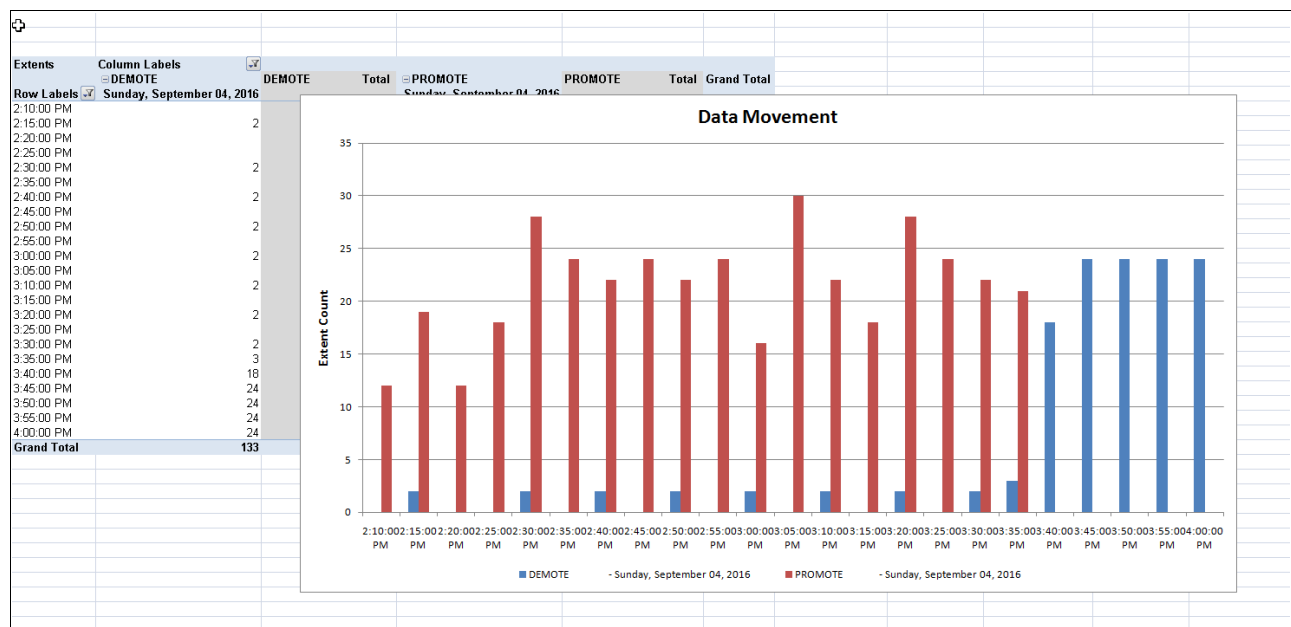


Figure 7-22 STAT Charting Utility Data Movement report

Using the pivot table features, you can obtain information at system and even single volume level. The data movement are classified in 5 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 on page 286.

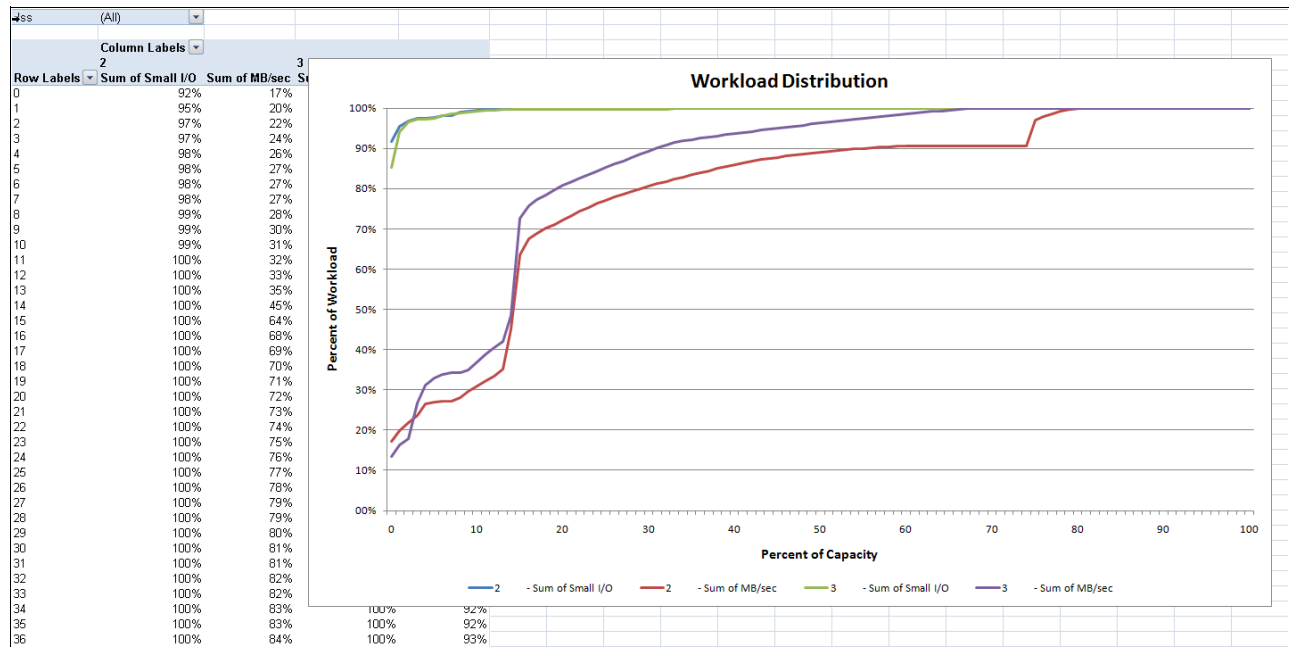


Figure 7-23 STAT Charting Utility Workload Skew report

This report provides detailed information about the workload distribution with reference to 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 aims to give you some suggestions and the basic concepts of how to start to implement a storage monitoring system for Spectrum Virtualize/Storwize using their specific functions or external IBM Tools.

8.1 Generic monitoring

With Spectrum Virtualize/Storwize you can implement generic monitoring using Spectrum Virtualize/Storwize specific functions that are integrated with the product itself without adding any external tools or cost.

8.1.1 Monitoring with GUI

The management GUI is the primary tool that is used to service your system.

Regularly monitor the status of the system 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 panel 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 there is an error that 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.

e-mail 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 the form of an event notification e-mail. 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 Simple Network Management Protocol (SNMP) messages sent by Spectrum Virtualize/Storwize.

Use this MIB file to configure a network management program to receive SNMP event notifications sent from a Spectrum Virtualize/Storwize system.

This MIB file is suitable for use with SNMP messages from all versions of Spectrum Virtualize/Storwize.

Spectrum Virtualize/Storwize MIB file can be downloaded at:

ftp://ftp.software.ibm.com/storage/san/sanvc/SVC_MIB_7.8.0.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 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 a Spectrum Virtualize/Storwize system, the space usage of Storage Pools and Thin Provisioned or Compressed Volumes can be monitored 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 disk 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 fit below the recommended new limit, you will be advised to do so. If volume capacity cannot be reduced to meet the recommended limit, you will be advised to create a non-compressed mirror of the data (if one does not already 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 it is for Spectrum Virtualize/Storwize as well.

The next sections aim to show what performance analysis tools are integrated with Spectrum Virtualize/Storwize systems, and what IBM external tool are available to collect performance statistics supplying an historical retention as well.

It is important to remember that performance statistics are not necessary only to debug or prevent some potential bottlenecks, but are required to make capacity planning for future growth easier.

8.2.1 Performance monitoring with the GUI

In 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 then can 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 impact performance.

The Volumes and MDisks graphs on the Performance panel 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 on page 291.

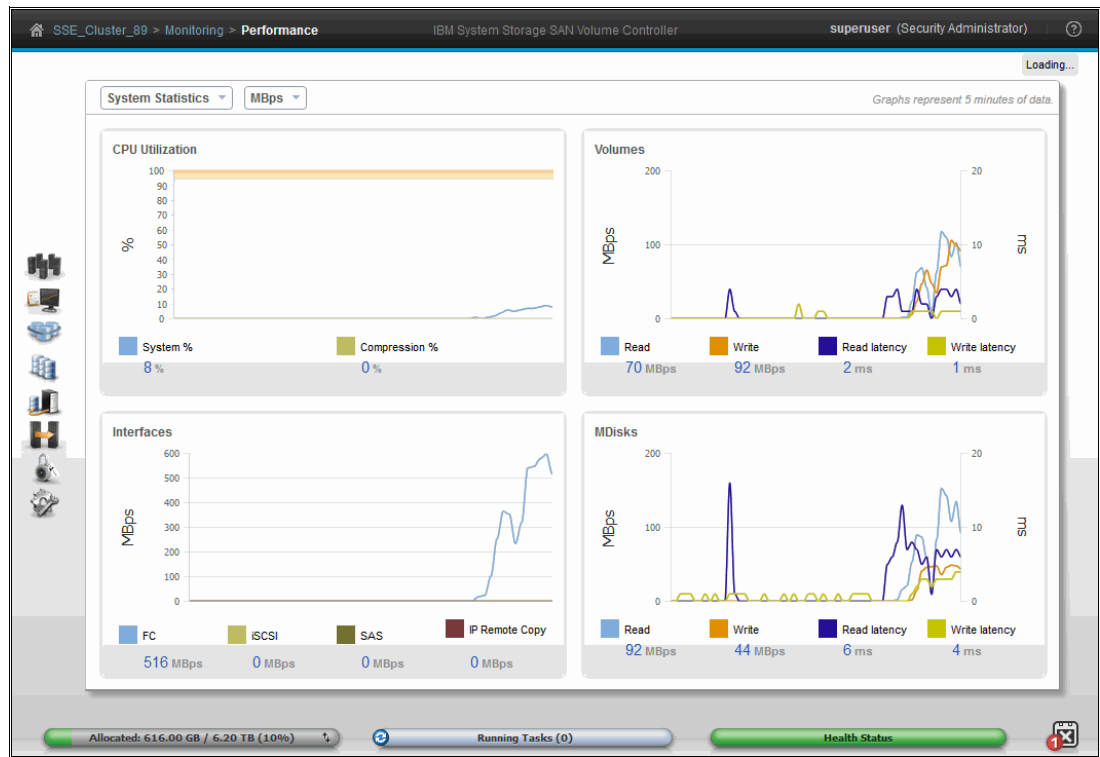


Figure 8-1 GUI example

Then you can choose the metrics you want to be displayed as shown in Figure 8-2.

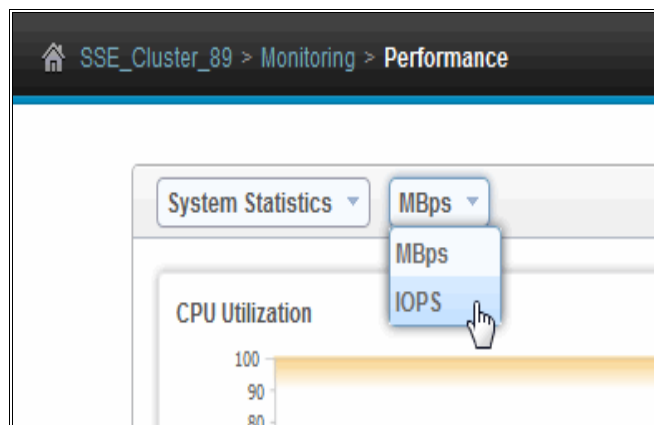


Figure 8-2 different metrics

8.2.2 Performance monitoring with IBM Spectrum Control

IBM Spectrum Control offers several reports that you can use to monitor Spectrum Virtualize/Storwize systems to identify performance problems. IBM Spectrum Control provides improvements to the web-based user interface that is designed to offer ease of use access to your storage environment.

IBM Spectrum Control provides a large amount of detailed information about Spectrum Virtualize/Storwize systems and in the next paragraph we will give you some basic suggestions about what metrics are strongly required 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 Tivoli Storage Productivity Center (including how to add a storage system), see this web site:

<http://www.ibm.com/support/docview.wss?uid=swg21986557>

and:

http://www.ibm.com/support/knowledgecenter/SS5R93_5.2.11/com.ibm.spectrum.sc.doc/fqz0_t_installing_main.html

Spectrum Control Dashboard

The performance dashboard provides Best Practice Performance Guidelines for the critical monitoring metrics.

These best practice guidelines do not represent the maximum operating limits of the related components, but are rather suggested limits 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 top of the chart will overlay the corresponding 24 hours for the previous day and 7 days prior. This allows for an immediate historical comparison of the respective metric.

Note that 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 which enables rapid detection of anomalous workloads and behaviors.

Figure 8-3 shows how to change the day of reference.

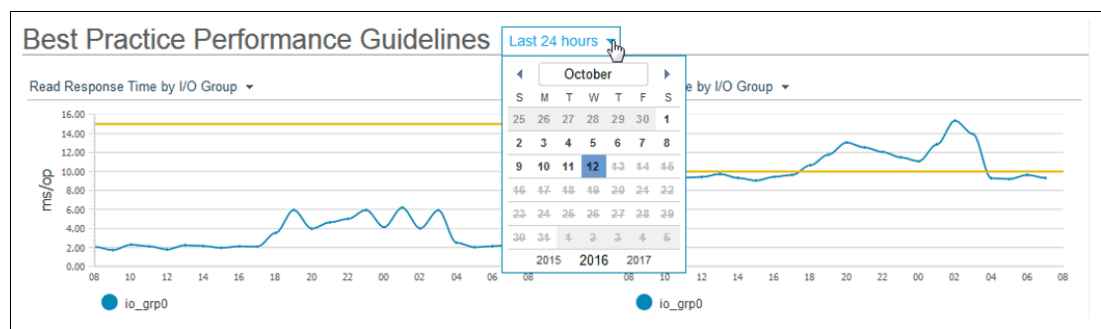


Figure 8-3 Change day of reference

Figure 8-4 on page 293 shows a metric that is exceeding the best practice limit (orange line).

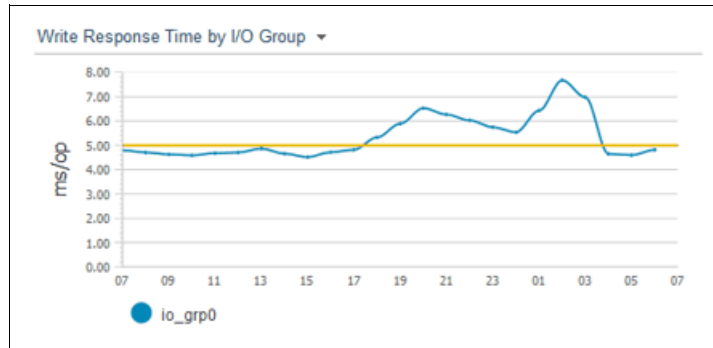


Figure 8-4 Metric exceeding best practice

Figure 8-5 shows the same chart as in Figure 8-4 with `io_grp0` selected which overlays the previous day and 7 days prior.

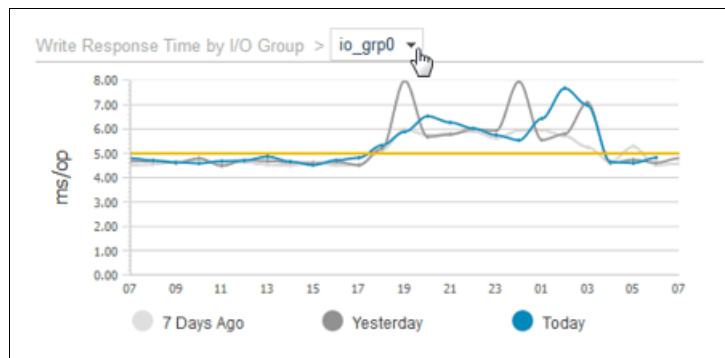


Figure 8-5 iogrp selection

From this we can quickly conclude that this exception occurs everyday 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 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 best practice 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 page in the web-based GUI, click Alerts in the General section, and then click Definitions.

8.2.3 Important metrics for debugging

Here are some of the most important metrics that need to be analyzed to debug performance problem in 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 I/Os/sec
- ▶ **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 given time. In case of Disk or Storage Subsystem is the amount of Data moved from a Host to a specific storage device. Typically measured in MB/sec
- ▶ **Responds time R/W:** 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 is the time used to complete a I/O operation. Typically measured in msec.
- ▶ **Cache Hit R/W:** Is the percentage of times where a read data o write data can be found already in cache or can find cache free space where to be written.

- ▶ **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 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 will be represented by less than 1 msec 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, for sure, but 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 MB per second for operations in which the port receives or send 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 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 Spectrum Virtualize/Storwize systems have limited number of GM I/O's that can be delivered and it is around 90.000 per each IO group.
- ▶ **Global Mirror (Secondary Write Lag):** The average number of additional 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. You 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 Spectrum Control from Spectrum Virtualize/Storwize systems. For more information about all metrics, refer to:

http://www.ibm.com/support/knowledgecenter/SS5R93_5.2.11/com.ibm.spectrum.sc.doc/t_pch_r_metrics_svc.html

8.2.4 Performance support package

In case you have performance issues on your system at any level (Host, Volume, Nodes, Pools etc.), you consult IBM Support who require detailed performance data about the Spectrum Virtualize/Storwize system to diagnose the problem. You are asked to 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 8-6.

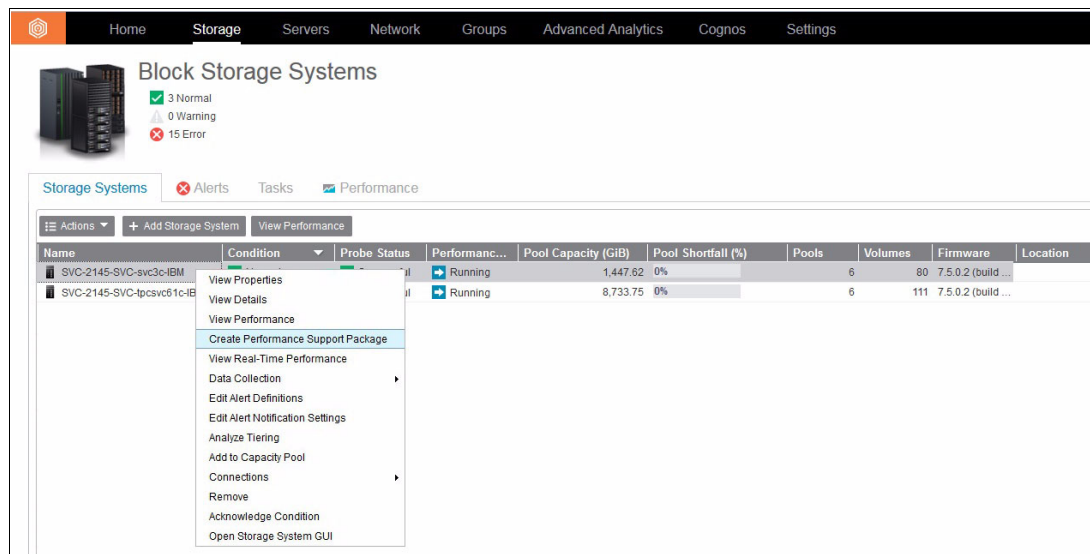


Figure 8-6 Performance support package creation

When the package has been created you will be requested to download in .zip format and the package will encompass different report in .csv format as shown in Figure 8-7.

log.txt	Text Document	1 KB
PerfReport_ITSO_SVC_ESC_Disks_20161017-233400_12hrs0mins.csv	CSV File	1 KB
PerfReport_ITSO_SVC_ESC_HostConnections_20161017-233400_12hrs0mins.csv	CSV File	1 KB
PerfReport_ITSO_SVC_ESC_IOGroups_20161017-233400_12hrs0mins.csv	CSV File	12 KB
PerfReport_ITSO_SVC_ESC_ManagedDisks_20161017-233400_12hrs0mins.csv	CSV File	18 KB
PerfReport_ITSO_SVC_ESC_Nodes_20161017-233400_12hrs0mins.csv	CSV File	18 KB
PerfReport_ITSO_SVC_ESC_Pools_20161017-233400_12hrs0mins.csv	CSV File	16 KB
PerfReport_ITSO_SVC_ESC_StoragePorts_20161017-233400_12hrs0mins.csv	CSV File	36 KB
PerfReport_ITSO_SVC_ESC_StorageSystem_20161017-233400_12hrs0mins.csv	CSV File	12 KB
PerfReport_ITSO_SVC_ESC_Volumes_20161017-233400_12hrs0mins.csv	CSV File	12 KB

Figure 8-7 Package files example

For more information about how to create a performance support package refer to:

http://www.ibm.com/support/knowledgecenter/en/SS5R93_5.2.11/com.ibm.spectrum.sc.doc/scn_per_create_support_package.html

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 even if it is available. 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 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 Spectrum Virtualize/Storwize systems are implemented in a DR solution with Metro Mirror (MM) or Global Mirror (GM) functions, is to check if MM or GM has been suspended because of a 1920 error or 1720 error.

As explained in Chapter 5, “Copy Services” on page 131, 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 can be caused by several factors.

Spectrum Virtualize/Storwize systems do not restart the MM or GM automatically, and this has to be done manually.

Without any alert or monitoring system implemented, your application RPO will start to increase and you will not want that.

How to set Spectrum Virtualize/Storwize systems alert monitoring is explained in 8.1.1, “Monitoring with GUI” on page 288.

When MM or GM is managed by IBM CSM and if a 1920 error occurs, IBM CSM has the ability to automatically restart MM or GM sessions and has the ability to set the delay time on the automatic restart option.

This delay allows some time for the situation to correct itself. Or, if you have several sessions, you can stagger them so that they do not all restart at the same time, which could cause an impact to the system.

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) 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 specific information about IBM Copy Service Manager, refer to:

http://www.ibm.com/support/knowledgecenter/en/SSESK4_6.1.3/com.ibm.storage.csm.help.doc/frg_r_session_properties_svc_mmffwp_gmffwp.html

8.3.1 Monitoring MM and GM with scripts

Spectrum Virtualize/Storwize system provides a complete command line interface (CLI), and you have the opportunity to interact with your systems using scripts.

Those scripts can run in the Spectrum Virtualize/Storwize shell but with a limited script command set available, or they can run out of the shell using any scripting language you prefer.

An example of script usage is one to check with a specific interval time if MM or GM are still active, or if any 1920 errors have occurred, or to react to an SNMP or e-mail alert received. Then the script could start some specific recovery action based on your recovery plan and environment.

Customers that do not use IBM Copy Service Manager have created their own scripts and sometimes these are supported by IBM as part of ITS professional services or IBM System Lab. services.

The best practice here is to tell your IBM representative what kind of monitoring you want to implement with scripts, and together try to find if one already exists in the IBM Intellectual Capital Management repository that can be re-used.

8.4 Monitoring Tier1 SSD

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 we suggest that you stay alert to the new fields 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 (2.42PB for the 15.36TB drive) Starts at 0, and can continue >100
write_endurance_usage_rate	Measuring / Low / Marginal / High Takes 160 Days to get initial measurement; Low ? 5.5 Years or more Marginal ? 4.5 – 5.5 Years High ? <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 prior to limit

In a case of these triggered events:

- ▶ SS_EID_VL_ER_SSD_WRITE_ENDURANCE_USAGE_RATE_HIGH
- ▶ SS_EID_VL_ER_SSD_DRIVE_WRITE_ENDURANCE_LIMITED

make sure that you contact your IBM Service Representative to put an action plan in place.



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 day-to-day activities of storage administration by using the Spectrum Virtualize software installed on IBM SAN Volume Controller, IBM Storwize family, and IBM FlashSystem V9000. This guidance can help you to maintain your storage infrastructure with the levels of availability, reliability, and resiliency that is demanded by today's applications, and to keep up with storage growth needs.

This chapter focuses on the most important topics to consider in Spectrum Virtualize administration so that you can use this chapter as a checklist. It also provides tips and guidance.

Important: The practices that are described here were effective in many Spectrum Virtualize installations worldwide for organizations in several areas. They all had one common need, which was the need 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 Spectrum Virtualize and SAN environment
- ▶ Storage management users
- ▶ Standard operating procedures
- ▶ Spectrum Virtualize code update
- ▶ SAN modifications
- ▶ Hardware upgrades for Spectrum Virtualize
- ▶ Adding expansion enclosures
- ▶ Spectrum Virtualize scripting

9.1 Documenting Spectrum Virtualize and SAN environment

This section focuses on the challenge of automating the documentation that is needed for a 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 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 Spectrum Virtualize and SAN environment documentation is 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 Spectrum Virtualize solution, use of these forms might also help you document how your Spectrum Virtualize solution was first configured.

Minimum documentation is needed for a 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 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 Spectrum Virtualize and SAN environment and system administrators can determine whether a name belongs to a volume, storage pool, MDisk, host, or host bus adapter (HBA) by looking at it. Also, because error messages often point to the device that generated an error, a good naming convention quickly highlights where to start investigating if an error occurs.

Typical Spectrum Virtualize and SAN component names limit the number and type of characters you can use. For example, Spectrum Virtualize names are limited to 63 characters, which makes creating a naming convention a bit easier than in previous versions of Spectrum Virtualize code.

Names: In previous versions of Spectrum Virtualize code, names were limited to 15 characters. Starting with version 7.1, the limit is 63 characters.

Many names in 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. (Server names are the exception, as explained in “Hosts” on page 304.) The naming examples that are used in the following sections are proven to be effective in most cases, but might not be fully adequate to your particular environment or needs. The naming convention to use is your choice, but you must implement it in the whole environment.

Storage controllers

Spectrum Virtualize names the storage controllers `controllerX`, with X being a sequential decimal number. If multiple controllers are attached to your 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 `controllerX`, you do not need to log in to Spectrum Virtualize to know which storage controller to check.

Note: Spectrum Virtualize detects controllers that are based on their WWNN. If you have a storage controller that has one WWNN for each WWPN this might lead to many `controllerX` names pointing to the same physical box. In this case, you should prepare a naming convention to cover this situation.

MDisks and storage pools

When the 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

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 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.

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 Spectrum Virtualize is as simple as changing any other 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, so 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, where:

- ▶ LL is the location, which might designate a city, data center, building floor, or room, and so on
- ▶ 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 worldwide port name (WWPN) in your SAN, and use these aliases in your zoning configuration. Consider the following examples:

- ▶ NYBIXTDB02_FC2: Interface fcs2 of AIX server NYBIXTDB02 (WWPN)
- ▶ SVC02_N2P4: SAN Volume Controller cluster SVC02, port 4 of node 2 (WWPN format 5005076XXXXXXXXX).

Be mindful of the 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.

- ▶ SVC02_I02_A: SAN Volume Controller cluster SVC02, ports group A for iogrp 2 (aliases *SVC02_N3P1*, *SVC02_N3P3*, *SVC02_N4P1*, and *SVC02_N4P3*)
- ▶ D8KXYZ1_I0301: DS8000 serial number 75VXYZ1, port I0301(WWPN)
- ▶ TL01_TD06: Tape library 01, tape drive 06 (WWPN)

If your SAN does not support aliases, for example, in heterogeneous fabrics with switches in some interoperations modes, use WWPNs in your zones all across. 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:

- ▶ `servername_svcclustername` (from a server to the SAN Volume Controller)
- ▶ `svcclustername_storagename` (from the SAN Volume Controller cluster to its back-end storage)
- ▶ `svccluster1_svccluster2` (for remote copy services)

9.1.2 SAN fabrics 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 it 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 email 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 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, you must explore the options that are provided to learn how to create a well-organized and useful diagram. Figure 9-1 shows an example of a poorly formatted diagram.

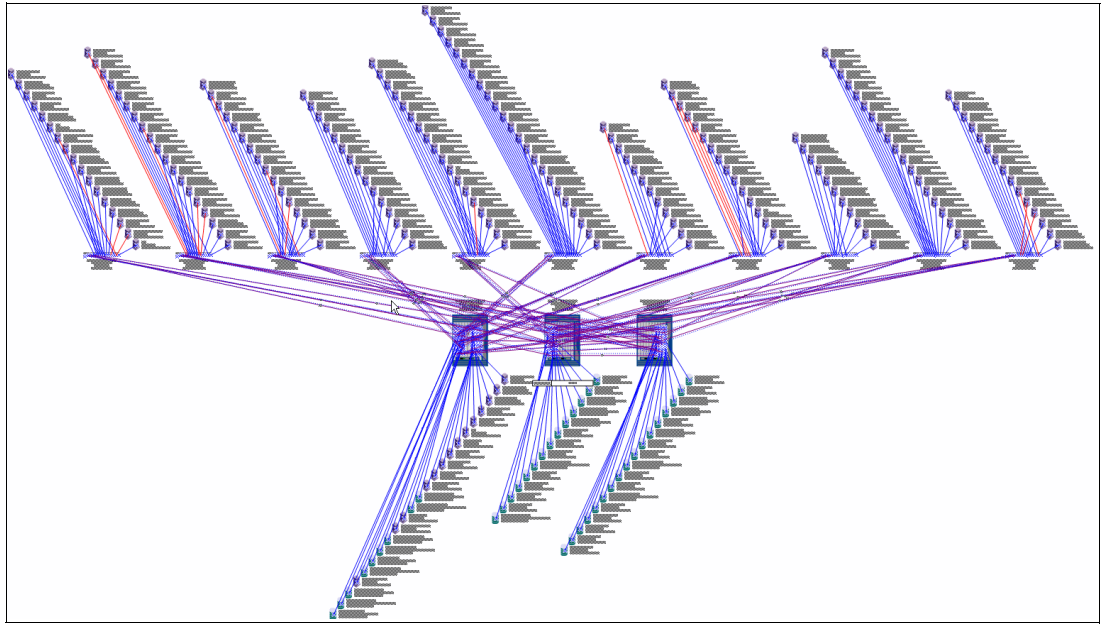


Figure 9-1 A poorly formatted SAN diagram

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.

SAN Health Options

General Options | Switch Diagnostics | Device Names | Report Format | Diagram Format | Audit Data File Upload

Excel Report Content

- ☐ Create a full length SAN Health report
- ☒ Create a shorter summary report showing only the SAN, Fabric and Switch summaries.
- ☐ Include performance graphs in the report
- ☐ Include the Visio place holder page that is inserted for page numbering consistency
- ☐ Use the device name rather than the port number on the performance graph
- ☐ Include empty ports in the report

Excel Report Formatting Options

- ☐ Format Pages For Letter Size Paper
- ☒ Format Pages For A4 Size Paper
- ☐ Add an image or company logo to the report page footer

File Name (*.bmp, *.gif, *.jpg)

Recommended Image Size: Width 180 x Height 60 Pixels

Figure 9-2 Brocade SAN Health Options window

SAN Health supports switches from manufactures other than Brocade, such as McData and Cisco. Both the data collection tool download and the processing of files are available at no cost, and 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 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 Spectrum Control reports, see the Spectrum Control documentation at:

<http://www.ibm.com/support/knowledgecenter/SS5R93>

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 Spectrum Virtualize documentation

You can back up the configuration data for a 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 **svcconfig backup** command to back up your configuration:
2. The command will output messages similar to the ones in Example 9-1.

Example 9-1 svcconfig 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 ...
.....
```

3. The **svcconfig 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
svc.config.backup.xml_<serial#>	Contains your configuration data.
svc.config.backup.sh_<serial#>	Contains the names of the commands that were issued to create the backup of the system.
svc.config.backup.log_<serial#>	Contains details about the backup, including any reported errors or warnings.

4. 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
```
5. If the process fails, resolve the errors, and run the command again.
6. 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 Spectrum Virtualize documentation spreadsheet.

The configuration backup file may contain too much data, for example it contains information about each internal storage drive installed in the system. Importing the file into your Spectrum Virtualize documentation spreadsheet might make it unreadable.

In this case, you might want to 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, see section 9.8, “Spectrum Virtualize scripting” on page 328. Then, use spreadsheet macros to import the collected data into your Spectrum Virtualize documentation spreadsheet.

When you are gathering 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 might 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 regularly. Make the resulting spreadsheet easy to consult and have it contain only the relevant information 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 311.

9.1.4 Storage documentation

Fully allocate all of the available space in the storage controllers that you use as back-end to the Spectrum Virtualize solution. This way, you can perform all your Disk Storage Management tasks by using 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 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 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 documentation that you might 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 might include the following details:

- ▶ 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, including the 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 unless the spreadsheet is password-protected.

- 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 is 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 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 email alerts that are automatically sent to the cell 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 them 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 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 Spectrum Virtualize this 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 Spectrum Virtualize support

Subscribing to 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.

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 Embedded Security Services (ESS)

Regardless of the authentication method you choose, perform 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 Spectrum Virtualize environment healthy and reliable.

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 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 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 and/or delete volumes) will be policed to prevent the removal of mappings to volumes and/or deletion of volumes which are considered active. Active means that the system has detected IO 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 in order 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. At all costs, avoid the use of the **-force** parameter in **rmvdisk**. If you issue the **svctask rmvdisk** command and it still has pending mappings, 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 destroying them (**rmvdisk**). The IBM internal Storage Technical Quality Review Process (STQRP) asks for a minimum of a 48-hour interval so that you can perform a quick backout if you later realize you still need some data in that volume.

9.3.2 Adding and removing hosts

When you add and remove host (or hosts) in Spectrum Virtualize, consider the following guidelines:

- ▶ Before you map new servers to Spectrum Virtualize, verify that they are all error free. Fix any errors that you find in your server and Spectrum Virtualize before you proceed. In 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; for example, 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 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 Spectrum Virtualize nodes and ports that you expected. When you create the host definition in 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 **cfgmgr** command in AIX.

9.4 Spectrum Virtualize code update

Because 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 Spectrum Virtualize update.

Most of the following sections explain how to prepare for the Spectrum Virtualize update. The last two sections present version-independent guidelines to update the Spectrum Virtualize system and disk drive.

9.4.1 Current and target Spectrum Virtualize code level

First, determine your current and target Spectrum Virtualize code level. Log in to your Spectrum Virtualize web-based GUI and find the current version, the specific tab to use will vary depending on the version itself. Alternatively, if you are using the CLI, run the **svcinfo 1ssystem** command.

Spectrum Virtualize code levels are specified by four digits in the format V.R.M.F, where:

- ▶ 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 Generally Available (GA) Spectrum Virtualize release unless you have a specific reason not to update, such as:

- ▶ The specific version of an application or other component of your SAN Storage environment has a known problem or limitation.
- ▶ The latest 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 V3700 Concurrent Compatibility and Code Cross-Reference:
<http://www.ibm.com/support/docview.wss?uid=ssg1S1004172>
- ▶ Storwize V5000 Concurrent Compatibility and Code Cross-Reference:
<http://www.ibm.com/support/docview.wss?uid=ssg1S1004336>
- ▶ 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 Spectrum Virtualize Upgrade Test Utility

Install and run the latest Spectrum Virtualize Upgrade Test Utility before you update the 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 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 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 7.7 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.

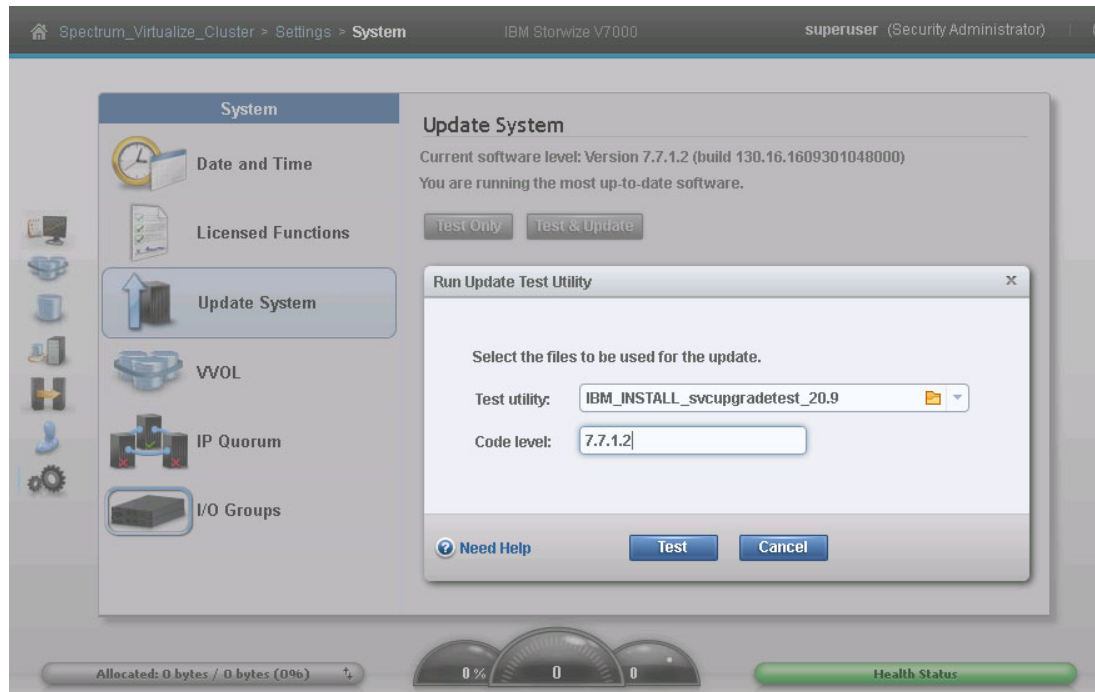


Figure 9-3 Spectrum Virtualize Upgrade Test Utility installation using the GUI

In Example 9-2 on page 315, we show how to install and run Upgrade Test Utility in CLI. In this case the Upgrade Test Utility found warnings and errors and indicates recommended actions.

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	2 in slot	19 in enclosure	1
ST9300603SS	B53E	B53B	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 Spectrum Virtualize hardware considerations

Before you start the update process, always check whether your Spectrum virtualize hardware and target code level are compatible.

If part or all your current hardware is not supported at the target code level you wish to update to, you will have to replace 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 Spectrum Virtualize code first.

9.4.4 Attached hosts preparation

If the appropriate precautions are taken, the Spectrum Virtualize update is not apparent to the attached servers and their applications. The automated update procedure updates one Spectrum Virtualize node at a time, while the other node in the I/O group covers for its designated volumes.

To ensure this, however, the *failover capability* of your servers' multipath software must be working properly. This can be mitigated by enabling NPIV if your current code level supports this functionality. For more information regarding NPIV refer to Chapter 6, "Hosts" on page 225.

Before you start Spectrum Virtualize update preparation, check the following items for every server that is attached to the Spectrum Virtualize 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
- ▶ The IBM Support page on SAN Volume Controller Flashes and Alerts (Troubleshooting):
[https://www.ibm.com/support/entry/myportal/all_troubleshooting_links/system_storage/storage_software/storage_virtualization/san_volume_controller_\(2145\)](https://www.ibm.com/support/entry/myportal/all_troubleshooting_links/system_storage/storage_software/storage_virtualization/san_volume_controller_(2145))
- ▶ The IBM Support page on Storwize V7000 Flashes and Alerts (Troubleshooting):
[https://www.ibm.com/support/entry/myportal/all_troubleshooting_links/system_storage/disk_systems/mid-range_disk_systems/ibm_storwize_v7000_\(2076\)](https://www.ibm.com/support/entry/myportal/all_troubleshooting_links/system_storage/disk_systems/mid-range_disk_systems/ibm_storwize_v7000_(2076))
- ▶ The IBM Support page on Storwize V5000 Flashes and Alerts (Troubleshooting):
https://www.ibm.com/support/entry/myportal/all_troubleshooting_links/system_storage/disk_systems/mid-range_disk_systems/ibm_storwize_v5000
- ▶ The IBM Support page on Storwize V3700 Flashes and Alerts (Troubleshooting):
https://www.ibm.com/support/entry/myportal/all_troubleshooting_links/system_storage/disk_systems/entry-level_disk_systems/ibm_storwize_v3700

Fix every problem or "suspect" that you find with the disk path failover capability. Because a typical Spectrum virtualize environment has several dozens of servers to a few hundred servers that are attached to it, the use of 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 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 a 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 Spectrum Virtualize environment must continue working through the other fabric and your applications must remain unaffected.

Because you are updating your 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 Spectrum Virtualize code level with all components of your SAN storage environment (SAN switches, storage controllers, servers 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 Spectrum Virtualize target code level is compatible.

The Spectrum Virtualize Supported Hardware List provides the complete information for using your 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>
- ▶ Support Information for IBM Storwize V5000:
<http://www.ibm.com/support/docview.wss?uid=ssg1S1004971>
- ▶ Support Information for IBM Storwize V3700:
<http://www.ibm.com/support/docview.wss?uid=ssg1S1004388>

By cross-checking the version of Spectrum Virtualize that are 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. Spectrum Virtualize system

5. Spectrum Virtualize internal disk drives

Attention: Do *not* update two components of your Spectrum Virtualize SAN storage environment simultaneously, such as the Spectrum Virtualize system and one storage controller, 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 Spectrum Virtualize participating in Metro Mirror or Global Mirror

When you update a 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 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 Spectrum Virtualize update

Adhere to the following version-independent guidelines for your Spectrum Virtualize code update:

- ▶ Schedule the Spectrum Virtualize code update for a low I/O activity time. The update process puts one node at a time offline, and 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 a 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 Spectrum Virtualize target code level on every computer that you intend to use to manage your Spectrum Virtualize.
- ▶ If you are planning for a major Spectrum Virtualize version update, update your current version to its latest fix level before you run the major update.

9.4.10 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 IOs 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 in case of 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 WWPNS 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 Spectrum Virtualize host definitions, it continues to access its previously mapped volumes exactly because the WWPNS 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 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 WWPNS** command to see output similar to what is shown in Example 9-3.

Example 9-3 Output of the pcmpath query WWPNS command

```
[root@nybixtdb02]> pcmpath query wwpns
```

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 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 10000000C925F5B0
node_logged_in_count 2
state active
WWPN 10000000C9266FD1
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

```
blg32sw1_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 Spectrum Virtualize `vdisk_uid` with the server LUN ID before you perform any modifications that involve 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 Spectrum Virtualize volume UID with 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 powered up and 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 320). 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 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.
There should be only one alias that uses each WWPN, and zones must reference this aliases.
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 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 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 show as inactive or offline.
8. Return to the server. Then, reconfigure the multipath software to recognize the new HBA and its associated SAN disk paths. Certify that all SAN LUNs have redundant, healthy disk paths through the good and the new HBAs.
9. Return to the 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 healthy, active ones.

By following these steps, you avoid removing your only good HBA in error.

9.6 Hardware upgrades for Spectrum Virtualize

The Spectrum Virtualize scalability features allow significant flexibility in its configuration. As a consequence, several scenarios are possible for its growth. The following sections describe:

- ▶ Adding Spectrum Virtualize nodes to an existing cluster
- ▶ Upgrading Spectrum Virtualize nodes in an existing cluster
- ▶ Moving to a new Spectrum Virtualize cluster
- ▶ Splitting a Spectrum Virtualize cluster

9.6.1 Adding Spectrum Virtualize nodes to an existing cluster

If your existing 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 Spectrum Virtualize code level first. For more information, see 9.4.3, “Spectrum Virtualize hardware considerations” on page 316.

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 Spectrum Virtualize. Although this 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 Spectrum Virtualize the new servers will use. Otherwise, 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 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 makes it difficult to identify and resolve eventual performance problems.

9.6.2 Upgrading 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 the SAN Volume Controller Knowledge Center for your current code level at this website:

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

From version 7.8 we also offer the following Storwize node canisters upgrade options:

- ▶ from Storwize V5010 to Storwize V5020
- ▶ from Storwize V5010 to Storwize V5030
- ▶ from Storwize V5020 to Storwize V5030
- ▶ from Storwize V7000 Gen2 to Storwize V7000 Gen2+

The new node canister will assume 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 the Knowledge Center for your product and current code level at these web sites:

- ▶ Storwize V5000 <http://www.ibm.com/support/knowledgecenter/STHGJ>
- ▶ Storwize V7000 <http://www.ibm.com/support/knowledgecenter/ST3FR7>

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 Spectrum Virtualize code, it will install the cluster version automatically during the node replacement procedure.

9.6.3 Moving to a new Spectrum Virtualize cluster

You might have a highly populated, intensively used Spectrum Virtualize cluster that you want to upgrade, and you also want to use the opportunity to overhaul your Spectrum Virtualize and SAN storage environment.

The following steps might make it easier is to replace your cluster entirely with a newer, bigger, and more powerful one:

1. Install your new Spectrum Virtualize cluster.
2. Create a replica of your data in your new cluster.
3. Migrate your servers to the new Spectrum Virtualize cluster when convenient.

If your servers can tolerate a brief, scheduled outage to switch from one Spectrum Virtualize cluster to another, you can use Spectrum Virtualize's remote copy services (Metro Mirror or Global Mirror) to create your data replicas, following these steps:

1. Select a host you want to move to the new Spectrum Virtualize cluster and find all the old volumes you will have to move.
2. Zone your host to the new Spectrum Virtualize cluster.
3. Create remote copy relationships from the old volumes in the old Spectrum Virtualize cluster to new volumes in the new Spectrum Virtualize cluster.
4. Map the new volumes from the new Spectrum Virtualize cluster to the host.
5. Discover new volumes on the host.
6. Stop all IO from the host to the old volumes from the old Spectrum Virtualize cluster.
7. Disconnect and remove the old volumes on the host from the old Spectrum Virtualize cluster.
8. Unmap the old volumes from the old Spectrum Virtualize cluster to the host.
9. Make sure remote copy relationships between old and new volumes in the old and new 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 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 will have to use host-based mirroring, following these steps:

1. Select a host you want to move to the new Spectrum Virtualize cluster and find all the old volumes you will have to move.
2. Zone your host to the new Spectrum Virtualize cluster.
3. Create new volumes in the new Spectrum Virtualize cluster of the same size as the old volumes in the old Spectrum Virtualize cluster.
4. Map the new volumes from the new 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, once 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 Spectrum Virtualize cluster.
9. Unmap the old volumes from the old Spectrum Virtualize cluster to the host.

This approach uses the server's computing resources (CPU, memory, and I/O) to replicate the data but can be done online if properly planned. Before you begin, make sure it has such resources to spare.

The biggest benefit to using either approach is that they easily accommodates (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 a Spectrum Virtualize cluster

Splitting a 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 Spectrum Virtualize system that handles all I/O operations, hosts, and subsystem storage attachments. The goal is to create a second 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 Spectrum Virtualize system. You can 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 workload onto the new system following the same approach described in 9.6.3, “Moving to a new Spectrum Virtualize cluster” on page 325.

It is uncommon to reduce the number of I/O groups. 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 planned well, when you buy a Spectrum Virtualize product that supports internal storage it has enough disk drives to run your business. But as time passes and your environment grows, there might be a need to add more storage to your system.

Depending on the Spectrum Virtualize product you have and on the code level installed, you can add different numbers of expansion enclosures to your system. Because all Spectrum Virtualize systems were designed to make managing and maintaining as simple as possible, adding an expansion enclosure also is an easy task. However, even with this ease, there are some guidance and preferred practices you should follow.

At the time of writing, the following 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 Spectrum Virtualize products support two chains of expansion enclosures:

- ▶ Storwize V5000
- ▶ Storwize V5030
- ▶ Storwize V7000 (Gen1, Gen2, Gen2+)
- ▶ FlashSystem 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 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, be mindful 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 extend time of problem resolution. 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, you should 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 is even more important on Storwize V7000 Gen2 and Storwize Gen2+, 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 Spectrum Virtualize scripting

Although the Spectrum Virtualize GUI is a powerful interface (similar to other GUIs), it is not well-suited to perform large amounts of specific operations. For complex, often-repeated operations, it is more convenient to use the Spectrum Virtualize CLI interface. CLI commands can be scripted by using any program that can pass text commands to the Spectrum Virtualize Secure Shell (SSH) connection.

On UNIX systems, you can use the `ssh` command to create an SSH connection with a Spectrum Virtualize system. On Windows systems, you can use the `plink.exe` utility (which is provided with the PuTTY tool) to achieve the same.

Create a Spectrum Virtualize user with the least privileges in order to access and perform batch operations on the system. Do not grant it Administrator privileges unless strictly necessary. Create and configure an SSH key specifically for it.

9.8.1 Connecting to Spectrum Virtualize using PuTTY predefined connection

The easiest way to create an SSH connection to the SAN Volume Controller is when the **plink.exe** utility can call a predefined PuTTY session. When you define a session, you include the following information:

- The auto-login user name, which you set to the user created specifically to perform batch operations. To set this parameter, click **Connection** → **Data** in the left pane of the PuTTY Configuration window, as shown in Figure 9-4 on page 329.

Note: In Figure 9-4 on page 329 we use the admin user but you should try to contain the privileges of the user dedicated at batch operations to the bare minimum.

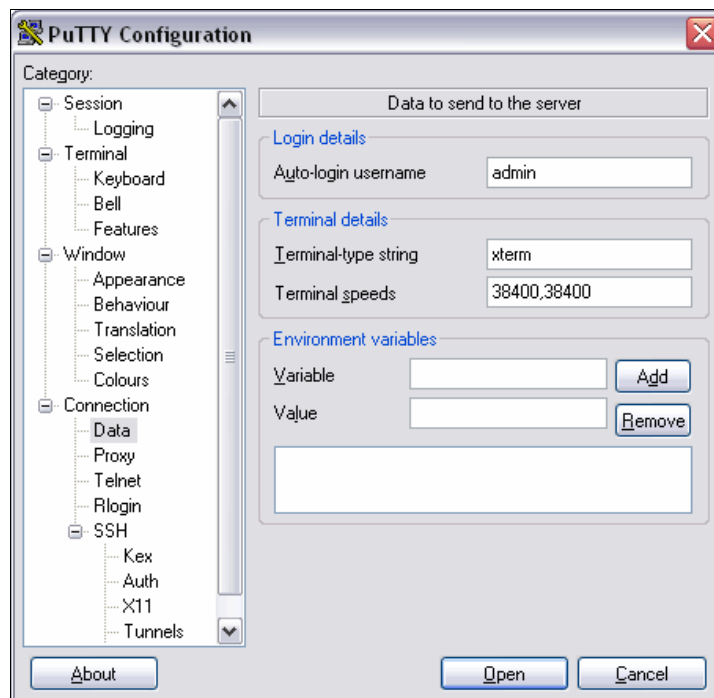


Figure 9-4 Configuring the auto-login user name

- The private key for authentication (for example, `icat.ppk`), which is the private key that you created. To set this parameter, select **Connection** → **SSH** → **Auth** in the left pane of the PuTTY Configuration window, as shown in Figure 9-5.

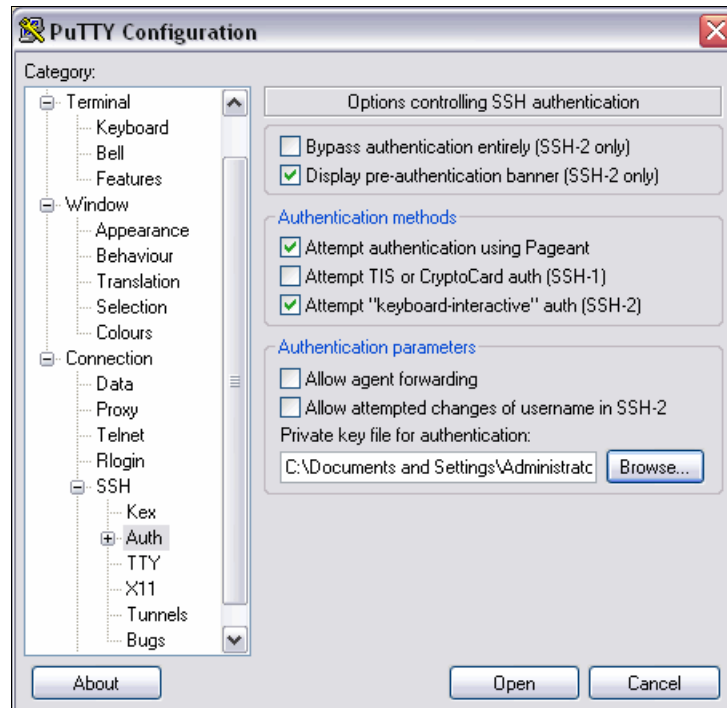


Figure 9-5 Configuring the SSH private key

- The IP address of the Spectrum Virtualize system. To set this parameter, select **Session** at the top of the left pane of the PuTTY Configuration window, as shown in Figure 9-6.

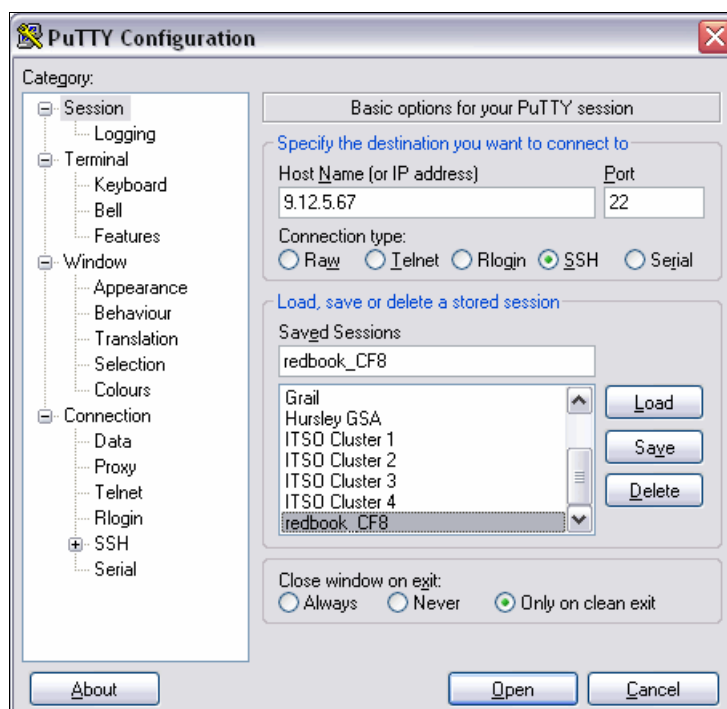


Figure 9-6 Specifying the IP address

When you are specifying the basic options for your PuTTY session, you need to set a session name, which in this example is `redbook_CF8`. To use the predefined PuTTY session, use the following syntax:

```
plink redbook_CF8
```

If you do not use a predefined PuTTY session, use the following syntax:

```
plink admin@<your cluster ip address> -i "C:\DirectoryPath\KeyName.PPK"
```

9.8.2 Run commands in the Spectrum Virtualize shell

You can run various limited scripts directly in the resuscitated Spectrum Virtualize shell. Example 9-8 shows a script to restart Global Mirror relationships and groups.

Example 9-8 Restarting Global Mirror relationships and groups

```
svcinfo lsrrconsistgrp -filtervalue state=consistent_stopped -nohdr | \
while read id name unused
do
    echo "Restarting group: $name ($id)"
    svctask startrrconsistgrp -force $name
done
svcinfo lsrrrelationship -filtervalue state=consistent_stopped -nohdr | \
while read id name master_cluster_id master_cluster_name master_vdisk_id \
    master_vdisk_name aux_cluster_id aux_cluster_name aux_vdisk_id \
    aux_vdisk_name primary consistency_group_id junk
do
    if [ "$consistency_group_id" == "" ]; then
        echo "Restarting relationship: $name ($id)"
        svctask startrrrelationship -force $name
    fi
done
```

9.8.3 Scripting toolkit

For more elaborate scripts the restricted Spectrum Virtualize shell might not be powerful enough. In those cases the processing should be performed on the machine connecting to the Spectrum Virtualize shell while the shell should be used only to run single commands.

IBM engineers developed a scripting toolkit that helps to automate Spectrum Virtualize operations. This scripting toolkit is based on Perl and is available at no-charge from the following IBM alphaWorks® website:

<https://www.ibm.com/developerworks/mydeveloperworks/groups/service/html/communityview?communityUuid=5cca19c3-f039-4e00-964a-c5934226abc1>

Attention: The scripting toolkit is available to users through the IBM alphaWorks website. As with all software that is available on the alphaWorks site, this toolkit was not extensively tested and is provided on an as-is basis. Because the toolkit is not supported in any formal way by IBM Product Support, use it at your own risk.



Troubleshooting and diagnostics

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 may cause some events.

This chapter provides an overview of common events that can occur in your environment. It describes situations that are related to Spectrum Virtualize, Storwize, the SAN environment, storage subsystems, hosts, and multipathing drivers. It also aims how to collect the necessary problem determination data.

This chapter includes the following sections:

- ▶ Common issues
- ▶ Collecting data and isolating the problem
- ▶ Recovering from problems
- ▶ Replacing a bad disk
- ▶ Health status during upgrade

10.1 Common issues

SANs, storage subsystems, and host systems can be complicated, often consisting 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.

With functions that are native in Spectrum Virtualize, administrators can easily locate any issue areas and take the necessary steps to fix any events. In many cases, Spectrum Virtualize and its service and maintenance features guide administrators directly, provide help, and suggest remedial action. Furthermore, Spectrum Virtualize probes whether the problem still persists or not.

When you experience events in the Spectrum Virtualize environment, ensure that all components that comprise the storage infrastructure are interoperable. In a Spectrum Virtualize environment, the Spectrum Virtualize support matrix is the main source for this information. For the latest Spectrum Virtualize V7.8 support matrix, see *V7.8 Supported Hardware List, Device Driver, Firmware and Recommended Software Levels*, available at:

Spectrum Virtualize

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009558>

Storwize V7000

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009559>

Although the latest Spectrum Virtualize code level is supported to run on older host bus adapters (HBAs), storage subsystem drivers, and code levels, IBM strongly recommends you to use the latest tested levels.

10.1.1 Host problems

From the host perspective, you can experience various situations that range from performance degradation to inaccessible disks. To diagnose these issues, you can check items from the host before you drill down to the SAN, Spectrum Virtualize, and storage subsystems.

Check the following areas on the host:

- ▶ Any special software that you are using
- ▶ Any recent change in the OS? Patching OS Upgrade and so on
- ▶ Operating system version and maintenance or service pack level
- ▶ Multipathing type and driver level
- ▶ Host bus adapter model, firmware, and driver level
- ▶ Host bus adapter connectivity issues

Based on this list, the host administrator must check and correct any problems.

For more information about managing hosts on the Spectrum Virtualize, see Chapter 6, “Hosts” on page 225.

10.1.2 Spectrum Virtualize events

Spectrum Virtualize has some 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. Every Spectrum Virtualize node maintains a

database of other devices that are visible in the SAN fabrics. This database is updated as devices appear and disappear.

Fast node reset

The Spectrum Virtualize Cluster software incorporates a *fast node reset* function. The intention of a fast node reset is to avoid I/O errors and path changes from the perspective of the host if a software problem occurs in one of the Spectrum Virtualize nodes. The fast node reset function means that Spectrum Virtualize software issues can be recovered without the host experiencing an I/O error and without requiring the multipathing driver to fail over to an alternative path. The fast node reset is performed automatically by the Spectrum Virtualize node. This node informs the other members of the cluster that it is resetting.

Other than Spectrum Virtualize node hardware and software events, failures in the SAN zoning configuration are a problem. A misconfiguration in the SAN zoning might lead to the Spectrum Virtualize cluster not working properly because the Spectrum Virtualize cluster nodes communicate with each other by using the Fibre Channel SAN fabrics.

You must check the following areas from a Spectrum Virtualize perspective:

- ▶ The attached hosts. For more information, see 10.1.1, “Host problems” on page 334.
- ▶ The SAN. For more information, see 10.1.3, “SAN events” on page 337.
- ▶ The attached storage subsystem. For more information, see 10.1.4, “Storage subsystem issues” on page 337.
- ▶ The local FC port masking. See 6.1.5, “Port masking” on page 228 for more details.

Spectrum Virtualize has several CLI commands that you can use to check the status of Spectrum Virtualize and the attached storage subsystems. 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 Spectrum Virtualize perspective:

▶ **svcinfo lscontroller *controllerid***

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 10.3.4, “Solving back-end storage issues” on page 345. The total of all *path_counts* must add up to the number of managed disks (MDisks) multiplied by the number of Spectrum Virtualize nodes.

▶ **svcinfo lsmdisk**

Check that all MDisk are online (not degraded or offline).

▶ **svcinfo lsmdisk *mdiskid***

Check several of the MDisk from each storage subsystem controller. Are they online? Do they all have *path_count* = number of backend ports in the zone to Spectrum Virtualize x number of nodes? (Example 10-1).

Example 10-1 issuing 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
fast_write_state empty
raid_status
raid_level
redundancy
strip_size
spare_goal
spare_protection_min
balanced
tier ssd
slow_write_priority
fabric_type fc
site_id
site_name
easy_tier_load very_high
encrypt yes
distributed no
drive_class_id
drive_count 0
stripe_width 0
rebuild_areas_total
rebuild_areas_available
rebuild_areas_goal
IBM_2145:itsosvcc11:superuser>
```

Example 10-1, “issuing lsmdisk command” on page 335 shows that the Flash900 has 8 ports zoned to Spectrum Virtualize, and the Spectrum Virtualize has 4 nodes, so $8 \times 4 = 32$.

► **svcinfolsvdisk**

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.

► **svcinfolshostvdiskmap**

Check that all volumes are mapped to the correct hosts. If a volume is not mapped correctly, create the necessary host mapping.

► **svcinfolsfabric**

Use this command with the various options, such as **-controller controllerid**. Also, you can check different parts of the Spectrum Virtualize configuration to ensure that multiple paths are available from each Spectrum Virtualize node port to an attached host or

controller. Confirm that all Spectrum Virtualize node port WWPNs are connected to the back-end storage consistently.

10.1.3 SAN events

Introducing the Spectrum Virtualize into your SAN environment and the use of its virtualization functions are not difficult tasks. Before you can use the Spectrum Virtualize in your environment, you must follow some basic rules. These rules are not complicated. However, 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 Spectrum Virtualize in your environment: a *host zone* and a *storage zone*. In addition, you must have a Spectrum Virtualize zone that contains all of the Spectrum Virtualize node ports of the Spectrum Virtualize cluster. This Spectrum Virtualize zone enables intracluster communication. For more information and important points about setting up the Spectrum Virtualize in a SAN fabric environment, see Chapter 1, “SAN topology” on page 1.

Because the Spectrum Virtualize is in the middle of the SAN and connects the host to the storage subsystem, check and monitor the SAN fabrics.

10.1.4 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 best 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 best practices are applied to the subsystem.
- ▶ Storage controller. Check the health and configurable settings on the controllers.
- ▶ 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 and configuration.
- ▶ Connectivity. Check the available paths (SAN environment).
- ▶ Layout and size of RAID arrays and LUNs. Performance and redundancy are contributing factors.

For more information about managing subsystems, see Chapter 2, “Back-end storage” on page 37.

Determining the correct number of paths to a storage subsystem

By using 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 10-2 shows how to obtain this information by using the **svcinfo lscontroller controllerid** and **svcinfo lsnode** commands.

Example 10-2 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>

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 10-2 shows that sixteen MDisks are present for the storage subsystem controller with ID 1, and four Spectrum Virtualize nodes are in the Spectrum Virtualize cluster. In this example, the following path_count is used of $16 \times 4 = 64$.

10.1.5 Port masking issues

Some situations of performance degradation and buffer-to-buffer credit exhaustion may be caused by incorrect local FC port masking and remote FC port masking. In order to have a healthy operation in your Spectrum Virtualize, IBM strongly recommends that you 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.

10.2 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 a Spectrum Virtualize environment are described.

10.2.1 Data collection

Data collection methods vary by operating system. You can collect the data for various major host operating systems.

First, collect the following basic information from the host:

- ▶ Operating system: Version and level
- ▶ HBA: Driver and firmware level
- ▶ Multipathing driver level

Regarding host, storage and SAN data collection, due to the dynamic changes that occur over time we recommend that you follow this IBM w3 Connections community (only available to IBMers):

<https://ibm.biz/emea-coc>

Note: This community is IBM internal only and you must use your intranet id and password. If you are not an IBM employee, you must contact your IBM representative or the vendor of your hardware and follow specific procedures for data collection.

The w3 Connections community has up-to-date procedures for several kind of devices, including hosts, storages and SAN, as shown in Figure 10-1 on page 340.

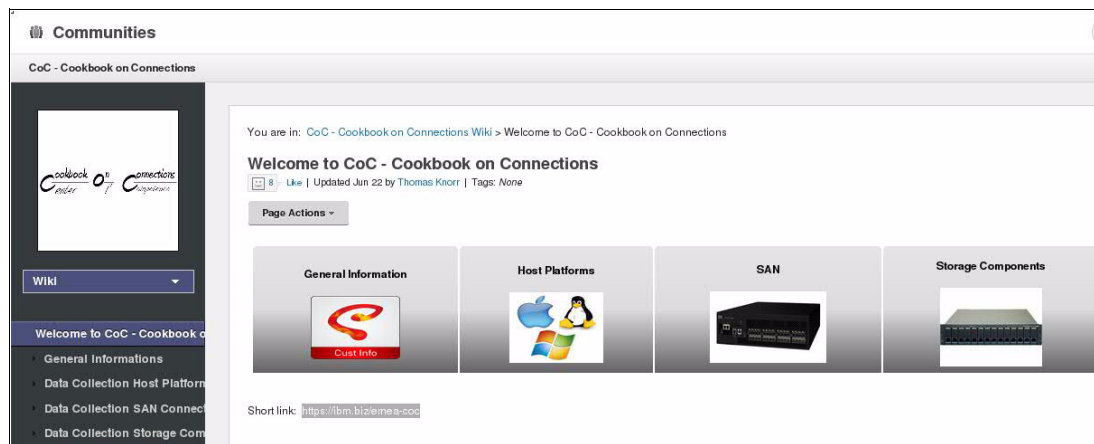


Figure 10-1 CoC - Cookbook on Connections internal wiki

10.3 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 IBM products, refer to the following IBM Support website:

<https://www-947.ibm.com/support/entry/myportal/support>

For more information about the latest flashes, concurrent code upgrades, code levels, and matrixes, see the following Spectrum Virtualize website:

<https://ibm.biz/BdRCzu>

10.3.1 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 10-3 shows two faulty paths that are reported by the SDD output on the host by using the **datapath query device -1** command. The faulty paths are the paths in the `close` state. Faulty paths can be caused by hardware and software problems.

Example 10-3 SDD output on a host with faulty paths

```
C:\Program Files\IBM\Subsystem Device Driver>datapath query device -1
```

```
Total Devices : 1
```

```
DEV#: 3 DEVICE NAME: Disk4 Part0 TYPE: 2145 POLICY: OPTIMIZED
SERIAL: 60050768018381BF2800000000000027
LUN IDENTIFIER: 60050768018381BF2800000000000027
```

```
=====
```

Path#	Adapter/Hard Disk			State	Mode	Select	Errors
0	Scsi	Port2	Bus0/Disk4 Part0	CLOSE	OFFLINE	218297	0
1 *	Scsi	Port2	Bus0/Disk4 Part0	CLOSE	OFFLINE	0	0
2	Scsi	Port3	Bus0/Disk4 Part0	OPEN	NORMAL	222394	0
3 *	Scsi	Port3	Bus0/Disk4 Part0	OPEN	NORMAL	0	0

Faulty paths can result from hardware issues, such as the following examples:

- ▶ Faulty small form-factor pluggable transceiver (SFP) on the host or SAN switch
- ▶ Faulty fiber optic cables
- ▶ Faulty HBAs

Faulty paths can result from software problems, such as the following examples:

- ▶ A back-level multipathing driver
- ▶ Obsolete HBA firmware and/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 in to 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.

Example 10-4 shows that one of the HBAs was experiencing a link failure because of a fiber-optic cable that had bent over too far. After changing the cable, the missing paths reappeared.

Example 10-4 Output from datapath query device command after fiber optic cable change

```
C:\Program Files\IBM\Subsystem Device Driver>datapath query device -l
```

Total Devices : 1

```
DEV#: 3  DEVICE NAME: Disk4 Part0  TYPE: 2145          POLICY: OPTIMIZED
SERIAL: 60050768018381BF2800000000000027
LUN IDENTIFIER: 60050768018381BF2800000000000027
```

Path#	Adapter/Hard Disk			State	Mode	Select	Errors
0	Scsi	Port3	Bus0/Disk4 Part0	OPEN	NORMAL	218457	1
1 *	Scsi	Port3	Bus0/Disk4 Part0	OPEN	NORMAL	0	0
2	Scsi	Port2	Bus0/Disk4 Part0	OPEN	NORMAL	222394	0

3 * Scsi Port2 Bus0/Disk4 Part0 OPEN NORMAL
0

10.3.2 Solving Spectrum Virtualize events

For any events in an environment that is implementing Spectrum Virtualize, use the Recommended Actions panel before you try to fix the problem anywhere else. Find the Recommended Actions panel by clicking **Monitoring** → **Events** → **Recommended Actions** in the Spectrum Virtualize Console GUI, as shown in Figure 10-2.

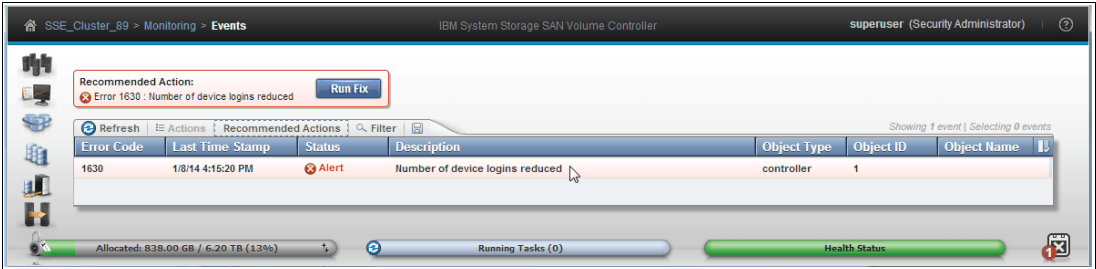


Figure 10-2 Recommended Action panel

The Recommended Actions panel shows event conditions that require actions and the procedures to diagnose and fix them. The highest-priority event is indicated with information about how long ago the event occurred. If an event is reported, you must select the event and run a fix procedure.

Complete the following steps to retrieve properties and sense about a specific event:

1. Select an event in the table.
2. Click **Properties** in the Actions menu, as shown in Figure 10-3.

Tip: You can also obtain access to the Properties by right-clicking an event.

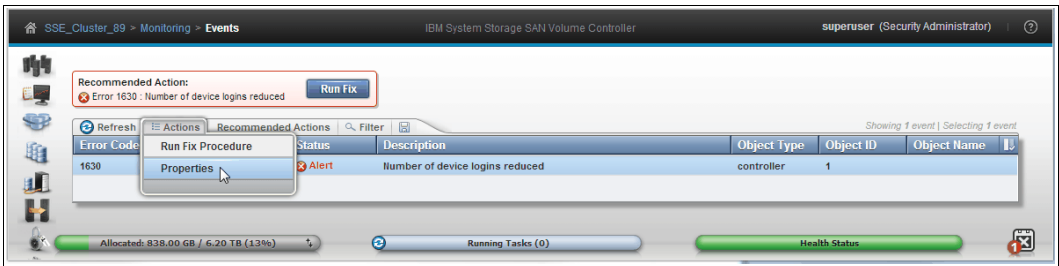


Figure 10-3 Event properties action

3. In the Properties and Sense Data for Event sequence_number window (see Figure 10-4, where sequence_number is the sequence number of the event that you selected in the previous step), review the information. Then, click **Close**.

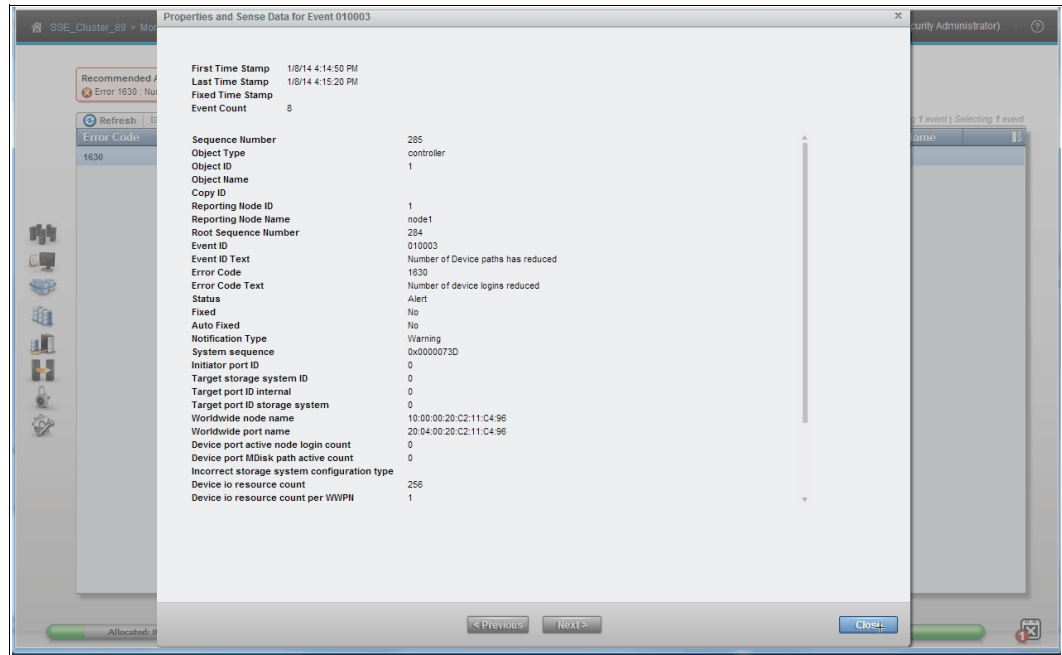


Figure 10-4 Properties and sense data for event window

Tip: From the Properties and Sense Data for Event Window, you can use the **Previous** and **Next** buttons to move between events.

You now return to the Recommended Actions panel.

Another common practice is to use the Spectrum Virtualize CLI to find issues. The following list of commands provides information about the status of your environment:

- ▶ **svctask detectmdisk**
Discovers changes in the back-end storage configuration.
- ▶ **svcinfo lscluster clustername**
Checks the Spectrum Virtualize cluster status.
- ▶ **svcinfo lsnode nodeid**
Checks the Spectrum Virtualize nodes and port status.
- ▶ **svcinfo lscontroller controllerid**
Checks the back-end storage status.
- ▶ **svcinfo lsmdisk**
Provides a status of all the MDisk.
- ▶ **svcinfo lsmdisk mdiskid**
Checks the status of a single MDisk.
- ▶ **svcinfo lsmdiskgrp**
Provides a status of all the storage pools.
- ▶ **svcinfo lsmdiskgrp mdiskgrpid**
Checks the status of a single storage pool.

► **svcinfolsvdisk**

Checks whether volumes are online and healthy,

Locating issues: Although the Spectrum Virtualize raises error messages, most events are not caused by the Spectrum Virtualize. Most issues are introduced by the storage subsystems or the SAN.

If the problem is caused by Spectrum Virtualize and you are unable to fix it by using the Recommended Action panel or the event log, collect the Spectrum Virtualize debug data as described in 10.2.1, “Data collection” on page 339.

To determine and fix other issues outside of Spectrum Virtualize, consider the guidance in the other sections in this chapter that are not related to Spectrum Virtualize.

10.3.3 Solving SAN issues

Some situations can cause issues in the SAN and on the 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 Spectrum Virtualize node ports must be zoned to one HBA with one port from each Spectrum Virtualize node. However, as shown in Example 10-5, 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. This incorrect zoning is shown in Example 10-5.

Example 10-5 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 10-6.

Example 10-6 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 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.

If you cannot fix the issue with these actions, use the method that is described in 10.2.1, “Data collection” on page 339, collect the SAN switch debugging data, and then contact IBM Support for assistance.

10.3.4 Solving back-end storage issues

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 error code. Other issues that are related to the storage subsystem are failures pointing to the managed disk I/O (error code 1310), disk media (error code 1320), and error recovery procedure (error code 1370).

However, all messages do not have only one explicit reason for being issued. Therefore, you must check multiple areas for issues and not the storage subsystem only. To determine the root cause of a problem, complete the following tasks:

- ▶ Check the Recommended Actions panel under Spectrum Virtualize.
- ▶ Check the attached storage subsystem for misconfigurations or failures.
- ▶ Check the SAN for switch problems or zoning failures.
- ▶ Collect all support data and involve IBM Support.

Complete the following steps:

1. Check the Recommended Actions panel by clicking **Monitoring** → **Events** → **Recommended Actions**, as shown in Figure 10-2 on page 342.

For more information about how to use the Recommended Actions panel, see the IBM System Storage Spectrum Virtualize Information Center at:

<https://ibm.biz/Bdsvxb>

2. Check the attached storage subsystem for misconfigurations or failures:
 - a. Independent of the type of storage subsystem, first check whether the system has any unfixed errors. Use the service or maintenance features that are provided with the storage subsystem to fix these issues.
 - b. Check whether the LUN masking is correct. When attached to the Spectrum Virtualize, ensure that the LUN masking maps to the active zone set on the switch. Create a similar LUN mask for each storage subsystem controller port that is zoned to the Spectrum Virtualize. Also, observe the Spectrum Virtualize restrictions for back-end storage subsystems, which can be found at:

Spectrum Virtualize

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009560>

Storwize V7000

<http://www-01.ibm.com/support/docview.wss?uid=ssg1S1009561>

If you need to identify which of the attached MDisk has which corresponding LUN ID, run the Spectrum Virtualize **svcinfo lsmdisk** CLI command (see Example 10-7). This command also shows to which storage subsystem a specific MDisk belongs (the controller ID).

Example 10-7 Determining the ID for the MDisk

```
IBM_2145:itsosvcc11:admin>svcinfo lsmdisk
id          name          status      mode
mdisk_grp_id mdisk_grp_name  capacity    ctrl_LUN_#
controller_name      UID
```

```
0          mdisk0          online          managed          0
MDG-1          600.0GB          0000000000000000
controller0
600a0b800017423300000059469cf845000000000000000000000000000000000
2          mdisk2          online          managed          0
MDG-1          70.9GB          00000000000000002
controller0
600a0b800017443100000096469cf0e8000000000000000000000000000000000
```

- Collect the support data for the involved SAN, Spectrum Virtualize, or storage systems as described in 10.2.1, “Data collection” on page 339.

For SAN issues or back-end storage issues, you can use the Spectrum Virtualize CLI to perform common error recovery steps.

- ▶ 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 configured or a zoning change occurs to ensure that the Spectrum Virtualize has recognized the changes.

- ▶ **svctask detectmdisk**
Discovers the changes in the back end.
- ▶ **svcinfo lscontroller and svcinfo lsmdisk**
Provides overall status of all controllers and MDisk.
- ▶ **svcinfo lscontroller controllerid**
Checks the controller that was causing the issue and verifies that all the WWPNs are listed as you expect.
- ▶ **svcinfo lsmdisk**
Determines whether all MDisk are now online.
- ▶ **svcinfo lscontroller controllerid**
Checks that the path_counts are distributed evenly across the WWPNs.

Finally, run the maintenance procedures on the Spectrum Virtualize to fix every error.

Always run directed maintenance procedures (dmp) to replace a failed disk. Spectrum Virtualize has a policy that it never writes to a disk unless the disk is defined. When a disk is replaced, the system identifies it as a new disk. To use the new disk, it marks the old disk as unused and the new disk as spare.

10.4 Health status during upgrade and known errors

When the software upgrade completes and the node canister firmware upgrade starts during the software and firmware upgrade process, the Health Status goes from Red to Orange to Green in the GUI until the upgrade is complete. This is normal behavior and is not an alarm.

While trying to upgrade a Spectrum Virtualize, you may get a message such as: **“The system is unable to install the update package. Error in verifying the signature of the update package.”** as shown in Figure 10-5.

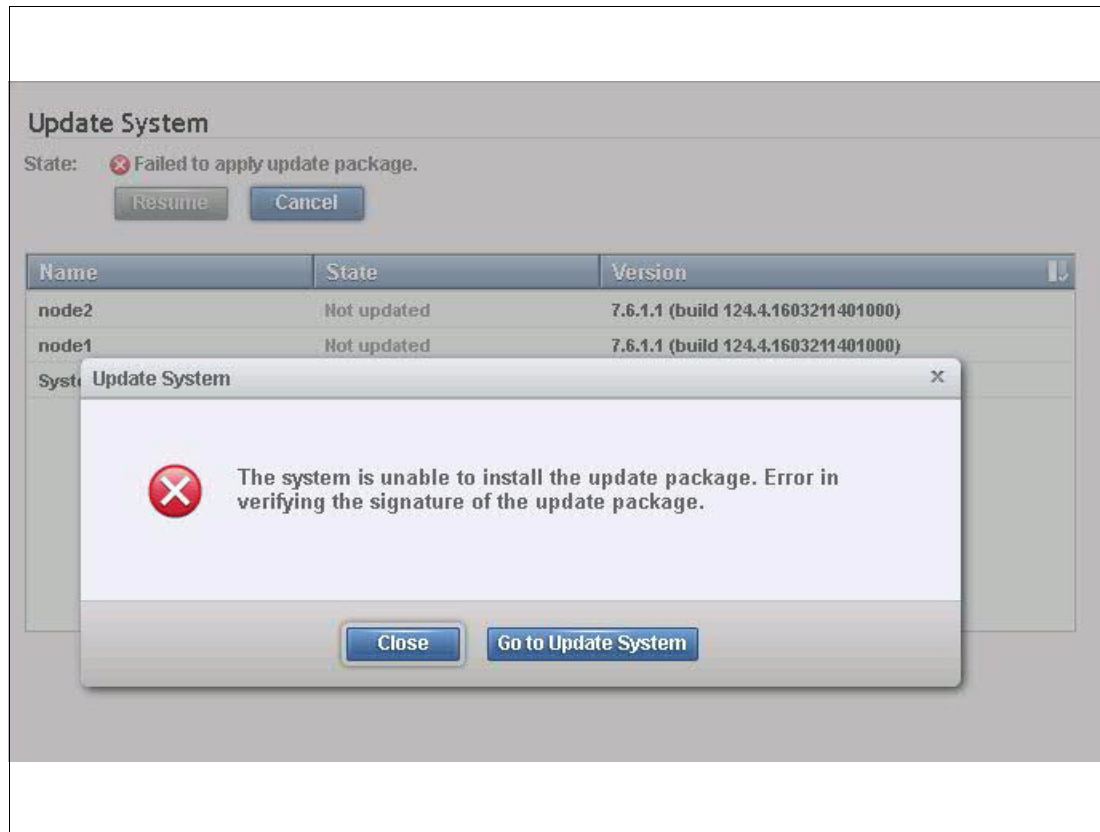


Figure 10-5 Error message during upgrade attempt

This does not mean you have a issue on the code package or in your system, our recommendation is to open a PMR and follow support instructions.



IBM Real-time Compression

This chapter highlights the best practises for using IBM Real-time Compression using the 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 adequate guidelines, and the factors to consider to achieve the best performance results and enjoy the compression savings that the Real-time Compression technology offers.

This chapter assumes the reader is already familiar with 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 and 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 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 Spectrum Virtualize version 7.6 the Comprestimator utility can be used directly from the 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 Spectrum Virtualize version 7.7 the Comprestimator utility can be used directly from the Spectrum Virtualize GUI. Figure 11-1 shows how to start a system wide analysis of compression estimates clicking on **Volumes** → **Actions** → **Space Savings** → **Estimate Compression Savings**.

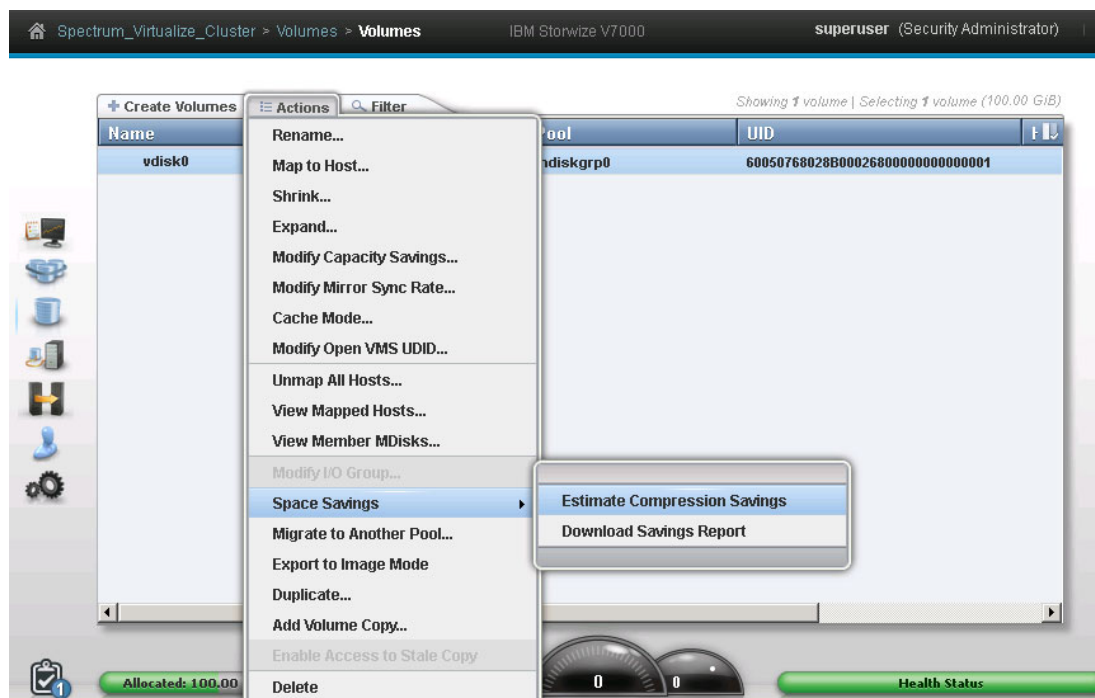


Figure 11-1 Estimating compression savings from the GUI

If using an older Spectrum Virtualize version or if you want to estimate the compression savings of a different storage system before migrating to 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 web site:

<http://www.ibm.com/webapp/set2/sas/f/comprestimator/home.html>

These are the best practices for using Comprestimator:

- ▶ Run the Comprestimator utility before implementing a Spectrum Virtualize solution or before implementing the Real-time Compression technology.
- ▶ Download the latest version of the utility from IBM if not using the version included with Spectrum Virtualize.
- ▶ Use Comprestimator to analyze volumes that contain as much active data as possible rather than volumes that are mostly empty. This 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 Recommended 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 above	>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 Spectrum Virtualize products, other IBM storage solutions, and other vendors storage subsystems. Disk Magic allows for in-depth environment analysis and is a recommended tool to estimate the performance of a system that is running Real-time Compression.

If you are an IBM Business Partner, more information together with the latest version can be found at this web site:

<http://www.ibm.com/partnerworld/wps/servlet/ContentHandler/SSPQ048068H83479I86>

If you are an IBM customer, please ask an IBM representative to evaluate the workload of your storage environment when implementing a Spectrum Virtualize Real-time Compression solution.

11.3 Verify available CPU resources

Before compression is enabled on Spectrum Virtualize systems, it is a good practice to 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			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 those 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.

Note: IBM recommends to customers who are planning to use Real-time Compression on 6-core SAN Volume Controller CG8 nodes to enhance their system with more CPU and cache memory resources that are dedicated to Real-time Compression. Because 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.

Note: IBM recommends to customers who are planning to use Real-time Compression on V7000 Gen2/Gen2+ to 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 IO group it is important to balance the compression workload. Consider a four-node (two IO groups) Spectrum Virtualize system with the following configuration:

- ▶ iogrp0: nodes 1 and 2 with 18 compressed volumes
- ▶ iogrp1: nodes 3 and 4 with 2 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 2 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 small number of compressed volumes (that is, 10 - 20)
- ▶ Alternative 2: Migrate compressed volumes from iogrp0 to iogrp1 and load balance 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.

The following recommendations are suggested:

- ▶ 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 default from version 6.4.1.5 and 7.1.0.1 and above). Use default grain size for FlashCopy with compressed volumes (64 KB) because this size reduces the performance impact when compressed FlashCopy targets are used.
- ▶ Consider the use of 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. 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, 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 nondisruptively 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 MDisk. 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 a Spectrum Virtualize system setup with some backend storage that supports compression (such as a Storwize product) and if you plan to implement compression, it is recommended to configure compression volumes on the Spectrum Virtualize system and not on the backend storage. This is to minimize 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.

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 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 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 best practise, it is recommended to always format new volumes during creation. This process zeros all blocks in the disks and eliminates traces of old data and it is the default behavior from version 7.7.

11.10 Mixed volumes in the same MDisk group

Note: Spectrum Virtualize version 7.3 onwards include a new cache architecture, which is not impacted 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 HDDs 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, run the following command to change to 600 ms system-wide:
 - `chsystem -compressiondestagemode on`

To check the current value:

- `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.

Starting with V7.1 and later, performance improvements were made that reduce the probability of a cache throttling situation. Yet, 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, it is recommended to separate compressed and non-compressed volumes to different storage pools. The compressed and non-compressed volumes do not share the cache partition and 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.

In this appendix, we provide preferred practice and guidelines for implementing Storwize family and SAN Volume Controller with IBM i.

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, we provide a short description of IBM i Storage management in this section.

Many host systems require you to take responsibility for how information is stored and retrieved from the disk units, along with providing 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 big 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, and so on. Therefore, application response time depends not only on disk response time but on many other factors, such as how large the IBM i storage pool is for the application, how frequently the applications closes files, whether it uses journaling, and so on.

A.1 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 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 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 duration of a batch job.

Planning for IBM i capacity

In order to correctly plan the disk capacity virtualized by SVC, or Storwize disk capacity for IBM i you need to 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 additional 9th sector; it stores the 8-byte headers of the 520 byte sectors in the 9th 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. Vice versa, 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 capacity 'overhead' on the storage subsystem so that the 8/9ths of the effective storage capacity which are available to IBM i, cover the needs of 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

In this section, we describe the guidelines and preferred practice for each type of connection.

Note: For updated and detailed information about the current requirements, refer to IBM System Storage Interoperation Center (SSIC) at the following web site:

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

and to the IBM i POWER® External Storage Support Matrix Summary at the following web site:

<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 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:

- ▶ 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 we recommend to connect SVC or Storwize to IBM i with Multipath using two or more FC adapters.

You can attach a maximum 64 LUNs to a port in the IBM i adapter, the LUNs 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 resides in POWER6® server or later. It requires IBM i V7.1 TR 6 or later when implemented in POWER6 or in POWER7, and it requires IBM i level V7.1 TR 8 or later when in POWER8. This type of connection requires switches that must be NPIV enabled.

For both resiliency and performance reasons we recommend to connect SVC or Storwize to IBM i in Multipath using two or more VIOS.

Following are the 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 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 attach a maximum of 64 LUNs to a port in the virtual FC adapter in IBM i, 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 virtual SCSI (VSCSI) mode requires that IBM i partition resides in POWER6 server or later. This type of connection requires IBM i V6.1.1 or later when IBM i resides in POWER6 or POWER7, and it requires IBM i V7.1 TR8 or later when in POWER8.

Using Multipath with two or more VIOS is recommended for resiliency and performance. When implementing Multipath with this type of connection keep in mind the following:

- ▶ IBM i Multipath is performed with two or more virtual SCSI (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 representing the same SVC/Storwize LUN.

- ▶ In addition to IBM i Multipath we also implement Multipath in each VIOS by using one of the Multipath drives, preferably SDDPCM driver. The paths connecting 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 32 LUNs to a port in the virtual SCSI adapter in IBM i, and with IBM i releases prior to 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 the 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

In this section we describe the values of certain attributes in VIOS that must be setup for Multipath, or are recommended to setup for the best performance.

FC adapter attributes

With either VIOS Virtual SCSI connection or NPIV connection we recommend to specify the following attributes for each SCSI I/O Controller Protocol Device (fscsi) device that connects a 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*

Setting of the specified values for the two attributes is 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 will be different, and it 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.

Setting *queue_depth* to 32 is recommended for performance reasons. By setting this value we make sure 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.

Setting *algorithm* to *load_balance* is recommended for performance reasons. By setting this value we ensure that the SDDPCM driver in VIOS will balance the I/O across available paths to Storwize or SVC.

Disk drives for IBM i

In this section we describe the best practice of how to implement internal disk drives in Storwize, and background storage of SVC and Storwize for IBM i host. Our 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.

When attaching Storwize with internal hard disk drives (HDDs) for IBM i, make sure that sufficient number of disk arms are provided to the IBM i workload. In general we recommend to implement HDDs with rotation 15 K RPM, however you can use also 10 K RPM HDDs providing that there are enough of them available for the IBM i host. Disk Magic modelling can help you determine the needed number of disk drives for expected IBM i performance.

When both Solid State Drives (SSD) and HDD are implemented make sure that a sufficiently large part of disk capacity resides on SSD. Typically we recommend at least 20% of IBM i capacity on SSD.

Exploitation of SSDs with Storwize or SVC is through Easy Tier. Even if you don't plan to install SSDs you can still use Easy Tier to evaluate your workload and provide information on the benefit you might gain by adding SSDs in the future.

We expect that IBM i workload will achieve the best performance when using disk capacity entirely from SSD.

With SVC, or when Storwize is implemented with background storage you should still consider enough HDD in the storage subsystem connected to Storwize or SVC, in order to accommodate IBM i workload peaks. When connecting both SSD or Flash disk and HDD as background storage make sure that a sufficient part of IBM i capacity resides on SSD or Flash storage.

We highly recommend using Disk Magic modelling *prior* to implementing a certain disk configuration for IBM i. In Disk Magic insert the current performance data of the IBM i workload, then insert the planned configuration, for example, Storwize with SSDs, Storwize with HDD and SSD with Easy Tier, SVC or Storwize with background storage of HDD, SSD, and Flash storage. When modelling Easy Tier specify the lowest skew level for IBM i workload.

Disk Magic will provide the predicted disk response time of IBM i on the planned disk configuration as well as the response time at workload growth.

Defining LUNs for IBM i

LUNs for IBM i host are defined from block-based storage, and we create them the same way as for open hosts. The minimum size of an IBM i LUN is 180 MB which provides 160 MB to IBM i due to block translation, and 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 reasons for this are:

- ▶ If more LUNs are attached to IBM i, the storage management uses more threads and thus enables better performance.
- ▶ The wait time component of disk response time, is lower when more LUNs are used, hence lower latency of disk IO operations.

On the other hand, 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 additional 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 best practices, our recommendations are as follows:

- ▶ For any IBM i disk pool (ASP) define all the LUNs to be the same size.
- ▶ 45 GB is the recommended minimum LUN size.
- ▶ You shouldn't define LUNs bigger than 200 GB.
- ▶ A minimum of 8 * LUNs for each ASP or LPAR is recommended.

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 prior to 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 may 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 as this 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, we recommend 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 may share the disk pool among IBM i workloads. Only very big critical workloads should reside in isolated disk pools.
- ▶ In Storwize using entirely SSD or Flash storage you may share the disk pool among IBM i workloads.
- ▶ We strongly recommend **not** to mix 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 running 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, refer to IBM System Storage Interoperation Center (SSIC) at the following web site:

<http://www.ibm.com/systems/support/storage/ssic/interoperability.wss>

and to the IBM i POWER External Storage Support Matrix Summary at the following web site:

<https://www.ibm.com/support/techdocs/atsmastr.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 (IO/sec) and data rate (MB/sec) that a port in a particular adapter can sustain at 70% utilization, and 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 may 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
IO/sec per port	52500 IO/sec	23100 IO/sec
Sequential throughput per port	1330 MB/sec	770 MB/sec
Transaction throughput per port	840 MB/sec	371 MB/sec

Zoning SAN switches

With native connection and the connection in VIOS_NPIV, zone the switches so that one World Wide 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 way we ensure resiliency for the IO to/from a LUN assigned to that WWPN: if the preferred node for that LUN fails the IO rate will continue 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, the two zones overlapping 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, keeping in mind that there are a maximum of 8 paths available from VIOS to SVC or Storwize. SVC or Storwize ports 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 8 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. We recommend to implement more than two paths for workloads where high IO 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 best 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 may 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/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 path through the non-preferred node are passive. Multipath employs the load balancing among the paths to a LUN that go through the node which is preferred for that LUN.

Boot from SAN

All connection options, Native, VIOS_NPIV, and VIOS Virtual SCSI, support IBM i Boot from SAN. IBM i boot disk (LoadSource) resides on a Storwize or SVC LUN which is connected the same way as the other LUNs; there aren't 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: first you add the LUNs from two virtual adapters with each adapter connecting one to-be mirrored half of LUNs. 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 which allows for a flexible and configurable RPO allowing GM to be maintained during peak periods of bandwidth constraint.

It is essential that you size the bandwidth of Metro Mirror or Global Mirror links to accommodate the peaks of IBM i workload, or else risk impact to 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 a sufficient number of disk drives are available to the FlashCopy target LUNs to keep good performance of production IBM i during the time that FlashCopy relationships are active. This 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 depending on the write operations to source or target LUNs.



Business continuity

Business continuity (BC) and continuous application availability are among the top 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 Spectrum Virtualize, and HyperSwap configurations specific for Storwize. Technical details or implementation guidelines are not presented in this chapter as they are described in separate publications.

11.11 Business Continuity with Stretched Cluster

Within standard implementations of 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 system 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 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 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,

Currently IBM supports Enhanced Stretched Cluster configurations, where nodes can be separated by a distance of up to 300 km, in specific configurations using Fibre Channel (FC) or Fiber Channel over IP (FC/IP) switch or Multi Protocol Router (MPR) inter-switch links (ISLs) between different locations.

11.12 Business Continuity with Enhanced Stretched Cluster

Software version 7.2 introduced the Enhanced Stretched Cluster (ESC) feature that further improved the Stretched Cluster configurations. 7.2 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 7.5, the site awareness concept has been extended to hosts. This enables more efficiency for host I/O traffic through the SAN, and easier host path management.

Version 7.6 introduced a new feature for stretched systems, the IP Quorum application. Using an IP-based quorum application as the quorum device for the third site, no Fibre Channel connectivity is required. Java applications run on hosts at the third site.

IP Quorum details can be found in the IBM Knowledge Center for SAN Volume Controller:

<https://ibm.biz/BdsvFN>

11.13 Business Continuity with HyperSwap

The HyperSwap high availability feature in the IBM Spectrum Virtualize and Storwize allows business continuity in the event of 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 300km 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 will automatically optimize itself to minimize data transmitted between sites and to minimize host read and write latency.

Some of the key features of HyperSwap are:

- ▶ Works with Spectrum Virtualize and IBM Storwize V7000, V5000, V7000 unified hardware as well.
- ▶ 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 Spectrum Virtualize I/O groups in a clustered system using the Active-Active Metro Mirror relationship under the covers, and appears as one volume to the host.
- ▶ Works with the standard multipathing drivers that are available on a wide variety of host types, with no additional host support required to access the highly available volume

11.14 IP Quorum Configuration

In a stretched configuration or HyperSwap configuration, you must use a third, independent site to house quorum devices. To use a quorum disk as the quorum device, this third site must use Fibre Channel connectivity together with an external storage system. Sometimes, Fibre Channel connectivity is not possible. In a local environment, no extra hardware or networking, such as Fibre Channel or SAS-attached storage, is required beyond what is normally always provisioned within a system.

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. To provide quorum resolution, use the **mkquorumapp** command to generate a Java application that is copied from the system and run on a host at a third site. The maximum number of applications that can be deployed is five. Currently, supported Java runtime environments (JREs) are IBM Java 7.1 and IBM Java 8

For details on IP Quorum requirements and installation, refer to the IP Quorum configuration section in the Knowledge Center here:

<https://ibm.biz/BdsvFN>

Note: The IP Quorum configuration process has been integrated into the Spectrum Virtualize GUI from v7.7.1 onwards.

Implementation of Stretched Cluster, Enhanced Stretched Cluster and HyperSwap

For further technical details and implementation guidelines on deploying Stretched Cluster or Enhanced Stretched Cluster refer to: *IBM Spectrum Virtualize and SAN Volume Controller Enhanced Stretched Cluster with VMware*, SG24-8211.

For further technical details and implementation guidelines on deploying HyperSwap, please refer to *IBM Storwize V7000, Spectrum Virtualize, HyperSwap, and VMware Implementation*, SG24-8317.

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 more information about the topics in this book. Note that some publications that are referenced in this list might be available in softcopy only:

- ▶ *Implementing the IBM System Storage SAN Volume Controller V7.2*, SG24-7933
- ▶ *Implementing the IBM Storwize V7000 V7.2*, SG24-7938
- ▶ *IBM b-type Gen 5 16 Gbps Switches and Network Advisor*, SG24-8186
- ▶ *Introduction to Storage Area Networks and System Networking*, SG24-5470
- ▶ *IBM SAN Volume Controller and IBM FlashSystem 820: Best Practices and Performance Capabilities*, REDP-5027
- ▶ *Implementing the IBM SAN Volume Controller and FlashSystem 820*, SG24-8172
- ▶ *Implementing IBM FlashSystem 840*, SG24-8189
- ▶ *IBM FlashSystem in IBM PureFlex System Environments*, TIPS1042
- ▶ *IBM FlashSystem 840 Product Guide*, TIPS1079
- ▶ *IBM FlashSystem 820 Running in an IBM StorwizeV7000 Environment*, TIPS1101
- ▶ *Implementing FlashSystem 840 with SAN Volume Controller*, TIPS1137
- ▶ *IBM FlashSystem V840 Enterprise Performance Solution*, TIPS1158
- ▶ *IBM Midrange System Storage Implementation and Best Practices Guide*, SG24-6363
- ▶ *IBM System Storage b-type Multiprotocol Routing: An Introduction and Implementation*, SG24-7544
- ▶ *IBM Tivoli Storage Area Network Manager: A Practical Introduction*, SG24-6848
- ▶ *Tivoli Storage Productivity Center for Replication for Open Systems*, SG24-8149
- ▶ *Tivoli Storage Productivity Center V5.2 Release Guide*, SG24-8204
- ▶ *Implementing an IBM b-type SAN with 8 Gbps Directors and Switches*, SG24-6116

You can search for, view, download, or order these documents and other Redbooks, Redpapers, Web Docs, draft, and other materials, at this website:

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

Other resources

The following publications also are 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, which is available at this website:
http://pic.dhe.ibm.com/infocenter/svc/ic/topic/com.ibm.storage.svc.console.doc/svc_bkmap_configuidebk.pdf
- ▶ *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*, which is available at this website:
<http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD105195>
- ▶ *Considerations and Comparisons between IBM SDD for Linux and DM-MPIO*, which is available at this website:
http://www.ibm.com/support/docview.wss?rs=540&context=ST52G7&q1=linux&uid=ssg1S7001664&loc=en_US&cs=utf-8&lang=en

Referenced websites

The following websites are also relevant as further information sources:

- ▶ IBM Storage home page:
<http://www.storage.ibm.com>
- ▶ IBM site to download SSH for AIX:
<http://oss.software.ibm.com/developerworks/projects/openssh>

- ▶ IBM Tivoli Storage Area Network Manager site:
<http://www-306.ibm.com/software/sysmgmt/products/support/IBMTivoliStorageAreaNetworkManager.html>
- ▶ IBM TotalStorage Virtualization home page:
<http://www-1.ibm.com/servers/storage/software/virtualization/index.html>
- ▶ SAN Volume Controller supported platform:
<http://www-1.ibm.com/servers/storage/support/software/sanvc/index.html>
- ▶ SAN Volume Controller Information Center:
<http://pic.dhe.ibm.com/infocenter/svc/ic/index.jsp>
- ▶ Cygwin Linux-like environment for Windows:
<http://www.cygwin.com>
- ▶ Microsoft Knowledge Base Article 131658:
<http://support.microsoft.com/support/kb/articles/Q131/6/58.asp>
- ▶ Microsoft Knowledge Base Article 149927:
<http://support.microsoft.com/support/kb/articles/Q149/9/27.asp>
- ▶ Open source site for SSH for Windows and Mac:
<http://www.openssh.com/windows.html>
- ▶ Sysinternals home page:
<http://www.sysinternals.com>
- ▶ Subsystem Device Driver download site:
<http://www-1.ibm.com/servers/storage/support/software/sdd/index.html>
- ▶ Download site for Windows SSH freeware:
<http://www.chiark.greenend.org.uk/~sgtatham/putty>

Help from IBM

IBM Support and downloads:

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

IBM Global Services:

<http://www.ibm.com/services>

To determine the spine width of a book, you divide the paper PPI into the number of pages in the book. An example is a 250 page book using Plainfield opaque 50# smooth which has a PPI of 526. Divided 250 by 526 which equals a spine width of .4752". In this case, you would use the .5" spine. Now select the Spine width for the book and hide the others: **Special>Conditional Text>Show/Hide>SpineSize(->Hide:>Set** . Move the changed Conditional text settings to all files in your book by opening the book file with the spine.fim still open and **File>Import>Formats** the Conditional Text Settings (ONLY!) to the book files.

Draft Document for Review February 15, 2017 9:43 am

7521spine.fim 375



IBM System Storage SAN Volume Volume

SG24-7521-04
ISBN DocISBN



(1.5" spine)
1.5"<-> 1.998"
789 <-> 1051 pages



IBM System Storage SAN Volume Controller and Storwize V7000

SG24-7521-04
ISBN DocISBN



(1.0" spine)
0.875"<->1.498"
460 <-> 788 pages

Redbooks

IBM System Storage SAN Volume Controller and Storwize V7000

SG24-7521-04
ISBN DocISBN



(0.5" spine)
0.475"<->0.873"
250 <-> 459 pages

Redbooks

IBM System Storage SAN Volume Controller and Storwize V7000 Best

(0.2" spine)
0.17"<->0.473"
90<->249 pages

(0.1" spine)
0.1"<->0.169"
53<->89 pages

To determine the spine width of a book, you divide the paper PPI into the number of pages in the book. An example is a 250 page book using Plainfield opaque 50# smooth which has a PPI of 526. Divided 250 by 526 which equals a spine width of .4752". In this case, you would use the .5" spine. Now select the Spine width for the book and hide the others: **Special>Conditional Text>Show/Hide>SpineSize(->Hide:>Set** . Move the changed Conditional text settings to all files in your book by opening the book file with the spine:fm still open and **File>Import>Formats** the Conditional Text Settings (ONLY!) to the book files.

Draft Document for Review February 15, 2017 9:43 am

7521spine.fm 376



IBM System Storage SAN Volume

SG24-7521-04
ISBN DocISBN

(2.5" spine)
2.5" <-> nnn.n"
1315<-> nnn pages



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

SG24-7521-04
ISBN DocISBN



(2.0" spine)
2.0" <-> 2.498"
1052 <-> 1314 pages



SG24-7521-04

ISBN DocISBN

Printed in U.S.A.

Get connected

